

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

VADEMÉCUM PRÁCTICO ESTADÍSTICA APLICADA



Vademécum Práctico
Estadística Aplicada



EJERCICIOS EXAMEN ESTADÍSTICA APLICADA

1. En una cadena de televisión se hizo una encuesta a 2500 personas para saber la audiencia de un debate y de una película que se emitieron en horas distintas: 2100 personas vieron la película, 1500 vieron el debate y 350 no vieron ninguno de los dos programas.

Eligiendo al azar a uno de los encuestados, se desea saber:

- Probabilidad de que viera la película y el debate.
- Probabilidad de que viera la película, sabiendo que vio el debate.
- Habiendo visto la película, probabilidad de que viera el debate.

2. Un psicólogo de una compañía aérea, por experiencias anteriores, conoce que el 90% de los tripulantes de cabina (TCP) que inician un determinado tratamiento técnico terminan con éxito. La proporción de TCPs con entrenamiento y con experiencia previa es del 10% de entre los que completaron con éxito su entrenamiento y del 25% de entre aquellos que no terminaron con éxito su entrenamiento. Se desea saber:

- Probabilidad de que un TCP con experiencia previa supere el entrenamiento con éxito.
- ¿La experiencia previa influye en el éxito del entrenamiento?

3. Sean tres sucesos incompatibles A, B y C, donde $P(A) = 0,5$, $P(B) = 0,25$ y $P(C) = 0,2$.

Se pide: $P(\bar{A} \cap \bar{B})$, $P(\bar{A} \cap \bar{B} \cap \bar{C})$, $P(\bar{A} \cap \bar{B} \cap C)$, $P(B - A)$ y probabilidad de que ocurra exactamente uno de los sucesos.

4. Tres tiradores hacen una descarga simultánea. Las probabilidades de hacer blanco son, respectivamente, 0,6, 0,5 y 0,4. Calcular las probabilidades de los sucesos:

- Algún tirador hace blanco.
- Exactamente dos tiradores hacen blanco.
- El tercer tirador hace blanco, sabiendo que dos primeros lo han hecho.

5. Sean A y B dos sucesos tales que $P(A) = 0,6$. Calcular $P(A \cap \bar{B})$ en cada caso:

- A y B mutuamente excluyentes
- A está contenido en B
- B está contenido en A y $P(B) = 0,3$
- $P(A \cap B) = 0,1$

6. Se sabe que $P(B / A) = 0,9$, $P(A / B) = 0,2$ y $P(A) = 0,1$

- Calcular $P(A \cap B)$ y $P(B)$
- ¿Son independientes los sucesos A y B?
- Calcular $P(A \cup \bar{B})$

7. Una empresa de mensajería sabe que en condiciones normales un paquete es entregado en plazo el 90% de las veces, aunque si hay sobrecarga de trabajo (que ocurre un 5% de las veces) el porcentaje de retrasos se eleva al 30%.

- Cuál es la probabilidad de que un paquete llegue en plazo a su destino.
- Sabiendo que se ha recibido una queja por retraso en el envío, el mensajero afectado aduce que ese día hubo sobrecarga de trabajo, aunque realmente no recuerda bien que sucedió. ¿Qué probabilidad hay de que efectivamente esté en lo cierto?.

8. En una región se cobra a los visitantes de los parques naturales, estimando que la variable aleatoria número de personas que visitan el parque en coche sigue la siguiente distribución:

x_i	1	2	3	4	5
$P(X = x_i)$	0,15	0,2	0,35	0,2	0,1

- Hallar el número medio de visitantes por vehículos.
- Hallar cuánto debe pagar cada visitante para que la ganancia por coche sea 2 euros.
- Si cada persona paga p euros, ¿cuál es la ganancia esperada en un día en que entran mil vehículos?

9. Un estudiante busca información del sector aeronáutico en tres manuales, las probabilidades de que esa información se encuentre en el primero, segundo o tercer manual, respectivamente, son iguales a 0,6 , 0,7 y 0,8.

¿Cuál es la probabilidad de que la información figure sólo en dos manuales?.

10. En el punto de partida de un laberinto hay tres orificios iguales A, B y C. Si una rata elige A vuelve al punto de partida después de recorrer dos metros. Si elige B recorre cinco metros y vuelve al mismo punto. Si elige C sale al exterior recorriendo un metro. ¿Por término medio que distancia recorre una rata antes de salir, si siempre elige un orificio distinto de los seleccionados en veces anteriores?

11. Sean A, B y C tres sucesos mutuamente independientes con $P(A) = P(B) = P(C) = p$, donde $0 < p < 1$. Calcular la probabilidad de que ocurran exactamente dos de los tres sucesos considerados.

12. En una cadena de televisión se hizo una encuesta a 3000 personas para saber la audiencia de un debate y de una película que se emitieron en horas distintas: 2400 personas vieron la película, 1600 vieron el debate y 400 no vieron ninguno de los dos programas. Se elige al azar uno de los encuestados:

- ¿Cuál es la probabilidad de que viera la película y el debate?.
- ¿Cuál es la probabilidad de que viera la película, sabiendo que vio el debate?.
- Sabiendo que vio la película, ¿Cuál es la probabilidad de que viera el debate?.

13. Si la probabilidad de que ocurra un suceso A es $1/5$

- ¿Cuál es el mínimo número de veces que hay que repetir el experimento para que la probabilidad de que ocurra al menos una vez sea mayor que $1/2$?
- ¿Cuál es la probabilidad de que ocurra al menos dos veces A al realizar 5 veces el experimento?

14. En una población de mujeres, las puntuaciones de un test de ansiedad-riesgo siguen una distribución normal $N(25, 10)$. Al clasificar la población en cuatro grupos de igual tamaño, ¿cuales serán las puntuaciones que delimiten estos grupos?

15. El departamento comercial de una industria alimenticia conoce que 2 de cada 10 consumidores reconocen su producto en una prueba a ciegas. ¿Cuántas pruebas ciegas de sabor deberían hacerse para que la proporción de que los que conocen la marca oscile entre el 16% y el 24% con una probabilidad mínima de 0,8?

16. Se está experimentando una nueva vacuna para la malaria que resulta efectiva en el 60% de los casos. Si se eligen al azar 45 personas. Se pide:

- Probabilidad de que en el grupo elegido la vacuna sea efectiva en 27 personas.
- Probabilidad de que sea efectiva en menos de 20 personas.

17. Una maca de frutas ha observado que el peso medio de los melones en gramos sigue una distribución normal $N(1700, 100)$. Hallar:

- Probabilidad de que los melones pesen menos de 1500 gramos y más de 2000 gramos.
- Sabiendo que son rechazados para la explotación aquellos melones que difieren más de 300 gramos del promedio. Determinar la proporción de melones rechazados.

18. Un ascensor limita el peso de sus cuatro ocupantes a 300 kilogramos. Si el peso de una persona sigue una distribución normal $N(71, 7)$, calcular la probabilidad de que el peso 4 personas supere los 300 kilogramos.

19. Una empresa aeronáutica vende cilindros para motor en paquetes de 25 unidades. Se sabe que el peso de cada cilindro es una variable aleatoria que tienen una media de 12 Kg y una desviación típica de 2 Kg. Se pide:

- ¿Cuál es la probabilidad de que en un paquete de 25 unidades escogidas aleatoriamente, el peso medio de cada cilindro sea superior a 13 kilogramos?
- ¿Cuál sería el resultado si la varianza poblacional fuera desconocida?, suponiendo entonces que la cuasidesviación típica es de 3 Kg?

20. El error cometido en expedir tiques por una maquina en el aeropuerto sigue una normal $N(0, \sigma)$.

a) ¿Cuál es la probabilidad de que el error cometido en valor absoluto de una medida cualquiera sea al menos σ ?

b) ¿Cuánto valdría la probabilidad si se toma como medida la media aritmética de 10 medidas independientes?

21. Las variables aleatorias en estudio son independientes.

Sean: $X_1 \sim N(10, 2)$, $X_2 \sim N(15, 2)$ y $X_3 \sim N(20, 2)$ donde $Y = 5X_1 + 4X_2 - 3X_3$ y

$$U = \left(\frac{X_1 - 10}{2} \right)^2 + \left(\frac{X_2 - 15}{2} \right)^2 + \left(\frac{X_3 - 20}{2} \right)^2.$$

Calcular: a) $P[20 \leq Y \leq 30]$ b) $P[U \leq 11]$

22. El número de millones de metros cúbicos que tiene un embalse sigue una distribución normal $N(980, 50)$. El consumo diario de las poblaciones que sirve es una normal $N(85, 30)$.

Si se sabe que durante una tormenta la cantidad de agua que se embalsa es una normal $N(50, 25)$. Un día han caído dos tormentas, calcular la probabilidad de que al final del día el agua embalsada sea menor o igual que 980 metros cúbicos.

23. Sean las variables aleatorias: $X \sim B(8, 0,4)$, $Y \sim B(7, 0,4)$, $U \sim B(55, 0,03)$, $V \sim B(65, 0,03)$

¿Qué distribución siguen las variables $X + Y$, $U + V$?

24. Se tienen dos variables aleatorias cuya distribución es normal: $X \sim N(5, 1)$, $Y \sim N(3, 2)$.

Obtener la esperanza matemática y la varianza de la variable $W = 2X - 3Y$ y de la variable $Z = (X - Y) / 2$

25. En cierta población, en una muestra de tamaño 100, la media muestral $\bar{x}$ de una característica sigue una distribución normal. Se conoce que $P(\bar{x} \leq 75) = 0,58$ y

$P(\bar{x} > 80) = 0,04$. Hallar la media y la desviación típica poblacional.

26. Para analizar el peso promedio de niños y niñas, siguiendo ambos pesos una distribución normal, se utiliza una muestra aleatoria de 20 niños y 25 niñas. El promedio de los pesos de los niños es 45 kg. con una desviación típica de 6,4 kg., mientras que el promedio del peso de las niñas es 38 kg. y una desviación típica de 5,6 kg. ¿Cuál es la probabilidad de que en la muestra el peso promedio de los niños sea al menos 10 kg. mayor que el de las niñas?

27. La renta disponible mensualmente de una familia rural es una variable aleatoria que sigue una distribución normal $N(905, 450)$, expresados en euros. Para realizar un estudio sobre los hábitos de esta comunidad rural, se ha tomado una muestra aleatoria simple de 1000 familias. Calcular la probabilidad de que la cuasidesviación típica de la muestra sea superior a 430 euros.

28. Una encuesta del periódico constó de 320 trabajadores de una multinacional que fueron despedidos entre 2017-2018, encontrando que el 20% habían estado sin trabajo por lo menos durante dos años. Si el periódico tuviera que seleccionar otra muestra aleatoria de 320 empleados despedidos entre 2017-2018.

¿Cuál sería la probabilidad de que el porcentaje muestral de trabajadores sin empleo durante por lo menos dos años, difiera del porcentaje obtenido en la primera encuesta, en un 5% o más?.

29. Un individuo juega con probabilidad de ganar igual a $1/2$ en cada partida. Si gana en una partida obtiene 5 euros y si pierde paga 5 euros. En una tarde juega 400 partidas. ¿Qué cantidad debe llevar si quiere tener una probabilidad de 0,95 de hacer frente a posibles pérdidas?

30. Un estudio de la DGT estima que el número de horas prácticas necesarias para la obtención del permiso de conducir sigue una distribución normal $N(24, 3)$. Se sabe que la autoescuela Fuenterrebollo ingresa por alumno una parte fija de 250 euros, más 23 euros por hora de práctica.

- ¿Qué probabilidad hay de obtener el permiso de conducir con 20 horas de prácticas o menos?
- ¿Cuántas horas de prácticas ha necesitado un conductor para obtener el permiso si el 68% de los conductores ha necesitado más horas de prácticas que él?
- Calcular el ingreso por alumno esperado.
- Calcular la desviación típica del ingreso por alumno.

31. Sea X una variable aleatoria que se distribuye según una normal $N(\mu, 4)$. Suponiendo que el tamaño muestral $n = 10$. Se pide:

- Determinar el número k tal que, exista una probabilidad de 0,99 de que el estadístico desviación típica muestral, sea mayor que él.
- Determinar el número k tal que, exista una probabilidad de 0,99 de que el estadístico cuasidesviación típica muestral, sea mayor que él.

32. En una muestra aleatoria de personas se analizan algunos hábitos de la vida, habiendo recogido datos de las siguientes variables:

$X_1 \equiv$ Estado general de salud: muy bueno (3), bueno (2), regular (1), malo (0)

$X_2 \equiv$ Sexo: mujer (1), hombre (0)

$X_3 \equiv$ Nivel del ejercicio diario: intenso (2), moderado (1), ninguno (0)

Realizadas las tablas de contingencia correspondientes, se calcularon los siguientes estadísticos para contrastar la asociación:

$$a) \chi^2(X_1, X_2) = 8 \quad b) \chi^2(X_2, X_3) = 4,5 \quad c) \chi^2(X_1, X_3) = 6,1$$

Con la información facilitada, a un nivel de significación del 5%, elaborar un diagnóstico para cada una de las parejas de variables.

33. La probabilidad de que un banco reciba un cheque sin fondos es 0.01

- a) Si en una hora reciben 20 cheques, ¿cuál es la probabilidad de que tenga algún cheque sin fondos?
- b) El banco dispone de 12 sucursales en la ciudad, ¿cuál es la probabilidad de que al menos cuatro sucursales reciban algún cheque sin fondos?
- c) La media del valor de los cheques sin fondos es de 600 euros. Sabiendo que el banco trabaja 6 horas diarias, ¿qué cantidad no se espera pagar?
- d) Si se computasen los 500 primeros cheques, ¿cuál es la probabilidad de recibir entre 3 y 6 (inclusive) cheques sin fondos?

34. Una compañía de seguros garantiza pólizas de seguros individuales contra retrasos aéreos de más de doce horas. Una encuesta ha permitido estimar a lo largo de un año que cada persona tiene una probabilidad de una de cada mil de ser víctima de un retraso aéreo que esté cubierto por este tipo de póliza y que la compañía aseguradora podrá vender una media de cuatro mil pólizas al año.

Se pide hallar las siguientes probabilidades:

- a) Que el número de retrasos cubiertos por la póliza no pase de cuatro por año
- b) Número de retrasos esperados por año
- c) Que el número de retrasos sea superior a dos por año
- d) Que ocurran doce retrasos por año

35. El jefe de recursos humanos de una empresa realiza un test de diez ítems a los aspirantes a un puesto, teniendo en cada ítems cuatro posibles respuestas, de las que sólo una es correcta. Suponiendo que los aspirantes teniendo la misma probabilidad de responder.

Se pide hallar las probabilidades para el aspirante:

- a) Conteste todos los ítems mal
- b) Conteste al menos cuatro ítems bien
- c) Conteste entre cuatro y seis ítems bien
- d) Conteste todos los ítems bien
- e) Conteste menos de tres ítems bien

36. Una fábrica aeronáutica produce en cada turno 100000 bolas para rodamientos, siendo la probabilidad de bola defectuosa 0,04. Las bolas se supervisan todas, depositando las defectuosas en un recipiente que se vacía al final de cada turno. ¿Cuántas bolas ha de contener el recipiente para que la probabilidad de que su capacidad no sea rebasada sea 0,95?

37. El 7% de los pantalones de una determinada marca salen con algún defecto. Se empaquetan en caja de 80 pantalones para diferentes tiendas. ¿Cuál es la probabilidad de que en una caja haya entre 8 y 10 pantalones defectuosos?

38. Un servicio dedicado a la reparación de electrodomésticos recibe por término medio 15 llamadas diarias. Determinar la probabilidad de que reciba un día más de 20 llamadas.

39. En una fábrica se sabe que la probabilidad de que r artículos sean defectuosos es

$$P[X = k] = \frac{4^k \cdot e^{-4}}{k!}. \text{ Determinar la probabilidad de que en 100 días el número de artículos}$$

defectuosos esté comprendido entre (400, 600)

40. Se ha realizado una muestra aleatoria simple (m.a.s.) de tamaño 10 de una población considerada normal, llegando a que la varianza muestral es 4.

Calcular la probabilidad $P[|\bar{x} - \mu| \leq 1,51]$

41. Se realiza una encuesta para conocer la proporción de españoles a los que no le gusta el fútbol, tomando una muestra aleatoria simple (m.a.s.) de tamaño 100.

Por análisis anteriores se conoce que dicha proporción es del 45%. Calcular la probabilidad de que la proporción muestral sea superior al 52%.

42. Calcular la media, varianza y coeficiente de variación de la variable aleatoria discreta que tiene como función de distribución:

$$F(x) = \begin{cases} 0 & x < 2 \\ 0,2 & 2 \leq x < 4 \\ 0,55 & 4 \leq x < 6 \\ 0,85 & 6 \leq x < 8 \\ 1 & x \geq 8 \end{cases}$$

43. La variable aleatoria discreta X tiene como distribución de probabilidad

$X = x_i$	1	2	3	4
$P(X = x_i)$	0,30	0,25	0,10	0,35

Se realiza un cambio de origen hacia la izquierda de dos unidades y un cambio de escala de tres unidades. Calcular y razonar propiedades:

a) Media y varianza de la X .

b) Media, varianza y coeficiente de variación de la variable transformada por el cambio de origen.

c) Media, varianza y coeficiente de variación de la variable transformada por el cambio de escala.

d) Media, varianza y coeficiente de variación de la variable transformada por el cambio de origen y escala.

44. Para aprobar la asignatura de estadística teórica se realiza un test con veinte ítems.

Sabiendo que una persona determinada tiene una probabilidad de 0,8 de contestar bien cada ítem. Se pide:

- Probabilidad de que la primera pregunta que contesta bien sea la tercera que hace.
- Para aprobar el test es necesario contestar diez ítems bien. ¿Cuál es la probabilidad de que apruebe al contestar el doceavo ítem?

45. Un dispositivo está formado por muchos elementos que trabajaban independientemente, siendo la probabilidad de fallo durante la primera hora de trabajo muy pequeña e iguales en todos los elementos.

Si la probabilidad de que en ese tiempo falle por lo menos un elemento es 0,98.

- Hallar la media y desviación típica del número de elementos que fallen en la primera hora.
- Calcular la probabilidad de que fallen a lo sumo dos elementos en ese tiempo.

46. En una caja hay 5 triángulos, 3 círculos y 2 rectángulos. Realizando extracciones con reemplazamiento, se piden las siguientes probabilidades:

- Al realizar 8 extracciones, se obtengan en 4 ocasiones un círculo.
- Se necesiten 8 extracciones para obtener 4 círculos.
- Que aparezca el primer círculo en la 8 extracción.
- Al realizar 8 extracciones aparezcan 3 triángulos, 3 círculos y 2 rectángulos.
- Al realizar 6 extracciones sin reemplazamiento aparezcan en 2 ocasiones un círculo.

47. Por prescripción médica, un enfermo debe hacer una toma de tres píldoras de un determinado medicamento. De las doce píldoras que contiene el envase hay cuatro en malas condiciones. Se pide:

- Probabilidad de que tome sólo una buena.
- Probabilidad de que de las tres píldoras de la toma al menos una esté en malas condiciones.
- ¿Cuál es el número de píldoras que se espera tome el enfermo en buenas condiciones en cada toma?
- Si existe otro envase que contenga cuarenta píldoras, de las que diez se encuentran en malas condiciones. ¿Qué envase sería más beneficiosos para el enfermo?

48. El consumo familiar de cierto artículo se distribuye uniformemente con esperanza 10 y varianza unidad. Determinar la probabilidad de que el consumo de dicho artículo se encuentre comprendido entre 8 y 12 unidades.

49. En un cine de verano hay instaladas 800 sillas, sabiendo que el número de asistentes es una variable aleatoria de media 600 y desviación típica 100.

¿Qué probabilidad existe de que el número de personas que vaya al cine un día cualquiera sea superior al número de sillas instaladas?

50. Se desea conocer el número de automóviles que se deben poner a la venta durante un periodo determinado para que se satisfaga una demanda media de 300 unidades con una desviación típica de 100 unidades, con una probabilidad no inferior al 75%.

51. La demanda media de un producto es de 100 unidades con una desviación típica de 40 unidades. Calcular la cantidad del producto que se debe tener a la venta para satisfacer la demanda de forma que puedan ser atendidos al menos el 80% de los clientes.

52. La variable discreta X tiene como distribución de probabilidad

$$P(X = k) = \frac{1}{10} \quad \text{siendo } k = 2, 3, \dots, 11$$

Se pide:

- | | |
|----------------------------|----------------------|
| a) Función de distribución | b) $P(X > 7)$ |
| c) $P(X < 5)$ | d) $P(3 \leq X < 7)$ |

53. Al lanzar cuatro monedas se considera el número de escudos obtenidos. De la variable aleatoria X así obtenida, se pide:

- Ley de probabilidad. Representación gráfica
- Función de distribución. Representación gráfica
- Esperanza matemática y varianza
- Mediana y moda de la distribución
- Probabilidad de obtener más de uno y menos de tres escudos

54. Sea (X, Y) una variable aleatoria bidimensional con función de densidad:

$$f(x, y) = \begin{cases} k & 0 < y < x < 1 \\ 0 & \text{restantes valores} \end{cases}$$

- Hallar k para que sea función de densidad.
- Hallar las funciones de densidad marginales. ¿Son X e Y independientes?
- Hallar las funciones de distribución marginales.
- Hallar las funciones de densidad condicionadas.

55. Una variable aleatoria continua X tiene por función de distribución:

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{2} & 0 \leq x \leq 1 \\ 2x - \frac{x^2}{2} - 1 & 1 < x \leq 2 \\ 1 & x > 2 \end{cases}$$

Se pide:

- Hallar la función de distribución y representarla
- Media, varianza, desviación típica y coeficiente de variación
- $P\left(\frac{1}{2} < X \leq \frac{3}{2}\right)$

56. El flujo de demanda de teléfonos móviles (en miles a la hora) de una determinada fábrica se ajusta a una variable aleatoria con la siguiente función de densidad:

$$f(x) = \begin{cases} k - x & \text{si } 0 \leq x \leq 1 \\ 0 & \text{en el resto} \end{cases}$$

- Hallar k para que $f(x)$ sea función de densidad
- Hallar la función de distribución $F(x)$
- Hallar la esperanza y la varianza
- $P\left[\frac{1}{2} \leq x \leq \frac{3}{4}\right]$

57. Una variable aleatoria continua X tiene por función de distribución:

$$F(x) = \begin{cases} 0 & x < 1 \\ x - 1 & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

- Calcular la función de densidad o función de cuantía
- Calcular la media, mediana y coeficiente de variación

58. Una variable aleatoria continua X tiene por función de densidad

$$f(x) = \begin{cases} 1-x & 0 \leq x < 1 \\ x-1 & 1 \leq x \leq 2 \\ 0 & \text{otros casos} \end{cases}$$

Se pide:

- Representa la función de densidad
- Hallar la función de distribución y su gráfica
- $P(0 \leq X \leq 1)$ $P(-2 \leq X \leq 2)$ $P\left(\frac{1}{2} \leq X < \infty\right)$

59. Sea (X, Y) una variable aleatoria bidimensional con función de densidad:

$$f(x, y) = \begin{cases} 1 & |y| < x, \ 0 < x < 1 \\ 0 & \text{restantes valores} \end{cases}$$

- Comprobar que $f(x, y)$ es función de densidad.
- Hallar las medias de X e Y .
- Hallar las probabilidades: $P\left[X < \frac{1}{2}; Y < 0\right]$ y $P\left[X > \frac{1}{2}; -\frac{1}{2} < Y < \frac{1}{2}\right]$

60. La función de densidad asociada a la emisión de billetes de una compañía aérea es:

$$f(x, y) = \begin{cases} x+y & 0 < x < 1 \quad 0 < y < 1 \\ 0 & \text{en el resto} \end{cases}$$

- Hallar la función de distribución.
- Hallar las funciones de densidad marginales de X e Y .
- ¿Son X e Y independientes?

61. La función de distribución asociada a un fenómeno de la naturaleza es:

$$F(x, y) = \begin{cases} (1-e^{-x}) \cdot (1-e^{-y}) & x > 0, \ y > 0 \\ 0 & \text{en el resto} \end{cases}$$

- Hallar la función de densidad.
- Hallar las funciones de densidad marginales de X e Y .
- Hallar las funciones de densidad condicionadas.
- Calcular el coeficiente de correlación.

62. La venta en un mercado de abastos lleva asociada la función:

$$f(x, y) = \begin{cases} k \left[\frac{xy}{2} + 1 \right] & 0 < x < 1 \quad -1 < y < 1 \\ 0 & \text{en el resto} \end{cases}$$

- Hallar k para que sea función de densidad
- Hallar la función de distribución
- Funciones de densidad marginales y condicionadas
- Se considera la transformación $Z = X - Y$ y $T = X + 2Y$, hallar la función de densidad de la variable (Z, T)

63. Gran número de fenómenos aeronáuticos tienen asociada una variable aleatoria con ley de probabilidad:

$$f(x) = \begin{cases} ke^{-kx} & 0 < x < \infty \quad k > 0 \\ 0 & \text{en otros casos} \end{cases}$$

Se pide:

- ¿Puede tomar k cualquier valor?
- Para $k = 0,1$ representar la función de densidad, la función de distribución y su gráfica
- Siendo $k = 0,1$ hallar $P(X > 10)$
- Para $k = 0,1$ calcular $P(50 < X \leq 100)$

64. La función de distribución asociada a la producción de una máquina, en miles de unidades, es del tipo:

$$F(x) = \begin{cases} 0 & x < 0 \\ x(2-x) & 0 \leq x \leq k \\ 1 & x > k \end{cases}$$

- Determinar k para que sea función de distribución
- Hallar la función de densidad
- Calcular la media, mediana, moda y varianza de la producción
- Hallar $P(X < 0,5)$ y $P(X > 0,25)$

65. Dada la variable aleatoria continua X con función de densidad:

$$f(x) = \begin{cases} k(x+2) & 0 \leq x \leq 4 \\ 0 & \text{en el resto} \end{cases}$$

Hallar:

- El valor de k para que sea realmente una función de densidad
- La función de distribución
- La varianza
- $P(2 \leq X \leq 3)$

66. Dada la función $f(x) = e^{-2x}$

- Comprobar si puede ser función de densidad de una variable aleatoria X cuando su campo de variación es el intervalo $x \geq 0$
- En caso de que no lo pueda ser, qué modificaciones habría que introducir para que lo fuera.

67. Sea X una variable aleatoria continua con función de densidad:

$$f(x) = \begin{cases} \frac{8}{7x^2} & 1 \leq x \leq 8 \\ 0 & \text{otro caso} \end{cases}$$

- Calcular el primer y tercer cuartil, el decil 7 y el percentil 85
- Calcular la mediana y moda

68. La demanda diaria de un determinado artículo es una variable aleatoria con función de densidad:

$$f(x) = \begin{cases} \frac{1}{8} & 0 < x \leq 4 \\ \frac{12-x}{64} & 4 < x \leq 12 \\ 0 & \text{otro caso} \end{cases}$$

Los beneficios diarios dependen de la demanda según la siguiente función:

$$B = \begin{cases} -5 & \text{si } x < 2 \\ 5 & \text{si } 2 < x \leq 4 \\ 10 & \text{si } 4 < x \leq 8 \\ 15 & \text{si } 8 < x \leq 12 \end{cases}$$

Calcular:

- Probabilidad de que en un día cualquiera la demanda sea superior a 10
- Probabilidad de que la demanda sea inferior a 3
- La esperanza y la varianza de la demanda
- Función de distribución de la demanda
- Función de cuantía y función de distribución de la variable aleatoria beneficios diarios
- Esperanza y varianza de la variable beneficios

69. Se ha verificado que la variable $X =$ "Peso en kilos de los niños al nacer" es una variable aleatoria continua con función de densidad

$$f(x) = \begin{cases} kx & 2 \leq x \leq 4 \\ 0 & \text{en otros casos} \end{cases}$$

Se pide:

- Hallar k para que $f(x)$ sea función de densidad. Representarla
- Hallar la función de distribución. Representarla
- Media, varianza y desviación típica
- Probabilidad de que un niño elegido al azar pese más de 3 kilos
- Probabilidad de que pese entre 2 y 3,5 kilos
- Qué debe pesar un niño para tener un peso igual o inferior al 90% de los niños

70. La función de densidad de una variable aleatoria es:

$$f(x) = \begin{cases} ax^2 + b & 0 < x < 2 \\ 0 & \text{en el resto} \end{cases} \quad \text{Sabido que } P\left[\frac{1}{2} < x < 1\right] = 0,1666$$

Determinar a y b .

71. Sea X una variable aleatoria continua, cuya función de densidad es

$$f_x(x) = \begin{cases} 3x^2 & 0 < x < 1 \\ 0 & \text{en otro caso} \end{cases}$$

Sea $Y = 1 - X^2$ una transformación de la variable aleatoria X . Se pide:

- Calcular la función de densidad de la variable aleatoria Y
- Calcular la función de distribución de la variable aleatoria Y

72. Sea X una variable aleatoria continua, cuya función de densidad es

$$f_x(x) = \begin{cases} \frac{1}{2} & -1 < x < 1 \\ 0 & \text{en otro caso} \end{cases}$$

Sea $Y = X^2$ una transformación de la variable aleatoria X . Se pide:

- Calcular la función de densidad de la variable aleatoria Y
- Calcular la función de distribución de la variable aleatoria Y

73. Sea X una variable aleatoria continua con función de densidad tal que

$$f(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & \text{otro caso} \end{cases} \quad \text{Se pide:}$$

- Función generatriz de los momentos (f.g.m.)
- Esperanza y varianza a partir de la f.g.m.
- Función característica

74. El tiempo de revisión del motor de un avión sigue aproximadamente una distribución exponencial, con media 22 minutos.

- Hallar la probabilidad de que el tiempo de la revisión sea menor de 10 minutos.
- El costo de la revisión es de 200 euros por cada media hora o fracción. ¿Cuál es la probabilidad de que una revisión cueste 400 euros?
- Para efectuar una programación sobre las revisiones del motor, ¿cuánto tiempo se debe asignar a cada revisión para que la probabilidad de que cualquier tiempo de revisión mayor que el tiempo asignado sea solo de 0,1?

75. Una variable aleatoria X se distribuye uniformemente en $(2, 4)$. Se pide:

- $P(X < 2,5)$
- $P(X > 3,2)$
- $P(2,2 < X < 3,5)$
- Esperanza y varianza

76. Una empresa produce un artículo que sigue una distribución uniforme entre 25000 y 30000 unidades. Sabiendo que vende cada unidad a 10 euros y la función de costes viene dada por $C = 100000 + 2X$, ¿cuál será el beneficio esperado?

77. La duración de vida de una pieza de un motor sigue una distribución exponencial, sabiendo que la probabilidad de que sobrepase las 100 horas de uso es 0,9. Se pide:

- Probabilidad de que sobrepase las 200 horas de uso.
- ¿Cuántas horas se mantiene funcionando con probabilidad 0,95?

78. La cabina de un avión tiene 30 dispositivos electrónicos $(E_1, E_2, \dots, E_{30})$, tan pronto como falla E_1 se activa E_2 , y así sucesivamente. El tiempo de fallo de cualquier dispositivo electrónico E_i es de tipo exponencial con parámetro $\lambda = 0,1$ hora.

Hallar la probabilidad de que el tiempo total de funcionamiento de todos los dispositivos supere las 350 horas.

79. Un corredor de bolsa adquiere 50 acciones diferentes, concertando con sus clientes una ganancia de 1200 euros por acción. Por experiencias anteriores, se sabe que los beneficios de cada acción son independientes y se distribuyen uniformemente en el intervalo $[1000, 2000]$. ¿Qué probabilidad tiene el corredor de no perder dinero?.

80. El número promedio de recepción de solicitudes en una ventanilla de atención al cliente es de tres al día.

- ¿Cuál es la probabilidad de que el tiempo antes de recibir una solicitud exceda cinco días?
- ¿Cuál es la probabilidad de que el tiempo antes de recibir una solicitud sea menor de diez días?
- ¿Cuál es la probabilidad de que el tiempo antes de recibir una solicitud sea menor de diez días, si ya han pasado cinco días sin recibir solicitudes?

81. Una web tiene un número promedio de 5 visitantes por hora. Se piden las probabilidades:

- Sea visitada por un mínimo de 6 personas.
- Qué pase más de una hora sin que sea visitada.

82. Sea X la variable aleatoria que describe el número de clientes que llega a un supermercado durante un día (24 horas). Sabiendo que la probabilidad de que llegue un cliente en un día es equivalente a 100 veces la que no llegue ningún cliente en un día, se pide:

- Probabilidad de que lleguen al menos 3 clientes al día.
- Si acaba de llegar un cliente, calcular la probabilidad que pase más de 25 minutos hasta que llegue el siguiente cliente (o hasta que llegue el primer cliente).
- En dos semanas, ¿cuál es la probabilidad aproximada de que lleguen como mucho 1300 clientes al supermercado?

83. La duración de vida de una pieza de un motor sigue una distribución exponencial, sabiendo que la probabilidad de que sobrepase las 100 horas de uso es 0,9. Se pide:

- Probabilidad de que sobrepase las 200 horas de uso.
- ¿Cuántas horas se mantiene funcionando con probabilidad 0,95?

84. En una tela las fallas siguen un proceso de Poisson a razón de 1 falla cada 10 metros. ¿Cuál es la probabilidad de que la distancia entre la cuarta y quinta falla sea mayor a un metro?

85. La ciudad de Segovia se comunica con UAM mediante 3 autobuses por hora, distribuidos según un proceso de Poisson.

¿Cuál es la probabilidad de tener que esperar un autobús más de 20 minutos?

86. El promedio de clientes que llegan a solicitar servicio a un banco es de cinco por minuto.

¿Cuál es la probabilidad de que dos clientes tarden entre 30 a 45 segundos en solicitar servicio?

87. El número promedio de recepción de solicitudes en una ventanilla de atención al cliente es de tres al día.

a) ¿Cuál es la probabilidad de que el tiempo antes de recibir una solicitud exceda cinco días?

b) ¿Cuál es la probabilidad de que el tiempo antes de recibir una solicitud sea menor de diez días?

c) ¿Cuál es la probabilidad de que el tiempo antes de recibir una solicitud sea menor de diez días, si ya han pasado cinco días sin recibir solicitudes?

88. Sea X la variable aleatoria que describe el número de clientes que llega a un supermercado durante un día (24 horas). Sabiendo que la probabilidad de que llegue un cliente en un día es equivalente a 100 veces la que no llegue ningún cliente en un día, se pide:

a) Probabilidad de que lleguen al menos 3 clientes al día.

b) Si acaba de llegar un cliente, calcular la probabilidad que pase más de 25 minutos hasta que llegue el siguiente cliente (o hasta que llegue el primer cliente).

c) En dos semanas, ¿cuál es la probabilidad aproximada de que lleguen como mucho 1300 clientes al supermercado?

89. El tiempo de revisión del motor de un avión sigue aproximadamente una distribución exponencial, con media 22 minutos.

a) Hallar la probabilidad de que el tiempo de la revisión sea menor de 10 minutos.

b) El costo de la revisión es de 200 euros por cada media hora o fracción. ¿Cuál es la probabilidad de que una revisión cueste 400 euros?

c) Para efectuar una programación sobre las revisiones del motor, ¿cuánto tiempo se debe asignar a cada revisión para que la probabilidad de que cualquier tiempo de revisión mayor que el tiempo asignado sea solo de 0,1?

90. Un sistema esta formado con dos procesadores, los procesadores 1 y 2 funcionan correctamente durante un tiempo exponencial de media de 5 y 6 años, respectivamente.

Se pide:

a) ¿Cuál es la probabilidad de que ambos procesadores funcionen correctamente más de 4 años?

b) ¿Cuál es la probabilidad de que el procesador 1 deje de funcionar correctamente antes que el procesador 2?

91. La variable aleatoria poblacional "renta de las familias" del municipio de Madrid se distribuye siguiendo un modelo $N(\mu, \sigma)$. Se extraen muestras aleatorias simples de tamaño 4. Como estimadores del parámetro μ , se proponen los siguientes:

$$\hat{\mu}_1 = \frac{x_1 + 2x_2 + 3x_3}{6} \quad \hat{\mu}_2 = \frac{x_3 - 4x_2}{-3} \quad \hat{\mu}_3 = \bar{x}$$

Se pide:

- Comprobar si los estimadores son insesgados
- ¿Cuál es el estimador más eficiente?
- Si tuviera que escoger entre ellos, ¿cuál escogería?. Razone su respuesta a partir del Error Cuadrático Medio.

92. El peso en kilos de los jamones vendidos por una empresa sigue una distribución normal con varianza 4 y peso medio desconocido. Se conoce que el peso medio de los jamones vendidos es superior a 5 kg, y se toman m.a.s. de tamaño 4 para estimar θ . ¿Cuál de los dos estimadores sería el mejor respondiendo a la insesgadez y eficiencia?

$$\hat{\theta}_1 = \frac{x_1 + x_2 + x_3}{4} \quad \hat{\theta}_2 = \frac{x_1 + x_2}{2}$$

93. La distribución del peso de las manzanas de una determinada cosecha sigue una distribución normal, cuyo peso medio es desconocido y cuya desviación típica es 7 gramos.

Se pide:

- Analizar cuál de los estimadores $\hat{\mu}_1$ y $\hat{\mu}_2$ del peso medio es mejor respecto del sesgo y de la eficiencia, para una muestra aleatoria simple de tamaño cinco.

b) Si $\hat{\mu}_1 = \frac{\sum_{i=1}^5 x_i}{5}$ y $\hat{\mu}_2 = x_1 + 2x_2 + 3x_3 - 4x_4 - x_5$, obtener los pesos medios estimados a partir de la siguiente muestra (125, 135, 130, 137, 142).

94. Sea x_1, x_2, x_3, x_4, x_5 una muestra aleatoria simple de una variable aleatoria X con distribución normal con media $(\mu - 5)$ y varianza σ^2 .

Se proponen los siguientes estimadores: $\hat{\mu}_1 = \sum_{i=1}^5 x_i$ $\hat{\mu}_2 = 8x_2 - 3x_5$

Determinar cuál es el mejor estimador para μ . Justificar la respuesta.

95. Sea una población con media μ de la que se extraen m.a.s. de tamaño n . Considere los

siguientes estimadores de la media: $\hat{\mu}_1 = \bar{x}$ $\hat{\mu}_2 = \frac{1}{n+1} \sum_{i=1}^n x_i$

- Estudie la insesgadez, la eficiencia relativa y la consistencia de ambos estimadores
- Elija uno de los dos en término del error cuadrático medio

96. Supongamos que la distribución de ingresos de una cierta población es una variable aleatoria con media μ desconocida y varianza σ^2 también desconocida.

Si se desea estimar el ingreso medio de la población mediante una m.a.s. (muestra aleatoria simple) de tamaño n , respecto de la insesgadez y de la eficiencia.

¿Cuál de los dos estimadores elegiríamos?. ¿Son consistentes?

$$\hat{\mu}_1 = \frac{\sum_{i=1}^n x_i}{n-1} \quad \hat{\mu}_2 = \frac{\sum_{i=1}^n x_i}{n}$$

97. La variable aleatoria X representa los gastos mensuales de una empresa, cuya función de densidad es $f(\theta, x) = \theta x^{\theta-1}$ con $\theta > 0$ y $0 < x < 1$. Se realiza una m.a.s. de tamaño 3, y se proponen tres estimadores:

$$\hat{\theta}_1 = \bar{x} \quad \hat{\theta}_2 = \frac{x_1^2 + 2x_2^2 + 3x_3^2}{6} \quad \hat{\theta}_3 = \frac{x_3 - 2x_1 + 4x_2}{6}$$

- Calcule los sesgos.
- Si la muestra que se obtiene es $(0,7, 0,1, 0,3)$, calcule las estimaciones puntuales.
- ¿Cuáles son las funciones estimadas para las estimaciones anteriores?

98. Un atleta olímpico de salto de altura se enfrenta a un listón de 2,3 metros. Su entrenador desea estudiar el comportamiento del saltador. Sabe que el número de saltos fallidos por hora es una variable aleatoria distribuida como una Poisson de parámetro λ . Se pide:

- Calcular el estimador máximo verosímil del parámetro λ
- Analizar las propiedades de Insesgadez, Consistencia y Eficiencia
- A partir de la eficiencia del estimador elaborar un intervalo de confianza para el parámetro poblacional λ

99. Una muestra aleatoria $(x_1, \dots, x_n)$ de la población tiene como función de densidad

$$f_{\theta}(x) = \begin{cases} \theta x^{\theta-1} & \text{si } x \in (0,1) \\ 0 & \text{en el resto} \end{cases} \quad \theta > 0$$

- Hallar un estadístico suficiente
- Estimador de máxima verosimilitud de θ

100. En una distribución $N(\mu, \sigma)$ se estima la media poblacional μ en una muestra aleatoria simple de tamaño n $(x_1, x_2, \dots, x_n)$ por medio de la función muestral:

$$\hat{\mu} = \frac{\sum_{i=1}^n i x_i}{n}$$

Estúdiese la eficiencia del estimador.

101. Una muestra aleatoria $(x_1, \dots, x_n)$ de la población tiene como función de densidad

$$f_{\theta}(x) = \frac{x}{\theta^2} e^{-x/\theta} \quad \text{siendo } x > 0$$

- Hallar un estadístico suficiente
- Estimador de máxima verosimilitud de θ

102. En un estacionamiento el número de veces que se abre la barrera en un intervalo de 10 minutos, para que pasen vehículos en un sector de seguridad, se considera una variable aleatoria con distribución de Poisson de parámetro λ desconocido.

a) En una muestra aleatoria de 8 intervalos de 10 minutos cada uno, elegidos de forma independiente, se registra para cada intervalo el valor que toma la variable en estudio.

3 5 8 7 4 5 6 2

Encontrar la estimación máximo verosímil de λ

b) Sea $(x_1, \dots, x_n)$ una muestra aleatoria de tamaño n que sigue una distribución de Poisson.

Sí $\hat{\lambda}_1 = \sum_{i=1}^n \frac{x_i}{n}$, $\hat{\lambda}_2 = \frac{x_1 + 3x_n}{4}$ son estimadores. Determinar el mejor estimador del parámetro λ

103. Sea $(x_1, x_2, \dots, x_n)$ una muestra aleatoria de tamaño n , distribuida según $f(x, \theta)$ con θ desconocido, donde X representa el tiempo máximo necesario para determinar un proceso en

$$\text{segundos: } f(X, \theta) = \begin{cases} (\theta + 1)x^{\theta} & 0 \leq x \leq 1, \theta > -1 \\ 0 & \text{otro caso} \end{cases}$$

- Determinar el estimador máximo verosímil de θ
- Determinar la estimación máximo verosímil de θ en una muestra aleatoria simple constituida por los datos: 0,7 0,9 0,6 0,8 0,9 0,7 0,9.

Estimar la probabilidad del tiempo máximo necesario para terminar un proceso, que no exceda de 0,25 segundos ni supere los 0,75 segundos.

- Determinar el estimador máximo verosímil de $\theta + 1$ y $\frac{2\theta + 1}{\theta - 1}$

104. En una distribución $N(\mu, \sigma)$, se estima la media poblacional μ mediante la media de una muestra aleatoria simple $(x_1, x_2, \dots, x_n)$. El estimador es insesgado y su varianza $V(\bar{x}) = \frac{\sigma^2}{n}$.

Demostrar que la media muestral es un estimador eficiente.

105. Sea la distribución $N(\mu, \sigma)$, con media μ conocida y varianza desconocida.

Calcular la estimación máximo-verosímil de la varianza en muestras aleatorias simples de tamaño n .

106. Calcular el estimador máximo verosímil del parámetro 'a' de las siguientes funciones:

a) $f(x; a) = a^2 e^{-ax}$ siendo $x \geq 0$ en muestras aleatorias simples de tamaño n .

b) $f(x; a) = ae^{-ax}$ para $x \geq 0$, $a > 0$ en muestras aleatorias simples de tamaño 2.

107. En la distribución $B(m, p)$ se considera como estimador del parámetro p el estadístico

$\hat{p} = \frac{\bar{x}}{n}$, siendo $\bar{x}$ la media muestral en muestras aleatorias simples de tamaño n . Hallar la eficiencia del estimador.

108. Una urna contiene bolas blancas y negras. Sea p la probabilidad de extraer una bola blanca cuando se realiza una extracción al azar. Asociado a este experimento aleatorio se encuentra la variable aleatoria X que puede tomar los valores:

$X = 1$ si la bola extraída es blanca $X = 0$ si la bola extraída es negra

Se selecciona una muestra aleatoria con reemplazamiento de tamaño 3 (x_1, x_2, x_3) , siendo X_i la variable aleatoria a la extracción i -ésima, y se supone que ha resultado la siguiente relación (B, N, B) . Como el parámetro p es desconocido se pretende saber entre los valores $p = 0,65$ y $p = 0,73$ qué valor hace más probable la aparición de dicha extracción.

109. A partir de una muestra aleatoria simple X de tamaño n , determinar la matriz de información de Fisher de una población con distribución $N(\mu, \sigma)$, respecto a los dos parámetros μ y σ .

Como consecuencia, determinar la distribución asintótica del Valor en Riesgo (Value at Risk) de una empresa que cotiza en Bolsa y considera como pérdidas los rendimientos inferiores a 3 euros por acción.

110. Sea una población definida por: $P(X = -1) = \frac{1-\theta}{2}$, $P(X = 0) = \frac{\theta+\lambda}{2}$, $P(X = 1) = \frac{1-\lambda}{2}$,

donde $0 < \theta < 1$, $0 < \lambda < 1$. Estimar los parámetros θ y λ por el método de los momentos, estudiando si son insesgados.

111. Una muestra aleatoria $(x_1, \dots, x_n)$ de la población tiene como función de densidad

$$f_{\theta}(x) = \begin{cases} e^{-x+\theta} & \text{si } x > 0 \\ 0 & \text{en el resto} \end{cases} \quad \text{Se pide:}$$

- Hallar un estimador por el método de los momentos de θ .
- Estudiar si el estimador encontrado en el apartado anterior es insesgado para estimar el parámetro θ .

112. El coseno X del ángulo con el que se emiten los electrones en un proceso radioactivo es

una variable aleatoria con función de densidad $f_{\theta}(x) = \begin{cases} (1+\theta x)/2 & -1 \leq x \leq 1 \\ 0 & \text{en el resto} \end{cases} \quad -1 \leq \theta \leq 1$

Considerando una muestra aleatoria $(x_1, \dots, x_n)$ de esta variable aleatoria, se pide:

- Obtener el estimador θ por el método de los momentos.
- Calcular la varianza de este estimador y demostrar que es consistente.

113. Sea una variable aleatoria X procedente de una población con densidad de probabilidad $N(\mu, 5)$. Efectuadas dos hipótesis sobre el valor de μ : $H_0: \mu = 12$ $H_1: \mu = 15$

Mediante un muestreo aleatorio simple de tamaño 25, se contrasta la hipótesis H_0 respecto de la hipótesis H_1 , estableciéndose que si la media muestral es menor que 14 se aceptaría la hipótesis nula. Determinar:

- La probabilidad de cometer el error tipo I
- La probabilidad de cometer el error tipo II
- La potencia del contraste

114. Las latas de mejillones de una determinada marca indican que el peso escurrido de dicho producto es de 250 gramos. No obstante, un consumidor está convencido de que el peso escurrido medio de dicho producto es menor que el que indican las latas. Si el peso escurrido sigue una ley normal con desviación típica 9 gramos. Se pide:

- Determinar, si existe, la mejor región crítica para contrastar, con un nivel de significación del 5% y muestras aleatorias simples de tamaño 100.
- Tomar una decisión acerca del rechazo o no de la hipótesis nula a partir de una muestra aleatoria simple de tamaño 100 en la cual se ha observado un peso escurrido promedio de 245 gramos.
- Determinar la función de potencia del contraste.

115. Sea una variable aleatoria X procedente de una población con densidad de probabilidad $N(\mu, 4)$. Se quiere contrastar la hipótesis nula $H_0: \mu = 10$ frente a la hipótesis alternativa $H_1: \mu = 12$, con un nivel de significación $\alpha = 0,05$, con un muestreo simple de tamaño 25.

Determinar:

- La probabilidad de cometer el error tipo II
- La potencia del contraste

116. Un agricultor sabe que el peso en kg. De las patatas sigue una distribución $N(\mu, 1)$. Una muestra de patatas dio un peso medio de 330 gramos. Con la muestra se realizó un contraste, con un nivel de significación del 5% y una potencia de 0,6406, en el que la hipótesis nula era $\mu = 0,4$ Kg y la alternativa $\mu = 0,3$ Kg. Se pide:

- ¿Cuál es el tamaño de la muestra utilizada por el agricultor?
- Qué hipótesis fue aceptada

117. Un laboratorio farmacéutico quiere lanzar un nuevo medicamento para la hipertensión, llamado Hipotensil. El director de dicho laboratorio cree que la eficacia del medicamento sería de un 95%, medida ésta como la proporción de pacientes a los que se les suministra y que experimentan una mejoría. Sin embargo, el inspector de sanidad del Ministerio no es tan optimista y opina que la eficacia es sólo del 85%. Para analizar la eficacia del medicamento antes de su comercialización, se selecciona una muestra aleatoria de 500 pacientes, a los que se les administra Hipotensil, de los cuales mejoran 467. ¿Tiene razón el director del laboratorio?. *Suponga un nivel de significación del 5%.*

118. Un portal e-business sabe que el 60% de todos sus visitantes a la web están interesados en adquirir sus productos pero no reacios al comercio electrónico y no realizan finalmente la compra vía internet. Sin embargo, en la dirección del portal se piensa que en el último año, el porcentaje de gente que está dispuesta a comprar por internet ha aumentado y esto se debe reflejar en sus resultados empresariales.

- Contrastar con un nivel de significación del 2%, si en el último año se ha reducido el porcentaje de gente que no está dispuesta a comprar por internet, si para ello se tomó una muestra de 500 visitantes para conocer su opinión y se observó que el 55% no estaba dispuesto a realizar compras vía on-line.
- Realizar el contraste con el p_valor

119. Se trata de determinar si en una ciudad el 20% o el 30% de las familias dispone de lavavajillas; para dilucidarlo se toma al azar una muestra de 400 familias de la mencionada ciudad y se adopta el criterio de si en la muestra hay menos de 100 familias con lavavajillas, se rechaza que el 20% de las familias poseen el mencionado electrodoméstico. Se pide:

- Nivel de significación del test.
- Potencia del test.

120. Las especificaciones de un tipo de báscula aseguran que los errores de los pesajes siguen una distribución $N(0, \sigma)$. Se quiere contrastar la afirmación sobre la dispersión que es igual a la unidad, frente a una hipótesis alternativa de que es el doble. Para ello se realizan 5 pesajes en las que el error cometido resultó ser: 1 0,9 -0,2 1,4 -0,7

Para un nivel de significación del 5% se pide enunciar una regla de decisión (obtener la región crítica) e indicar que hipótesis resulta aceptada.

121. En una población $N(5, \sigma)$ se quiere contrastar la hipótesis nula $H_0: \sigma^2 = 2$ frente a la hipótesis alternativa $H_1: \sigma^2 = 3$, con un nivel de significación $\alpha = 0,025$, con una muestra aleatoria simple de tamaño 10:

5,1 6,2 4 2,8 2,9 5,6 3,7 3,4 2,5 5,2

Hallar la potencia del contraste.

122. En una población con distribución $N(\mu, \sigma)$ se desea contrastar la hipótesis nula $H_0: \sigma^2 = 25$ frente a la hipótesis alternativa $H_1: \sigma^2 = 36$, se toma una muestra aleatoria de tamaño 16, con varianza igual a 27.

- Tiene sentido utilizar el nivel de significación del 1%
- En caso negativo, utiliza el nivel de confianza más grande posible
- Hallar la potencia del contraste

123. El propietario de un automóvil sospecha que su vehículo según una ley normal tiene un consumo medio de combustible en carretera superior a los 5,6 litros /100 km., que es lo que el fabricante indica en su publicidad. Para apoyar empíricamente su sospecha observa el consumo medio en 11 viajes seleccionados aleatoriamente entre todos los que realiza en el año, obteniendo los siguientes resultados: 6,1 6,5 5,1 6 5,9 5,2 5,8 5,3 6,2 5,9 6,3
Se pide:

- ¿Están fundadas las sospechas del propietario a un nivel de significación del 1%?
- Calcula el p_valor
- ¿En cuántas ocasiones debería observarse el consumo medio para que con un nivel de confianza del 99% se detectase un consumo medio de 5,9 litros/100 km?

124. El directorio de uno de los grandes operadores de Internet está considerando la posibilidad de ofrecer tarifa plana a sus clientes sabiendo que siguen una ley normal. Según sus conocimientos sobre el tema, sabe que está trabajando con una variable aleatoria que se distribuye como una normal. Mantiene la hipótesis de que los hogares que tienen Internet se conectan con una media de 5 horas mensuales. No obstante, existen otros estudios que sostienen que el tiempo de conexión es más alto.

Para evaluar, a un 10% de significación, dicha hipótesis, el directorio decide encuestar a una muestra aleatoria de 30 hogares, obteniendo una media de 5,34 horas de conexión, con una dispersión de 7,24 horas.

- Formular el contraste a realizar
- Determinar la mejor región crítica del contraste
- ¿Se puede rechazar la hipótesis nula?
- Calcular y representar la función de potencia. ¿Qué representa la función de potencia?. Para facilitar los cálculos, suponer que el tamaño muestral es 300 hogares.

125. El número de averías de un determinado tipo de avión se considera una variable aleatoria con distribución de Poisson de media 2 averías al mes. El equipo de mantenimiento intenta reducir esta media incorporando algunas mejoras. Para comprobar si con estas medidas se reduce el número medio de averías, se decide observar el número medio de averías en los 25 meses siguientes a la introducción de las mejoras. Si el número medio de averías en esos 25 meses fue de 1,5.

¿Qué decisión debe adoptar el servicio técnico a un nivel de significación del 1%?

¿Y si el servicio técnico relaja su nivel de exigencia al 85% de confianza?. ¿Cambiaría su decisión?.

126. En los días previos a unas elecciones municipales, el candidato de un partido político está convencido de obtener el 60% de los votos electorales. No obstante, su partido encarga una encuesta entre 100 votantes potenciales, resultando que el 52% de ellos dijeron tener intención de votar a dicho candidato. Con un nivel de significación del 5%, se pide contrastar:

- a) $H_0: p = 0,60$ frente a $H_1: p = 0,50$
- b) $H_0: p = 0,60$ frente a $H_1: p \neq 0,60$
- c) Potencia del contraste efectuado en el apartado (a)

127. El dueño de los cines CINEFILÓN considera que, dado el aforo de la sala, una afluencia diaria a la misma del 85% sería óptima, en el sentido de que los clientes se sientan cómodos y para que a la vez no haya pérdidas económicas. Durante un período de tiempo, se analiza la afluencia a los cines, observándose que, en media, se ocupan 171 de las 200 butacas.

- a) ¿Con qué confianza podrá afirmar el dueño de CINEFILÓN que la asistencia a sus cines es óptima?
- b) ¿Qué pasaría si el dueño quisiera estar más seguro de su decisión, y ampliar el nivel de confianza al 99%?

Nota: La regla de la decisión adoptada es que si hay una desviación inferior al 5% de la cantidad óptima, se aceptaría la hipótesis de que la afluencia es, efectivamente, óptima.

128. En una población $N(10, \sigma)$ se desea contrastar la hipótesis nula $H_0: \sigma^2 = 3$ frente a la hipótesis alternativa $H_1: \sigma^2 \neq 3$ con un nivel de significación del 5%. En una muestra aleatoria simple de tamaño 4 se obtuvieron los resultados: 10 8 12 14

129. En una población $N(\mu, 5)$ se formula la hipótesis nula $H_0: \mu = 18$ frente a la hipótesis alternativa $H_1: \mu \neq 18$, con un nivel de significación $\alpha = 0,01$, con una muestra de tamaño 10 con los datos recogidos: 16 12 15 16 20 25 14 18 17 22

- a) Contrastar la hipótesis nula
- b) Realizar el contraste con el p_valor
- c) Realizar el contraste con métodos afines al p_valor

130. Una empresa de neumáticos afirma que una nueva gama sigue una ley normal y en promedio duran más de 28.000 km. Las pruebas con 64 neumáticos dan como resultado una duración media de 27.800 km, con una desviación estándar de 1.000 km.

- a) Si se usa un nivel de significación del 5%, comprobar si hay evidencia suficiente para rechazar la afirmación de la empresa.
b) ¿Cuál es el p_valor?

131. Una línea de producción sigue una distribución normal, funciona con un peso de llenado de 16 gramos por envase. El exceso o defecto de peso en el llenado presenta graves problemas, debiendo parar la línea de producción. Un inspector de calidad toma una muestra de 25 artículos, y de acuerdo con los resultados obtenidos, ¿qué decisión debe tomar?.

- a) Para un nivel de significación de 0,05, en la muestra se obtiene $\bar{x} = 16,32$ gramos y una desviación $s_x = 0,8$ gramos.
b) ¿Cuál es el p-valor?.

132. El análisis laboral que la U.E ha realizado para toda Europa, señala que en España, el salario mensual de los varones, en algunos sectores económicos, supera en más de 100 euros el salario de las mujeres que desempeñan las mismas tareas.

El Ministerio de Trabajo español decide considerar el salario mensual como una variable aleatoria normalmente distribuida con desviación típica de 39,6 euros para los trabajadores masculinos y de 36 euros para las trabajadoras de dichos sectores, siendo el salario de cada población independiente del de la otra. Para tratar de verificar lo publicado, se elige una muestra aleatoria simple de 500 trabajadores y de 700 trabajadoras, obteniéndose unos salarios medios mensuales de 1.500 y 1.370 euros respectivamente.

¿Están fundamentadas las conclusiones de la U.E al 1% de significación?

133. Con un nivel de significación del 4,72%, se desea contrastar la hipótesis nula de igualdad de medias de dos poblaciones $N(\mu_1, 4)$ y $N(\mu_2, 4,5)$. Para ello, se han tomado dos muestras aleatorias simples e independientes, respectivamente, obteniéndose los siguientes valores:

x_i	20,4	10,2	7,3	12,8	13,4	9,4
y_j	19,8	9,7	14,6	15,7	8,4	

134. La directora del departamento de personal de una corporación está buscando empleados para un puesto en el extranjero. Durante el proceso de selección, la administración le pregunta cómo va la incorporación de empleados, y ella contesta que sigue una ley normal y que la puntuación promedio en la prueba de aptitudes será de 90 puntos.

Cuando la administración revisa 19 de los resultados de la prueba, encuentra que la puntuación media es de 86,25 puntos con una desviación estándar de 11.

Con un nivel de confianza del 90%, ¿lleva razón la directora?. Calcula su p_valor.

135. Una empresa ubicada en Madrid tiene dos conductores para trasladar a los empleados a Segovia. Los conductores deben anotar la duración de cada trayecto. En una muestra aleatoria simple de 50 partes de incidencias por conductor, el conductor A registra un tiempo medio de trayecto de 62,30 minutos con una desviación típica de 10,325 minutos, mientras que el conductor B tiene un tiempo medio de trayecto de 60,02 minutos con una desviación típica de 8,625 minutos.

El tiempo medio empleado por el conductor A en el trayecto sigue una ley normal $N(\mu_1, 9)$ y el empleado por el conductor B se distribuye según una ley $N(\mu_2, 8)$.

Con un nivel de significación del 5%, se pide contrastar:

- a) $H_0: \mu_1 = \mu_2$ frente a $H_1: \mu_1 \neq \mu_2$
b) $H_0: \mu_1 - \mu_2 \geq 2$ frente a $H_1: \mu_1 - \mu_2 < 2$

136. Una compañía de refrescos presenta un nuevo producto en el mercado afirmando que posee menos calorías que su homólogo más antiguo y conserva el resto de propiedades. Para verificar la afirmación de la compañía se eligieron al azar 14 botes del refresco nuevo y se calculó su media, 20 calorías por bote, y su desviación estándar muestral, tres calorías. De modo independiente, se tomó otra muestra aleatoria de 16 botes del refresco antiguo, obteniéndose una media de 28 calorías por bote con una desviación estándar muestral de 5 calorías. Suponiendo que la cantidad de calorías por bote sigue una distribución normal en ambos refrescos, pero con desviaciones típicas diferentes, ¿existe alguna razón para no creer en la afirmación de la compañía con un nivel de significación del 2,5%?

137. El estacionamiento de corta estancia en el Aeropuerto Adolfo Suárez está cerca de la terminal, de modo que al recoger a un pasajero solo hay que caminar una corta distancia al área de recuperación de equipajes. Para decidir si el estacionamiento tiene suficientes lugares, el gerente del aeropuerto necesita saber si el tiempo medio de permanencia es superior a 40 minutos. Una muestra aleatoria de una población normal mostró que 12 clientes estuvieron en el estacionamiento los siguientes lapsos de tiempo, en minutos:

55 47 27 64 53 56
49 53 39 48 48 37

Con un nivel de significación 0,05, ¿se puede afirmar que el tiempo medio en el estacionamiento es superior a 40 minutos?

138. Muebles Quintana manifiesta que la satisfacción de sus empleados sigue una ley normal y es idéntica en las distintas secciones. El jefe de recursos humanos para decidir si la afirmación es correcta toma una muestra aleatoria de 100 trabajadores de barnizado encontrando 80 satisfechos y otra de 200 trabajadores de cortado encontrando 140 empleados satisfechos. Con un nivel de significación del 5% se puede admitir la afirmación de Muebles Quintana.

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

139. Se quiere comparar las puntuaciones en estadística teórica de dos Universidades que siguen, respectivamente, una distribución normal, con varianzas desconocidas y se supone que estadísticamente iguales. Se toman al azar dos muestras iguales de estudiantes de ambas Universidades. Los resultados obtenidos en SPSS son:

Prueba de muestras independientes

	Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
	F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ de la diferencia	95% Intervalo de confianza para la diferencia	
								Inferior	Superior
Calificación Se han asumido varianzas iguales	,152	,699	,443	28	,661	2,200	4,962	-7,964	12,364

Con un nivel de significación de 0,05, se pide:

- ¿Existe diferencia significativa sobre las puntuaciones de estadística teórica en las dos Universidades?
- ¿Cuál es la media ponderada de las cuasivarianzas muestrales?
- Cuál sería el p_valor si se hubieran contrastado las puntuaciones

$$H_0 : \mu_{\text{Autónoma}} \leq \mu_{\text{Complutense}} \quad H_1 : \mu_{\text{Autónoma}} > \mu_{\text{Complutense}}$$

140. El precio de los productos vendidos por una empresa es una variable aleatoria con función de densidad $f(x) = \theta x^{\theta-1}$, donde $0 < x < 1$, $\theta > 0$, que depende del parámetro desconocido θ .

Se quiere contrastar sobre el valor de dicho parámetro la hipótesis nula $H_0 : \theta = 1$ frente a la alternativa $H_1 : \theta = 2$. Para ello, se toma una muestra aleatoria simple de tamaño dos.

Determinar el nivel de significación y la potencia del contraste, si se toma como región crítica $x_1 x_2 \leq 0,6$

141. El volumen diario de ventas de una empresa, en cien mil euros, es una variable aleatoria X sobre cuya función de densidad se establece la hipótesis nula $H_0 : f(x) = x/2$, $0 < x < 2$, frente a la hipótesis alternativa $H_1 : f(x) = 1/2$, $0 < x < 2$. Para realizar el contraste se toma una muestra aleatoria simple de dos días en los que los volúmenes de venta fueron de 50.000 euros y 100.000 euros. Con un nivel de significación del 5 por ciento, se pide:

- Hallar la mejor región crítica.
- Calcular la potencia del contraste.
- Qué hipótesis se acepta.

142. El peso (en gramos) de las cajas de cereales de una determinada marca sigue una distribución $N(\mu, 5)$. Se han tomado los pesos de 16 cajas seleccionadas aleatoriamente, y los resultados obtenidos han sido:

506, 508, 499, 503, 504, 510, 497, 512, 514, 505, 493, 496, 506, 502, 509, 496.

- Obtener el intervalo de confianza del 95% para la media poblacional.
- Determinar cuál sería el tamaño muestral necesario para conseguir, con un 95% de confianza, un intervalo de longitud igual a 2 gramos.
- Suponiendo ahora que σ es desconocida, calcular el intervalo de confianza para la media al 99%.

143. El gerente de un centro comercial de Madrid desea estimar la cantidad media que gastan los clientes que visitan el centro comercial. Sabiendo que el gasto sigue una distribución normal, una muestra de 20 clientes señala las siguientes cantidades en euros:

37,9 46,9 41,8 46,8 49,2 52,7 50,8

52,6 61,8 23,7 42,2 61,5 58,8 54,8

48,5 61,9 51,4 48,1 51,3 43,9

Determinar un intervalo de confianza del 95% e interpretar el resultado. ¿Se podría afirmar que la media poblacional es de 50 o 60 euros?

144. Un fabricante de televisores está desarrollando un nuevo modelo de televisor en color, y para este fin se pueden utilizar dos tipos de esquemas transistorizados. El fabricante selecciona una muestra de esquemas transistorizados del primer tipo de tamaño 16, y otra del segundo tipo de tamaño 13. Los datos muestrales respecto a la vida media de cada esquema son los siguientes:

$$\begin{cases} \bar{x}_1 = 1400 \text{ horas} & s_1 = 30 \text{ horas} & n_1 = 16 \\ \bar{x}_2 = 1500 \text{ horas} & s_2 = 17 \text{ horas} & n_2 = 13 \end{cases}$$

Construir un intervalo de confianza del 90% para la diferencia de vida media de cada tipo de esquema.

145. Un instituto de investigaciones agronómicas siembra, en cinco parcelas diferentes, dos tipos de maíz híbrido. Las producciones en quintales métricos por hectárea son:

Híbrido I	90	85	95	76	80
Híbrido II	84	87	90	92	90

Construir un intervalo de confianza del 90% para la diferencia entre las producciones medias.

146. Un equipo de investigación biológica está interesado en ver si una nueva droga reduce el colesterol en la sangre. Con tal fin toma una muestra de diez pacientes y determina el contenido de colesterol en la sangre antes y después del tratamiento. Los datos muestrales expresados en miligramos por 100 mililitros son los siguientes:

Antes	217	252	229	200	209	213	215	260	232	216
Después	209	241	230	208	206	211	209	228	224	203

Construir un intervalo de confianza del 95 por 100 para la diferencia del contenido medio de colesterol en la sangre antes y después del tratamiento.

147. Según los dirigentes del partido A, la intención de voto del partido rival B, en Andalucía, es la misma que la que tiene en Madrid. Se realiza una encuesta a 100 personas en Andalucía de los que 25 mostraron su apoyo al partido B, y a otras 100 personas en Madrid de las que 30 se inclinaron por el partido B.

- Construir un intervalo de confianza del 90% para la proporción de personas que votarían al partido B en Andalucía.
- ¿A cuántas personas habría que encuestar para obtener un margen de error o error de estimación $\pm 2\%$, al nivel de confianza anterior?.
- Construir un intervalo de confianza al 90% para la diferencia de proporciones en la estimación del voto del partido B en las dos comunidades. ¿Podemos afirmar que los dirigentes del partido A tienen razón?.

148. Las puntuaciones de estadística teórica se distribuyen según una ley normal, con μ y σ desconocidas. Se desea contrastar $H_0: \mu = 74$ frente a $H_1: \mu \neq 74$. Para ello, se introducen en SPSS las puntuaciones de quince alumnos escogidos aleatoriamente. Los resultados han sido:

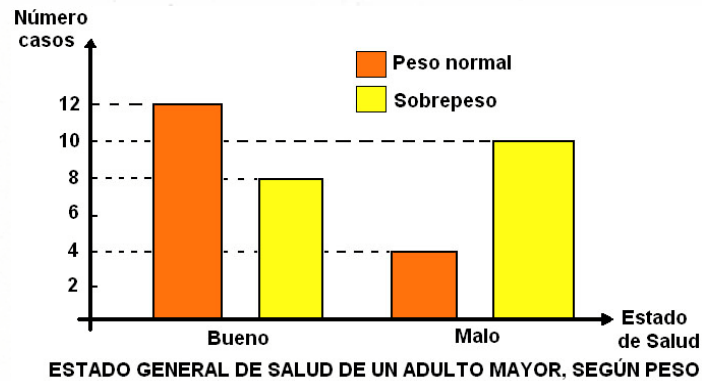
Prueba para una muestra

	Valor de prueba = 74					
	t	gl	Sig. (bilateral)	Diferencia de medias	95% Intervalo de confianza para la diferencia	
					Inferior	Superior
Autónoma	1,923	14	,075	7,200	-,83	15,23

Con un nivel de significación de 0,05, se pide:

- ¿Qué se puede decir del contraste?
- Media y cuasidesviación típica muestral
- Define el p-valor y como calcular la significación bilateral
- Calcula el p_valor al contrastar $H_0: \mu \leq 74$ frente a $H_1: \mu > 74$
- Calcula el p_valor al contrastar $H_0: \mu \geq 74$ frente a $H_1: \mu < 74$

149. En el gráfico se presenta la evaluación del estado general de salud de una muestra de personas adultas mayores, según sea su peso normal o sobrepeso.



Analizar la existencia de una relación significativa entre el peso y el estado general de salud en el adulto mayor, con un nivel de significación del 5%.

150. El número de defectos en las tarjetas de circuito impreso sigue una distribución Poisson. Se reúne una muestra aleatoria de 60 tarjetas de circuito impreso y se observa el número de defectos. Los resultados obtenidos son los siguientes:

Número de defectos	0	1	2	3 o más
Frecuencia observada	32	15	9	4

¿Muestran los datos suficiente evidencia para decir que provienen de una distribución de Poisson con una fiabilidad del 95%?

151. Para comprobar si los operarios encontraban dificultades con una prensa manual de imprimir, se hizo una prueba a cuatro operarios anotando el número de atascos sufridos al introducir el mismo número de hojas, dando lugar a la siguiente tabla:

Operario	A	B	C	D	Total
Obstrucciones	6	7	9	18	40

Con un nivel de significación del 5%, ¿existe diferencia entre los operarios?

152. La tabla refleja el número de accidentes mortales de tráfico que se producen en una carretera a lo largo de un período de tiempo.

Accidentes mortales por día	0	1	2	3	4	5
Número de días	132	195	120	60	24	9

¿Se ajustan los datos a una distribución de Poisson?. Utilizar un nivel de significación 0,05

153. Para curar cierta enfermedad se sabe que existen cuatro tratamientos diferentes. Aplicados por separado a un grupo distinto de enfermos, se han observado los siguientes resultados:

Enfermo Tratamientos	Curados	No curados	Total
A	60	23	83
B	46	10	56
C	70	17	87
D	54	30	84

¿Se puede considerar que la eficacia de los cuatro tratamientos es la misma con un nivel de confianza del 95 por 100?

154. Un experimento para investigar el efecto de vacunación de animales de laboratorio refleja la siguiente tabla:

Vacuna	Animal Laboratorio	
	Enfermo	No Enfermo
Vacunado	9	42
No Vacunado	18	28

Con un nivel de significación de 0,05, ¿Es conveniente vacunar?

155. Para analizar la repercusión que tienen los debates televisivos en la intención de voto, un equipo de investigación recogió datos entre 240 individuos antes y después del debate, resultando la siguiente tabla:

Antes del debate (candidatos)	Después del debate (candidatos)		Total
	A	B	
A	46	50	96
B	85	59	144
Total	131	109	240

Se desea saber si el debate televisivo cambió la intención de voto, con un nivel de significación del 5%.

156. Se desea analizar si los estudiantes de universidades privadas preferentemente son de los estratos económicos altos del país. Para ello, se ha tomado la siguiente muestra:

Universidades	Grupos socioeconómicos			
	Alto	Medio alto	Medio bajo	Bajo
Estado	13	17	4	3
Privadas	38	19	2	2

- a) Para validar el análisis con un nivel de confianza del 95%, realizar un contraste por la razón de verosimilitud (test G).
b) Estudiar el grado de dependencia entre el tipo de universidad y el estrato socioeconómico.

157. En un estudio sobre la opinión de fumar en lugares públicos se realiza una encuesta a 350 personas, obteniendo los siguientes resultados:

	Muy en contra	En contra	A Favor	Muy a favor	$n_{i.}$
Fumador	60	50	20	10	140
No Fumador	10	30	70	100	210
$n_{.j}$	70	80	90	110	350

Con un nivel de significación de 0,05 se desea conocer si existe diferencia de opinión entre fumadores y no fumadores.

158. Se ha medido el perímetro craneal a niños de edad comprendida entre los dos y tres años, obteniéndose los siguientes datos en centímetros:

Intervalos	38 – 40	40 – 42	42 – 44	44 – 46	46 – 48
Frecuencia observada	11	7	7	6	4
Frecuencia esperada	5,348		9,2855	5,9255	

Comprobar si las observaciones craneales son o no distribuidas según una ley normal con un nivel de confianza del 99%.

159. Para conocer la opinión de los ciudadanos sobre la actuación del alcalde de una determinada ciudad, se realiza una encuesta a 404 personas, cuyos resultados se recogen en la siguiente tabla:

	Desacuerdo	De acuerdo	No contestan
Mujeres	84	78	37
Varones	118	62	25

Contrastar, con un nivel de significación del 5%, que no existen diferencias de opinión entre hombres y mujeres ante la actuación del alcalde.

160. En dos ciudades, A y B, se observó el color del pelo y de los ojos de sus habitantes, encontrándose las siguientes tablas:

Ciudad A		
Pelo \ Ojos	Rubio	No Rubio
Azul	47	23
No azul	31	93

Ciudad B		
Pelo \ Ojos	Rubio	No Rubio
Azul	54	30
No azul	42	80

- a) Hallar los coeficientes de contingencia de las dos ciudades.
b) ¿En cuál de las dos ciudades se puede afirmar que hay mayor dependencia entre el color del pelo y de los ojos?

161. Para una muestra aleatoria simple de 350 días, el número de urgencias tratadas diariamente en un hospital queda reflejado en la siguiente tabla:

Número urgencias	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	Total días
Número de días	20	65	100	95	60	10	350

Contrastar con un nivel de significación del 5% si la distribución del número de urgencias tratadas diariamente en el hospital se ajusta a una distribución normal.

162. La tabla adjunta refleja un análisis de la obesidad en 14 sujetos. Con un nivel de significación de 0,05, se desea analizar si existen diferencias en la prevalencia de obesidad entre hombres y mujeres o si, por el contrario, el porcentaje de obesos no varía entre sexos.

Sexo	Obesidad		Total
	Sí	No	
Mujeres	1	4	5
Hombres	7	2	9
Total	8	6	14

163. En un estudio sobre la opinión de fumar en lugares públicos se realiza una encuesta a 350 personas, obteniendo los siguientes resultados:

	Opinión				$n_{i.}$
	Muy en contra	En contra	A Favor	Muy a favor	
Fumador	60	50	20	10	140
No Fumador	10	30	70	100	210
$n_{.j}$	70	80	90	110	350

Con un nivel de significación de 0,05 se pide:

- 1) ¿Existe diferencia de opinión entre fumadores y no fumadores?
- 2) ¿Cuál es el grado de dependencia en la opinión?
- 3) Coeficiente Phi y Coeficiente V de Cramer
- 4) Test G da la razón de verosimilitud
- 5) Coeficiente Lambda de Goodman y Kruskal
- 6) Coeficiente Tau de Goodman y Kruskal

164. En la tabla se refleja la edad de los empleados de una empresa y el grado de satisfacción en el trabajo, con un nivel de significación del 5%, se solicita:

Edad	Satisfacción en el trabajo				
	A	B	C	D	E
< 25	10	10	20	40	70
25 - 36	20	10	15	20	30
> 36	60	50	30	10	5

- a) Contrastar si el grado de satisfacción en el trabajo no depende de la edad de los empleados mediante el test de Chi-cuadrado y el test G de la razón de verosimilitud.
- b) Analizar el grado de asociación entre las variables con los estadísticos Phi, V de Cramer y Coeficiente de contingencia.
- c) Calcular la Gamma de Goodman y Kruskal , Coeficiente Tau-C , Tau-B de Kendall , Coeficiente D de Somers y el Coeficiente Lambda de Goodman y Kruskal.
- d) Calcular el Coeficiente Tau de Goodman y Kruskal y el Coeficiente de Incertidumbre.

165. Para analizar la repercusión que tienen los debates televisivos en la intención de voto, un equipo de investigación recogió datos entre 240 individuos antes y después del debate, resultando la siguiente tabla:

Antes del debate (candidatos)	Después del debate (candidatos)		Total
	A	B	
A	46	50	96
B	85	59	144
Total	131	109	240

¿El debate televisivo cambió la intención de voto, con un nivel de significación del 5%.

166. Un agricultor desea saber si existe diferencia entre diez abonos en el cultivo en una determinada zona. Para ello abona seis matas con cada abono, observa el número de kilos y anota el tipo de abono X_i y la varianza de cada abono σ_i^2 , obteniendo los siguientes resultados:

X_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
σ_i^2	9	3	4	3	5	4	2	4	5	3

¿Hay diferencia entre los abonos a un nivel de $\alpha = 0,01$? . ¿Y aun nivel de $\alpha = 0,05$?

167. Un inversor tiene acciones en diferentes sectores del mercado continuo de valores. Después de quedarse sin asesor, para tener idea sobre la cotización de sus acciones seleccionó nueve sociedades en las que participaba y revisó sus cotizaciones de cierre el pasado viernes:

Sociedades	Cotización compra	Cotización actual
Acerinox	10,07	9,92
Bankinter	7,70	7,85
Bankia	3,98	4,37
Caixabank	2,56	1,44
Mapfre	3,05	3,05
BBVA	4,41	3,32
IAG	6,73	6,46
Telefónica	7,85	8,09
Mercadona	2,95	2,95

A un nivel de significación del 5%, ¿puede afirmarse que no hay diferencia significativa entre el valor de las acciones en las sociedades en las que participa este inversor?

Nota: Utilizar la prueba de rangos con signos de Wilcoxon

168. Se emplean dos estrategias de ventas (A y B) de un mismo producto en veinte ciudades. El volumen de ventas anuales del mismo producto en cada ciudad (en miles de euros) por tipo de estrategia, está reflejado en la tabla adjunta. ¿Existe alguna diferencia significativa entre las estrategias de ventas a un nivel de significación del 5%?

Ventas (miles euros)			Ventas (miles euros)		
Ciudad	A	B	Ciudad	A	B
1	12,36	11,32	11	10,15	10,03
2	15,65	10,21	12	13,15	13,02
3	9,83	9,79	13	14,65	12,05
4	18,37	19,21	14	11,47	12,62
5	11,72	11,43	15	13,15	12,96
6	13,04	12,96	16	15,67	14,97
7	10,15	9,81	17	13,12	13,65
8	15,41	14,65	18	10,17	9,87
9	11,07	11,79	19	14,78	14,17
10	12,31	11,05	20	9,63	10,17

169. Una empresa inicialmente establece una jerarquía entre diez agentes vendedores de acuerdo a los resultados obtenidos en un test (el más idóneo es 1, el peor es 10). Un año después se replantea una nueva jerarquía con el mismo criterio de acuerdo al volumen de ventas anuales. Los resultados obtenidos se plasman en la tabla adjunta. ¿Existe alguna diferencia estadísticamente significativa entre ambos controles a un nivel de significación del 5%?

Jerarquía empresarial		
Vendedor	Test	Ventas
Alonso	4	1
Espinosa	5	6
Jiménez	2	2
Fernández	10	8
Crespo	7	9
García	1	3
Mateos	8	4
Cortes	3	5
Rodenas	9	10
Cabildo	7	7

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

170. De un cierto producto se tomaron 12 muestras lo más parecidas posibles y se dividieron en 5 familias con 4, 2, 2, 3, 1 muestras, que se almacenaron bajo diferentes condiciones.

Se trata de ver con los datos de hidratación de la tabla adjunta si con un nivel de confianza del 95% hay diferencias significativas entre los métodos de almacenaje.

Método	1	2	3	4	5
	8,3	7,4	8,1	7,9	7,1
	7,6	7,1	6,4	9,5	
	8,4			10,0	
	8,3				

171. Se someten 24 muestras de agua a 4 tratamientos de descontaminación diferentes y asignados al azar. Para cada muestra se mide un indicador de la calidad del agua (cuanto más alto está el indicador, mayor es la calidad del agua). Con un nivel de confianza del 95% ¿son los tratamientos igualmente efectivos?.

T1	T2	T3	T4
62	63	68	56
60	67	66	62
63	71	71	60
59	64	67	61
	65	68	63
	66	68	64
			63
			59

172. En un restaurante desean contratar a un camarero profesional cualificado. En una preselección se eligieron a cuatro personas de prueba. Para evaluar su rendimiento y el grado de satisfacción de los clientes se utilizó las cantidades entregadas como propinas. En la tabla adjunta se refleja el porcentaje de propina entregado sobre el importe total de la factura de los distintos clientes.

1	2	3	4
8	10	8,5	13
7,5	9,5	5	10,5
10,9	11,4	15	9,5
10	10,3	11,1	12
12	12	10,5	16
6,5		7	8
10		11	

Se supone que los porcentajes de propinas sobre el importe total de la factura se pueden considerar normalmente distribuidos. Con un nivel de significación del 10% se pide:

- Se puede aceptar la hipótesis de homocedasticidad de las distribuciones correspondientes al porcentaje de propina.
- Se pueden considerar los cuatro candidatos igualmente cualificados para el trabajo.

173. Un inversor tiene acciones en diferentes sectores del mercado continuo de valores. Después de quedarse sin asesor, para tener idea sobre la cotización de sus acciones seleccionó nueve sociedades en las que participaba y revisó sus cotizaciones de cierre el pasado viernes:

Sociedades	Cotización compra	Cotización actual
Acerinox	10,07	9,92
Bankinter	7,70	7,85
Bankia	3,98	4,37
Caixabank	2,56	1,44
Mapfre	3,05	3,05
BBVA	4,41	3,32
IAG	6,73	6,46
Telefónica	7,85	8,09
Mercadona	2,95	2,95

A un nivel de significación del 5%, ¿puede afirmarse que no hay diferencia significativa entre el valor de las acciones en las sociedades en las que participa este inversor?

Nota: Utilizar la prueba de rangos con signos de Wilcoxon

174. Se emplean dos estrategias de ventas (A y B) de un mismo producto en veinte ciudades. El volumen de ventas anuales del mismo producto en cada ciudad (en miles de euros) por tipo de estrategia, está reflejado en la tabla adjunta. ¿Existe alguna diferencia significativa entre las estrategias de ventas a un nivel de significación del 5%?

Ventas (miles euros)			Ventas (miles euros)		
Ciudad	A	B	Ciudad	A	B
1	12,36	11,32	11	10,15	10,03
2	15,65	10,21	12	13,15	13,02
3	9,83	9,79	13	14,65	12,05
4	18,37	19,21	14	11,47	12,62
5	11,72	11,43	15	13,15	12,96
6	13,04	12,96	16	15,67	14,97
7	10,15	9,81	17	13,12	13,65
8	15,41	14,65	18	10,17	9,87
9	11,07	11,79	19	14,78	14,17
10	12,31	11,05	20	9,63	10,17

175. Una empresa inicialmente establece una jerarquía entre diez agentes vendedores de acuerdo a los resultados obtenidos en un test (el más idóneo es 1, el peor es 10). Un año después se replantea una nueva jerarquía con el mismo criterio de acuerdo al volumen de ventas anuales. Los resultados obtenidos se plasman en la tabla adjunta. ¿Existe alguna diferencia estadísticamente significativa entre ambos controles a un nivel de significación del 5%?

Jerarquía empresarial		
Vendedor	Test	Ventas
Alonso	4	1
Espinosa	5	6
Jiménez	2	2
Fernández	10	8
Crespo	7	9
García	1	3
Mateos	8	4
Cortes	3	5
Rodenas	9	10
Cabildo	7	7

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

176. El director de una agencia de viajes, para diseñar un conjunto de estrategias publicitarias, está interesado en tener información sobre las edades de los compradores de un determinado paquete de viajes. Por este motivo, decide registrar la edad de las personas que compran este tipo de oferta de viajes, obteniendo la secuencia:

34 26 27 31 28 32 23 26 29 36
40 25 42 23 20 29 31 34 28 24
23 35 29 34 32 31 25 24 30 31
37 29 28 27

Con una significación del 5% se pide:

- Se admite la aleatoriedad en la serie de edades
- Observando los datos, ¿se puede afirmar que las personas que compran este paquete de viajes tienen al menos 30 años?

177. Un cuestionario sobre innovación empresarial, entre otros aspectos, recoge datos sobre la edad (x_i) y los años de experiencia profesional (y_i), obteniendo la siguiente tabla:

x_i	35	37	42	31	33	47	52	50	60	39	43	34	34	41	49	40	34	57	53	32	45
y_i	5	7	10	6	8	12	17	16	23	7	13	4	9	11	12	14	9	21	28	7	10

- Contrastar que la edad de innovadores es una muestra aleatoria simple de una población continua y simétrica, con mediana igual o mayor que 40 para un nivel de significación del 5%.
- Contrastar que la experiencia profesional es una muestra aleatoria simple de una población continua y simétrica, con mediana 16 para un nivel de significación del 5%.

178. En la tabla adjunta aparecen dos muestras aleatorias simples que reflejan los costes laborales unitarios de dos sectores de producción A y B::

SECTOR A	SECTOR B
1200	1263
1150	1370
1326	1614
1480	1416
1235	1139
1215	1041
1515	1493
1375	1817
1296	1388
1463	1110
1514	
1409	
1268	
1377	
1091	

Con un nivel de significación del 5%, utilizando la prueba de Ansari-Bradley, se pide:

- Contrastar que los costes unitarios de los dos sectores tienen la misma dispersión, suponiendo que las medianas de las dos poblaciones son iguales.
- Contrastar que los costes unitarios de los dos sectores tienen la misma dispersión, suponiendo que las medianas de las dos poblaciones son distintas.

179. En la tabla adjunta se presentan los gastos diarios medios de los turistas que visitan dos ciudades españolas A y B.

x_{iA}	300	351	326	482	532	227	211	300	326	469	614	129	265	377
x_{iB}	128	300	189	97	223	341	493	63						

Sabiendo que la mediana de la población A es 400 y la de B es 250, con un nivel de significación del 5%, contrasta que los gastos medios de los turistas con destino A tienen una dispersión igual o menor que los del destino B. Utilizar la prueba de Ansari-Bradley.

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

180. En tres estaciones del metro de Madrid se analiza la puntualidad del servicio. Para ello, se toma una muestra en cada una de ellas, midiendo el tiempo que se retrasa cada tren que llega. Se han obtenido los resultados siguientes:

A	30	25	29	32	33	20	19	5	18	23		
B	26	21	22	37	38	13	10	9	3	14	16	34
C	31	35	42	15	17	11	12	24				

Contrastar si los retrasos en las tres estaciones son semejantes, con un nivel de significación del 5%. Utilizar la prueba de Kruskal-Wallis.

181. Un fabricante de juguetes desea conocer si existen diferencias en cuanto a la calidad de cuatro marcas de pilas extendidas en el mercado, con el fin de recomendarlas para su utilización en un nuevo juguete que se va a promocionar. Para comparar las cuatro marcas, toma muestras aleatorias de pilas de cada una de las marcas y controla el tiempo que permanecen funcionando el juguete en cuestión. Los resultados obtenidos fueron:

A	87	95	100	79	85	98	100	84	
B	97	62	59	37	90	42	44	100	94
C	100	45	50	95	70	78	80	47	69
D	75	50	54	65					

Utilizando un nivel de significación del 5%, ¿puede decirse que existen diferencias significativas en las calidades de estas marcas de pilas alcalinas?.

Utilizar la prueba de Kruskal-Wallis.

182. En la siguiente tabla se adjuntan los datos sobre el ROE (ratio utilizado para realizar análisis mediante la medición de la rentabilidad obtenida por la entidad sobre sus recursos propio) y ROA (beneficio obtenido antes de intereses e impuestos - activos totales) de diez entidades financieras que cotizan en Bolsa.

Entidades	ROE	ROA
1 Endesa	13,68	1,80
2 Repsol	14,64	1,73
3 BBVA	20,28	1,66
4 Banesto	17,52	1,25
5 Carrefour	20,34	2,21
6 Ibercaja	15,12	2,28
7 Aceralia	18,6	1,75
8 Unicaja	18,72	2,83
9 BBK	19,32	2,62
10 Bankia	13,8	1,32

Contrastar que ambos conceptos no están relacionados empleando la τ de Kendall, con un nivel de significación del 5%.

183. La tabla adjunta presenta las cotizaciones máximas de las acciones de una compañía durante los últimos diez días del mes de octubre:

14,47 12,63 15,12 17,75 18,05 17,1 13,42 20,4 19,3 17,4

- a) Con un nivel de significación del 10%, ¿se puede decir que la muestra es aleatoria con respecto a la media de cotizaciones alcanzada en mes anterior que fue de 17,5 euros?
- b) Contrastar si las acciones se pueden considerar normalmente distribuidas, con un nivel de significación del 5%. Utilizar el contraste de Lilliefors y Shapiro-Wilk

184. En la tabla aparecen los resultados de una encuesta que se realizó a cien personas sobre las veces que viajaban en avión durante un año.

x_i	1	2	3	4	5	6	7	8	9	10	11	12	14
n_i	2	5	10	14	16	12	11	8	5	6	8	2	1

Con una significación del 5%, determinar si la muestra se ajusta a una distribución de Weibull de parámetros $b = 6$ y $c = 2$, empleando la prueba de Kolmogorov-Smirnov.

185. Una compañía dedicada al envasado, fabricación y venta de productos lácteos pretende analizar el consumo anual de leche en una ciudad. Para realizar el estudio, decide llevar a cabo una encuesta a cien personas, clasificando por edades los litros de leche medios consumidos, obteniendo los siguientes datos:

$L_{i-1} - L_i$	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50	50 – 60
n_i	26	21	16	15	12	10

Con un nivel de significación del 5%, empleando la prueba de Kolmogorov-Smirnov, determinar si la muestra se ajusta a una distribución uniforme $U(0,60)$

186. En la tabla aparecen las puntuaciones obtenidas en un test de estrés que se ha pasado en dos ciudades.

A	18	13	22	9	12	21	14	17	15	10	16	19	11
B	12	16	11	15	9	8	13	6	9	13	7	22	

Con un nivel de significación del 1%, a partir de los datos, puede afirmarse que el estrés es mayor en la ciudad A que en la ciudad B. Utilizar el test de la U de Wilcoxon-Mann-Whitney, el contraste de Kolmogorov-Smirnov y la prueba de la Mediana.

187. Durante los meses de declaración de la renta una sucursal bancaria decidió dar cita previa a los clientes para evitar largas esperas. A pesar de la media, los clientes esperan un tiempo medio de 5 minutos antes de ser atendidos. Determinado día se eligió al azar anotar los tiempos de espera de cada uno de los clientes, contabilizando los tiempos:

4 0 3 9 6,5 7,5 4 6 5,5

Para un nivel de significación del 5%, ¿se ajusta la muestra a una distribución exponencial?

188. En un test de coeficiente de inteligencia se anotaron las respuestas incorrectas que figuran en la tabla siguiente:

Intervalos $L_{i-1} - L_i$	50 – 60	60 – 70	70 – 80	80 – 90	90 – 100	100 – 110
Frecuencia n_i	30	25	120	10	10	5

Con un nivel de significación del 5%, ¿se ajusta la muestra a una distribución exponencial? Utilizar el test de Kolmogorov-Smirnov y la prueba de la Chi-cuadrado.

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

189. En un vuelo se adjudican los asientos a pasajeros que proceden de dos agencias de viajes (A y B) de la forma que se presenta:

Asiento	Agencia viajes
1	A
2	B
3	B
4	B
5	A
6	B
7	A
8	B
9	B
10	B
11	B
12	A
13	A
14	B
15	B
16	A
17	B
18	A
19	A
20	A
21	B
22	B
23	B
24	B
25	A
26	A
27	B
28	A
29	B
30	B

¿Se han asignado los asientos aleatoriamente, con un nivel de significación del 5%?

190. La asociación de mutuas de accidentes analiza el número de bajas ocurridas en el último mes debidas a accidentes laborales, en empresas constructoras que realizan su actividad en la capital de provincia. Los datos recogidos aparecen en la tabla siguiente:

Intervalos $L_{i-1} - L_i$	0 – 2	2 – 4	4 – 6	6 – 8	8 – 12
Frecuencia n_i	22	35	32	3	3

Con un nivel de significación del 5%, ¿se ajusta la muestra a una distribución normal?. Utilizar la prueba de la Chi-cuadrado y la prueba de Lilliefors.

191. Para tratar de evaluar las diferentes capacidades por sexo, se somete a un examen lógico a una muestra de hombres y mujeres, obteniendo las siguientes puntuaciones:

Hombres	52,48	72,98	73,8	22,14	36,9	82	73,8	40,18	24,6	57,4	123
Mujeres	65,6	48,38	82	123	143,5	73,8	24,6	26,24	49,2	65,6	131,2

Con una significación del 5%, ¿existen diferencias significativas entre hombres y mujeres en cuanto a la dispersión de las puntuaciones?. Utilizar el contraste de Siegel-Tukey.

192. En un estudio sobre las sanciones que recaen sobre empresas por incumplimiento de normas de seguridad en el trabajo se encontró que las sanciones medias eran las mismas en dos comunidades autónomas (X e Y). En cada comunidad se eligieron al azar expedientes sancionadores, presentando en la tabla adjunta los importes (expresados en cientos de euros).

X	78,4	119	32,2	89,6	105	128,8	63	53,2	93,8	56	102,2	18,2	72,8
Y	84	81,2	70	44,8	96,6	116,2	50,4	67,2	77	123,2	98	28	

Con un 10% de significación, ¿Podría afirmarse que existen diferencias significativas en cuanto a la dispersión de las sanciones de seguridad en el trabajo en las dos comunidades?

Utilizar el contraste de Siegel-Tukey.

193. Se tienen cuatro razas de gatos. Se toman tres muestras elegidas al azar de forma que en cada una de ellas hay un gato de cada raza. Se dispone de tres tipos de vitaminas distintas y se administra una de ellas solamente a cada muestra. Se elige como variable de respuesta el peso ganado por cada uno de los gatos, información que se recoge en la tabla:

Razas	Tratamientos		
	Vitamina A	Vitamina B	Vitamina C
1	80	70	82
2	60	32	35
3	90	82	85
4	95	89	78

Con una significación del 5%, se quiere saber:

- 1) Si la ganancia media de pesos es equivalente para las tres vitaminas
- 2) Si el peso medio ganado es significativamente distinto de una raza a otra

194. Se usaron tres clases de gasolina en cuatro marcas de automóviles. ¿Indican los valores de la tabla (kilómetros recorridos por litro) una diferencia significativa entre las clases de gasolina?. ¿Y entre las marcas de automóviles?. Significación del 5%.

Gasolina	Automóviles			
	A1	A2	A3	A4
G1	6,290	6,460	6,418	6,205
G2	6,503	6,290	7,225	6,843
G3	7,310	7,183	7,438	7,055

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

195. Analizar el efecto de las variables Edad y Fumar sobre la Ansiedad Social, con un nivel de significación del 5%.

Edad	Fumar	
	Si	No
15 - 25	3,91	4,83
	5,01	3,95
	4,44	4,04
	3,33	3,66
	4,71	9,42
25 - 35	5,65	9,66
	6,49	7,68
	5,47	9,57
	5,72	7,93
	5,44	7,39
35 45	4,94	5,92
	7,13	5,45
	5,54	5,19
	5,92	6,12
	6,16	4,45

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día

196. Se realiza un experimento para estudiar el efecto del nivel del agua sobre la longitud del tallo de dos tipos de plantas de guisantes, utilizando una significación del 5%. Para ello se utilizaron tres niveles de agua, los datos obtenidos se reflejan en la tabla adjunta, figurando en rojo el orden temporal de la toma de datos.

		Nivel del agua		
		Nivel I	Nivel II	Nivel III
Tipo Planta	Tipo 1	71,3 (1)	107,5 (2)	123,1 (11)
		75,1 (4)	96,1 (8)	125,2 (16)
		69,0 (5)	103,6 (18)	125,7 (19)
		73,2 (7)	100,7 (21)	121,0 (20)
		74,4 (22)	102,3 (26)	122,9 (30)
	Tipo 2	70,4 (13)	88,1 (6)	109,0 (3)
		73,2 (15)	85,8 (10)	101,1 (9)
		71,1 (23)	86,0 (17)	103,2 (12)
		71,2 (24)	87,5 (25)	109,7 (14)
		69,2 (27)	81,0 (29)	106,1 (28)

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día

RESOLUCIÓN DINÁMICA



PROBLEMAS RESUELTOS EXAMEN ESTADÍSTICA APLICADA

1. En una cadena de televisión se hizo una encuesta a 2500 personas para saber la audiencia de un debate y de una película que se emitieron en horas distintas: 2100 personas vieron la película, 1500 vieron el debate y 350 no vieron ninguno de los dos programas.

Eligiendo al azar a uno de los encuestados, se desea saber:

- Probabilidad de que viera la película y el debate.
- Probabilidad de que viera la película, sabiendo que vio el debate.
- Habiendo visto la película, probabilidad de que viera el debate.

Solución:

Sean los sucesos: P = "Ver la película", D = "Ver el debate"

Organizando los datos en una tabla de doble entrada:

	D	$\bar{D}$	
P	1450	650	2100
$\bar{P}$	50	350	400
	1500	1000	2500

$$a) P(P \cap D) = \frac{1450}{2500} = 0,58$$

$$b) P(P / D) = \frac{P(P \cap D)}{P(D)} = \frac{1450}{1500} = 0,97$$

$$c) P(D / P) = \frac{P(D \cap P)}{P(P)} = \frac{1450}{2100} = 0,69$$

2. Un psicólogo de una compañía aérea, por experiencias anteriores, conoce que el 90% de los tripulantes de cabina (TCP) que inician un determinado tratamiento técnico terminan con éxito. La proporción de TCPs con entrenamiento y con experiencia previa es del 10% de entre los que completaron con éxito su entrenamiento y del 25% de entre aquellos que no terminaron con éxito su entrenamiento. Se desea saber:

- Probabilidad de que un TCP con experiencia previa supere el entrenamiento con éxito.
- ¿La experiencia previa influye en el éxito del entrenamiento?.

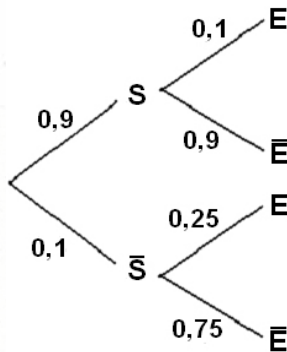
Solución:

Sean los sucesos:

S = "Supera el entrenamiento con éxito" E = "Tiene experiencia previa"

$$P(S) = 0,9 \quad P(\bar{S}) = 0,1 \quad P(E / S) = 0,1 \quad P(E / \bar{S}) = 0,25$$

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....



$$a) P(S/E) = \frac{P(S \cap E)}{P(E)} = \frac{0,9 \times 0,1}{0,9 \times 0,1 + 0,1 \times 0,25} = 0,78$$

$$b) \begin{cases} P(S/E) = 0,78 \\ P(S) = 0,9 \end{cases} \rightarrow P(S) > P(S/E)$$

La experiencia previa influye desfavorablemente en el éxito del tratamiento.

3. Sean tres sucesos incompatibles A, B y C, donde $P(A) = 0,5$, $P(B) = 0,25$ y $P(C) = 0,2$. Se pide: $P(\bar{A} \cap \bar{B})$, $P(\bar{A} \cap \bar{B} \cap \bar{C})$, $P(\bar{A} \cap \bar{B} \cap C)$, $P(B - A)$ y probabilidad de que ocurra exactamente uno de los sucesos.

Solución:

Al ser los sucesos A, B y C incompatibles: $P(A \cap B) = 0$, $P(A \cap C) = 0$, $P(B \cap C) = 0$ y $P(A \cap B \cap C) = 0$

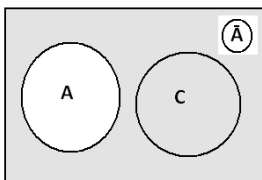
De otra parte,

$$\begin{aligned} P(A \cup B \cup C) &= P[(A \cup B) \cup C] = P(A \cup B) + P(C) - P[(A \cup B) \cap C] = \\ &= P(A) + P(B) - P(A \cap B) + P(C) - P[(A \cap C) \cup (B \cap C)] = \\ &= P(A) + P(B) - P(A \cap B) + P(C) - [P(A \cap C) + P(B \cap C) - P(A \cap B \cap C)] = \\ &= P(A) + P(B) - P(A \cap B) + P(C) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C) = \\ &= P(A) + P(B) + P(C) - \cancel{P(A \cap B)} - \cancel{P(A \cap C)} - \cancel{P(B \cap C)} + \cancel{P(A \cap B \cap C)} \end{aligned}$$

con lo que,

$$P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B}) = 1 - P(A \cup B) = 1 - [P(A) + P(B) - \cancel{P(A \cap B)}] = 1 - 0,5 - 0,25 = 0,25$$

$$P(\bar{A} \cap \bar{B} \cap \bar{C}) = P(\overline{A \cup B \cup C}) = 1 - P(A \cup B \cup C) = 1 - [P(A) + P(B) + P(C)] = 1 - (0,5 + 0,25 + 0,2) = 0,05$$



$$C \subset \bar{A} \rightarrow P(\bar{A} \cap C) = P(C) = 0,2$$

$$C \subset \bar{B} \rightarrow P(\bar{B} \cap C) = P(C) = 0,2$$

$$P(\bar{A} \cap \bar{B} \cap C) = P(C) = 0,2$$

$$P(B - A) = P(B \cap \bar{A}) = P(B) = 0,25$$

4. Tres tiradores hacen una descarga simultánea. Las probabilidades de hacer blanco son, respectivamente, 0,6 , 0,5 y 0,4. Calcular las probabilidades de los sucesos:

- Algún tirador hace blanco.
- Exactamente dos tiradores hacen blanco.
- El tercer tirador hace blanco, sabiendo que dos primeros lo han hecho.

Solución:

$$\text{a) Sean los tiradores A, B y C } \begin{cases} P(A) = 0,6 & P(\bar{A}) = 0,4 \\ P(B) = 0,5 & P(\bar{B}) = 0,5 \\ P(C) = 0,4 & P(\bar{C}) = 0,6 \end{cases}$$

Se resuelve por el suceso contrario, que no haga blanco ningún tirador:

$$P(\bar{A} \cap \bar{B} \cap \bar{C}) = P(\bar{A}) \times P(\bar{B}) \times P(\bar{C}) = 0,4 \times 0,5 \times 0,6 = 0,12$$

$$P(A \cup B \cup C) = 1 - P(\bar{A} \cap \bar{B} \cap \bar{C}) = 1 - 0,12 = 0,88$$

b) Exactamente dos tiradores hacen blanco

$$\begin{aligned} P[(A \cap B \cap \bar{C}) \cup (A \cap \bar{B} \cap C) \cup (\bar{A} \cap B \cap C)] &= P[A \cap B \cap \bar{C}] + P[A \cap \bar{B} \cap C] + P[\bar{A} \cap B \cap C] = \\ &= (0,6 \times 0,5 \times 0,6) + (0,6 \times 0,5 \times 0,4) + (0,4 \times 0,5 \times 0,4) = 0,38 \end{aligned}$$

c) Sea el suceso D = "Los dos primeros tiradores hacen blanco" $D = (A \cap \bar{B}) \cup (\bar{A} \cap B)$

$$P[C/D] = \frac{P(C \cap D)}{P(D)} = \frac{P[(A \cap \bar{B} \cap C) \cup (\bar{A} \cap B \cap C)]}{P(D)} = \frac{(0,6 \times 0,5 \times 0,4) + (0,4 \times 0,5 \times 0,4)}{0,38} = 0,526$$

5. Sean A y B dos sucesos tales que $P(A) = 0,6$. Calcular $P(A \cap \bar{B})$ en cada caso:

- a) A y B mutuamente excluyentes
- b) A está contenido en B
- c) B está contenido en A y $P(B) = 0,3$
- d) $P(A \cap B) = 0,1$

Solución:

a) A y B mutuamente excluyentes $\Rightarrow A \cap B = \emptyset \Rightarrow P(A \cap B) = 0$

$$P(A \cap \bar{B}) = P(A) - P(A \cap B) = P(A) - 0 = 0,6$$

b) $A \subset B \Rightarrow P(A \cap B) = P(A) \Rightarrow P(A \cap \bar{B}) = P(A) - P(A \cap B) = P(A) - P(A) = 0$

c) $B \subset A \Rightarrow P(A \cap B) = P(B) \Rightarrow$

$$\Rightarrow P(A \cap \bar{B}) = P(A) - P(A \cap B) = P(A) - P(B) = 0,6 - 0,3 = 0,3$$

d) $P(A \cap \bar{B}) = P(A) - P(A \cap B) = 0,6 - 0,1 = 0,5$

6. Se sabe que $P(B / A) = 0,9$, $P(A / B) = 0,2$ y $P(A) = 0,1$

- a) Calcular $P(A \cap B)$ y $P(B)$
- b) ¿Son independientes los sucesos A y B?
- c) Calcular $P(A \cup \bar{B})$

Solución:

$$a) P(B / A) = \frac{P(B \cap A)}{P(A)} = 0,9 \Rightarrow P(B \cap A) = 0,9 \times P(A) = 0,9 \times 0,1 = 0,09$$

$$P(A / B) = \frac{P(A \cap B)}{P(B)} = 0,2 \Rightarrow P(A \cap B) = 0,2 \times P(B) = 0,09 \Rightarrow P(B) = \frac{0,09}{0,2} = 0,45$$

$$P(\bar{B}) = 1 - P(B) = 1 - 0,45 = 0,55$$

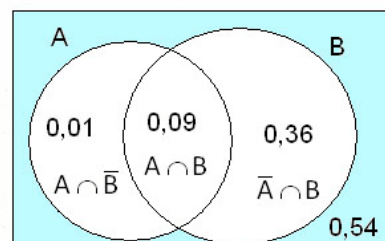
b) Los sucesos no son independientes, dado que:

$$P(A \cap B) = 0,09 \neq P(A) \times P(B) = 0,1 \times 0,45$$

o bien, no son independientes porque

$$P(A / B) = 0,2 \neq P(A) = 0,1$$

c) $P(\bar{B}) = 1 - P(B) = 1 - 0,45 = 0,55$

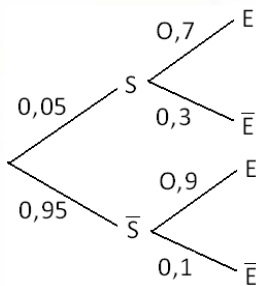


$$P(A \cup \bar{B}) = \begin{cases} = 0,1 + 0,54 = 0,64 \\ = P(A) + P(\bar{B}) - P(A \cap \bar{B}) = 0,1 + 0,55 - 0,01 = 0,64 \end{cases}$$

7. Una empresa de mensajería sabe que en condiciones normales un paquete es entregado en plazo el 90% de las veces, aunque si hay sobrecarga de trabajo (que ocurre un 5% de las veces) el porcentaje de retrasos se eleva al 30%.

- a) Cuál es la probabilidad de que un paquete llegue en plazo a su destino.
b) Sabiendo que se ha recibido una queja por retraso en el envío, el mensajero afectado aduce que ese día hubo sobrecarga de trabajo, aunque realmente no recuerda bien que sucedió. ¿Qué probabilidad hay de que efectivamente esté en lo cierto?.

Solución:



Sean los sucesos:

E = " Entrega a tiempo del paquete"

S = " Hay sobrecarga de trabajo"

a) $P(E) = P(S) \times P(E/S) + P(\bar{S}) \times P(E/\bar{S}) = 0,05 \times 0,7 + 0,95 \times 0,9 = 0,89$

b) $P(S/\bar{E}) = \frac{P(S \cap \bar{E})}{P(\bar{E})} = \frac{0,05 \times 0,3}{1 - 0,89} = 0,136$

8. En una región se cobra a los visitantes de los parques naturales, estimando que la variable aleatoria número de personas que visitan el parque en coche sigue la siguiente distribución:

x_i	1	2	3	4	5
$P(X = x_i)$	0,15	0,2	0,35	0,2	0,1

- a) Hallar el número medio de visitantes por vehículos.
b) Hallar cuánto debe pagar cada visitante para que la ganancia por coche sea 2 euros.
c) Si cada persona paga p euros, ¿cuál es la ganancia esperada en un día en que entran mil vehículos?

Solución:

- a) X = " Número de personas que visitan el parque en coche"

$$E(X) = \sum_{i=1}^5 x_i P(X = x_i) = 1 \times 0,15 + 2 \times 0,2 + 3 \times 0,35 + 4 \times 0,2 + 5 \times 0,1 = 2,9 \text{ visitante/vehículo}$$

b) Sea la variable aleatoria $Y = \text{"Ganancia por coche"}$

$$E(Y) = E(pX) = p \times E(X) = p \times 2,9 = 2 \quad \mapsto \quad p = \frac{2}{2,9} = 0,689$$

c) Sea la variable aleatoria $U = \text{"Ganancia de un día"}$

$$U = \underbrace{\text{precio}}_{\text{visitante}} \underbrace{\times}_{\text{p}} \underbrace{\text{número de}}_{\text{vehículos}} \underbrace{\times}_{\text{c}} \underbrace{\text{número personas}}_{\text{por vehículo}} \underbrace{\text{X}}_{\text{X}}$$

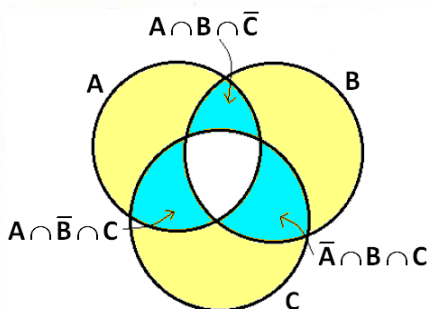
$$\text{Ganancia esperada día: } E(U) = E(p \times c \times X) = p \times c \times E(X) = p \times 1000 \times 2,9 = 2900 p$$

Sí $p = 0,689$ la ganancia esperada en un día será: $2900 \times 0,689 = 1998,1$ euros

9. Un estudiante busca información del sector aeronáutico en tres manuales, las probabilidades de que esa información se encuentre en el primero, segundo o tercer manual, respectivamente, son iguales a 0,6, 0,7 y 0,8.

¿Cuál es la probabilidad de que la información figure sólo en dos manuales?

Solución:



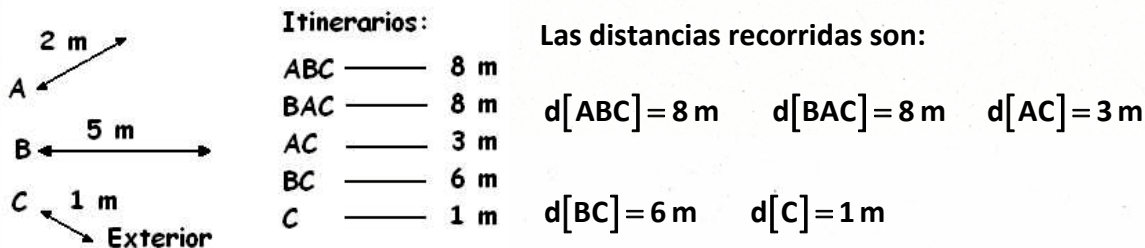
Sean A, B y C los tres sucesos en cuestión tales que $P(A) = 0,6$, $P(B) = 0,7$ y $P(C) = 0,8$.

Cada uno de los sucesos A, B y C es independiente con los otros dos, además los sucesos $(A \cap B \cap \bar{C})$, $(A \cap \bar{B} \cap C)$ y $(\bar{A} \cap B \cap C)$ son incompatibles.

$$\begin{aligned} P[(A \cap B \cap \bar{C}) \cup (A \cap \bar{B} \cap C) \cup (\bar{A} \cap B \cap C)] &= \\ &= P[A \cap B \cap \bar{C}] + P[A \cap \bar{B} \cap C] + P[\bar{A} \cap B \cap C] = \\ &= P(A) P(B) P(\bar{C}) + P(A) P(\bar{B}) P(C) + P(\bar{A}) P(B) P(C) = \\ &= 0,6 \cdot 0,7 \cdot 0,2 + 0,6 \cdot 0,3 \cdot 0,8 + 0,4 \cdot 0,7 \cdot 0,8 = 0,452 \end{aligned}$$

10. En el punto de partida de un laberinto hay tres orificios iguales A, B y C. Si una rata elige A vuelve al punto de partida después de recorrer dos metros. Si elige B recorre cinco metros y vuelve al mismo punto. Si elige C sale al exterior recorriendo un metro. ¿Por término medio que distancia recorre una rata antes de salir, si siempre elige un orificio distinto de los seleccionados en veces anteriores?

Solución:



Las probabilidades de las distancias recorridas:

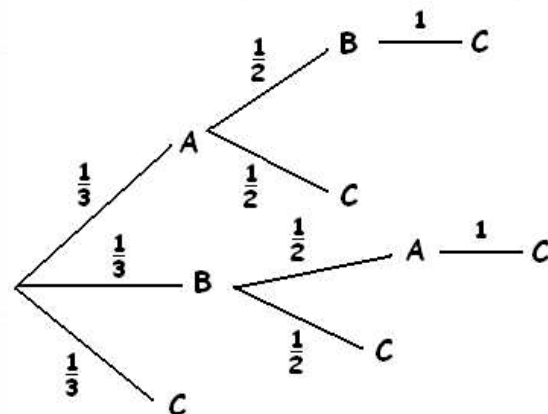
$$P[ABC] = P[A] \cdot P[B/A] \cdot P[C] = \frac{1}{3} \cdot \frac{1}{2} \cdot 1 = \frac{1}{6}$$

$$P[BAC] = P[B] \cdot P[A/B] \cdot P[C] = \frac{1}{3} \cdot \frac{1}{2} \cdot 1 = \frac{1}{6}$$

$$P[AC] = P[A] \cdot P[C/A] = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$$

$$P[BC] = P[B] \cdot P[C/B] = \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$$

$$P[C] = \frac{1}{3}$$



La distribución de probabilidad:

	ABC	BAC	AC	BC	C
d_i	8	8	3	6	1
p_i	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{3}$

La distancia media recorrida es:

$$\mu = E[d] = \sum_{i=1}^4 d_i \cdot p_i = 8 \cdot \frac{1}{6} + 8 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} + 1 \cdot \frac{1}{3} = \frac{27}{6} = 4,5 \text{ m}$$

11. Sean A, B y C tres sucesos mutuamente independientes con $P(A) = P(B) = P(C) = p$, donde $0 < p < 1$. Calcular la probabilidad de que ocurran exactamente dos de los tres sucesos considerados.

Solución:

Sea el suceso E = 'ocurran exactamente dos de los sucesos A, B, C'

$$E = (A \cap B \cap \bar{C}) \cup (A \cap \bar{B} \cap C) \cup (\bar{A} \cap B \cap C)$$

♦ Por ser incompatibles $(A \cap B \cap \bar{C})$, $(A \cap \bar{B} \cap C)$ y $(\bar{A} \cap B \cap C)$:

$$\begin{aligned} P(E) &= P[(A \cap B \cap \bar{C}) \cup (A \cap \bar{B} \cap C) \cup (\bar{A} \cap B \cap C)] = \\ &= P(A \cap B \cap \bar{C}) + P(A \cap \bar{B} \cap C) + P(\bar{A} \cap B \cap C) \end{aligned}$$

♦ Como son independientes A, B y C $\rightarrow \begin{cases} A, B, \bar{C} \text{ son independientes} \\ A, \bar{B}, C \text{ son independientes} \\ \bar{A}, B, C \text{ son independientes} \end{cases}$

$$\begin{aligned} P(E) &= P(A) \cdot P(B) \cdot P(\bar{C}) + P(A) \cdot P(\bar{B}) \cdot P(C) + P(\bar{A}) \cdot P(B) \cdot P(C) = \\ &= p^2(1-p) + p^2(1-p) + p^2(1-p) = 3p^2(1-p) \end{aligned}$$

12. En una cadena de televisión se hizo una encuesta a 3000 personas para saber la audiencia de un debate y de una película que se emitieron en horas distintas: 2400 personas vieron la película, 1600 vieron el debate y 400 no vieron ninguno de los dos programas. Se elige al azar uno de los encuestados:

- ¿Cuál es la probabilidad de que viera la película y el debate?.
- ¿Cuál es la probabilidad de que viera la película, sabiendo que vio el debate?.
- Sabiendo que vio la película, ¿Cuál es la probabilidad de que viera el debate?.

Solución:

Se organizan los datos en una tabla de doble entrada, completando los que faltan.

Sean los sucesos: P = "Ver la película" D = "Ver el debate"

	D	$\bar{D}$	
P	1400	1000	2400
$\bar{P}$	200	400	600
	1600	1400	3000

$$a) P(P \cap D) = \frac{1400}{3000} = \frac{7}{15} = 0,466$$

$$b) P(P / D) = \frac{P(P \cap D)}{P(D)} = \frac{1400}{1600} = \frac{7}{8} = 0,875$$

$$c) P(D / P) = \frac{P(D \cap P)}{P(P)} = \frac{1400}{2400} = \frac{7}{12} = 0,583$$

13. Si la probabilidad de que ocurra un suceso A es $1/5$

a) ¿Cuál es el mínimo número de veces que hay que repetir el experimento para que la probabilidad de que ocurra al menos una vez sea mayor que $1/2$?

b) ¿Cuál es la probabilidad de que ocurra al menos dos veces A al realizar 5 veces el experimento?

Solución:

a) $X =$ "Número de éxitos en n pruebas", $X \sim B(n, 0,2)$, $p = 0,2$, $q = 0,8$

$$P(X \geq 1) > \frac{1}{2} \rightarrow P(X < 1) \leq \frac{1}{2} \Rightarrow P(X = 0) \leq \frac{1}{2}$$

$$P(X = 0) = \binom{n}{0} \cdot 0,2^0 \cdot 0,8^n = 0,8^n \leq \frac{1}{2} \xrightarrow{0,8^4 = 0,4096} n = 4$$

Para que el suceso A tenga al menos una probabilidad mayor que $1/2$ hay que realizar el proceso un mínimo de 4 veces.

b) Se trata de una distribución binomial $B(5, 0,2)$

$$P(X \geq 2) = 1 - [P(X = 0) + P(X = 1)] = 1 - \left[\binom{5}{0} \cdot 0,2^0 \cdot 0,8^5 + \binom{5}{1} \cdot 0,2^1 \cdot 0,8^4 \right] =$$

$$= 1 - (0,3277 + 0,4096) = 0,2627$$

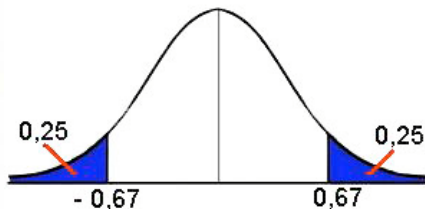
14. En una población de mujeres, las puntuaciones de un test de ansiedad-riesgo siguen una distribución normal $N(25, 10)$. Al clasificar la población en cuatro grupos de igual tamaño, ¿cuales serán las puntuaciones que delimiten estos grupos?

Solución:

Siendo la variable aleatoria $X =$ "Puntuaciones en un test de ansiedad-riesgo"

Las puntuaciones que delimitan estos cuatro grupos serán el primer Q_1 , segundo Q_2 y tercer cuartil Q_3 de la distribución.

$$P(X \leq Q_1) = 0,25 \rightarrow P\left(\frac{X-25}{10} \leq \frac{Q_1-25}{10}\right) = P\left(z \leq \frac{Q_1-25}{10}\right) = 0,25$$



$$\frac{Q_1-25}{10} = -0,67 \mapsto Q_1 = 25 - 0,67 \times 10 = 18,3$$

$$P(z \leq -0,67) = 0,25 \quad P(z \geq 0,67) = 0,25$$

En la distribución normal la media y la mediana son iguales: $\mu = M_e = Q_2 = 25$

$$P(X \leq Q_3) = 0,75 \rightarrow P\left(\frac{X-25}{10} \leq \frac{Q_3-25}{10}\right) = P\left(z \leq \frac{Q_3-25}{10}\right) = 0,75 \rightarrow P\left(z \geq \frac{Q_3-25}{10}\right) = 0,25$$

$$\frac{Q_3-25}{10} = 0,67 \rightarrow Q_3 = 25 + 0,67 \times 10 = 31,7$$

Por consiguiente, el primer grupo serían las mujeres con puntuaciones inferiores o iguales a 18,3. El segundo grupos son aquellas mujeres con puntuaciones entre 18,3 y 25. El tercer grupo son las mujeres con puntuaciones entre 25 y 31,7. El cuarto grupo son mujeres que tengan puntuaciones superiores a 31,7.

15. El departamento comercial de una industria alimenticia conoce que 2 de cada 10 consumidores reconocen su producto en una prueba a ciegas. ¿Cuántas pruebas ciegas de sabor deberían hacerse para que la proporción de que los que conocen la marca oscile entre el 16% y el 24% con una probabilidad mínima de 0,8?

Solución:

Reconocen el producto el 20%: $p = 0,2 \quad P(0,16 \leq \hat{p} \leq 0,24) \geq 0,8$

$$\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right) \quad \hat{p} \sim N\left(0,2, \sqrt{\frac{0,2 \times 0,8}{n}}\right) = N\left(0,2, \frac{0,4}{\sqrt{n}}\right)$$

$$P\left(\frac{0,16-0,2}{0,4/\sqrt{n}} < z < \frac{0,24-0,2}{0,4/\sqrt{n}}\right) = P\left(-0,1\sqrt{n} < z < 0,1\sqrt{n}\right) = 0,8$$

$$P\left(-0,1\sqrt{n} < z < 0,1\sqrt{n}\right) = 1 - 2P\left(z > 0,1\sqrt{n}\right) = 0,8 \quad \mapsto \quad P\left(z \geq 0,1\sqrt{n}\right) = 0,1$$

$$0,1\sqrt{n} = 1,282 \quad \rightarrow \quad n = 165$$

16. Se está experimentando una nueva vacuna para la malaria que resulta efectiva en el 60% de los casos. Si se eligen al azar 45 personas. Se pide:

- Probabilidad de que en el grupo elegido la vacuna sea efectiva en 27 personas.
- Probabilidad de que sea efectiva en menos de 20 personas.

Solución:

Denotando por X = 'Vacuna para la malaria', esta variable aleatoria sigue una distribución binomial, con $n = 45$ y $p = 0,6$, siendo este último valor la probabilidad de que sea efectiva la vacuna.

$$X \sim B(45, 0,6) \text{ con } \begin{bmatrix} n \cdot p = 27 > 5 \\ n \cdot p \cdot q = 10,8 \end{bmatrix} \rightarrow X \sim N(27, \sqrt{10,8}) \equiv N(27, 3,28)$$

$$\text{a) } P[X = 27] \xrightarrow{\text{factor corrección}} P[26,5 < X < 27,5] = P\left[\frac{26,5-27}{3,28} < \frac{X-27}{3,28} < \frac{27,5-27}{3,28}\right] =$$

$$= P[-0,15 < z < 0,15] = P[z > -0,15] - P[z > 0,15] = P[z < 0,15] - P[z > 0,15] =$$

$$= P[z < 0,15] - P[z > 0,15] = 1 - 2P[z > 0,15] = 1 - 2 \cdot 0,4404 = 0,1192$$

$$\text{b) } P[X < 20] = P\left[\frac{X-27}{3,28} < \frac{20-27}{3,28}\right] = P[z < -2,13] = P[z > 2,13] = 0,0166$$

17. Una maca de frutas ha observado que el peso medio de los melones en gramos sigue una distribución normal $N(1700, 100)$. Hallar:

- Probabilidad de que los melones pesen menos de 1500 gramos y más de 2000 gramos.
- Sabiendo que son rechazados para la explotación aquellos melones que difieren más de 300 gramos del promedio. Determinar la proporción de melones rechazados.

Solución:

a) Sea la variable aleatoria $X = \text{"Peso de los melones en gramos"}$, $X \sim N(1700, 100)$

$$\begin{aligned} P[(X < 1500) \cup (X > 2000)] &= P(X < 1500) + P(X > 2000) = \\ &= P\left[\frac{X - 1700}{100} < \frac{1500 - 1700}{100}\right] + P\left[\frac{X - 1700}{100} > \frac{2000 - 1700}{100}\right] = P(z < -2) + P(z > 3) = \\ &= P(z > 2) + P(z > 3) = 0,0228 + 0,00135 = 0,02415 \end{aligned}$$

$$\begin{aligned} \text{b) } P(1400 \leq X \leq 2000) &= P\left[\frac{1400 - 1700}{100} \leq \frac{X - 1700}{100} \leq \frac{2000 - 1700}{100}\right] = P(-3 \leq z \leq 3) = \\ &= 1 - 2P(z \geq 3) = 1 - 2 \times 0,00135 = 0,9973 \text{ melones aceptados} \end{aligned}$$

$$\text{melones rechazados: } 2P(z \geq 3) = 2 \times 0,00135 = 0,0027$$

18. Un ascensor limita el peso de sus cuatro ocupantes a 300 kilogramos. Si el peso de una persona sigue una distribución normal $N(71, 7)$, calcular la probabilidad de que el peso 4 personas supere los 300 kilogramos.

Solución:

Si el peso de una persona sigue una distribución normal $N(71, 7)$, la media muestral de 4 personas sigue

$$\text{una distribución normal } \bar{x} \sim N\left(71, \frac{7}{\sqrt{4}}\right) \equiv N(71, 3,5)$$

$$\begin{aligned} P[(X_1 + X_2 + X_3 + X_4) > 300] &= P\left[\frac{X_1 + X_2 + X_3 + X_4}{4} > \frac{300}{4}\right] = P[\bar{x} > 75] = \\ &= P[\bar{x} > 75] = P\left[\frac{\bar{x} - 71}{3,5} > \frac{75 - 71}{3,5}\right] = P[z > 1,143] = 0,1265 \end{aligned}$$

$$\text{Interpolando: } \frac{0,1271 - 0,1251}{1,14 - 1,15} = \frac{x - 0,1251}{1,143 - 1,15} \mapsto \frac{0,002}{-0,01} = \frac{x - 0,1251}{-0,007}$$

$$x = 0,1251 + \frac{0,002 \cdot 0,007}{0,01} = 0,1265$$

Otro Método:

$$\text{Si } x_i \sim N[71, 7] \mapsto \sum_{i=1}^4 x_i \sim N(\underbrace{4 \times 71}_{n \cdot \mu}, \underbrace{7 \times \sqrt{4}}_{\sigma \cdot \sqrt{n}}) = N[284, 14]$$

$$P\left[\sum_{i=1}^4 x_i \geq 300\right] = P\left[\frac{\sum_{i=1}^4 x_i - 284}{14} \geq \frac{300 - 284}{14}\right] = P[z \geq 1,143] = 0,1265$$

19. Una empresa aeronáutica vende cilindros para motor en paquetes de 25 unidades. Se sabe que el peso de cada cilindro es una variable aleatoria que tienen una media de 12 Kg y una desviación típica de 2 Kg. Se pide:

- ¿Cuál es la probabilidad de que en un paquete de 25 unidades escogidas aleatoriamente, el peso medio de cada cilindro sea superior a 13 kilogramos?
- ¿Cuál sería el resultado si la varianza poblacional fuera desconocida?, suponiendo entonces que la cuasidesviación típica es de 3 Kg?

Solución:

Sea la variable aleatoria $X =$ "Peso de cada cilindro"

Se trata de una muestra de una población normal $N[12, 2]$ con varianza conocida.

Al ser la varianza poblacional conocida, aunque la muestra es pequeña se supone que el peso

$$\text{medio muestral } \bar{x} \sim N\left[12, \frac{2}{\sqrt{25}}\right] \equiv N[12, 0,4]$$

$$\text{a) } P[\bar{x} \geq 13] = P\left[\frac{\bar{x} - 12}{0,4} \geq \frac{13 - 12}{0,4}\right] = P[z \geq 2,5] = 0,00621$$

b) En este caso, se trata de una población normal $N[12, \sigma]$ con varianza desconocida, en una muestra de tamaño $n = 25$, el peso medio muestral $\bar{x}$ sigue una t de Student con $n - 1 = 24$ grados de libertad, es decir:

$$\bar{x} \sim t_{n-1}\left(\mu, \frac{s_x}{\sqrt{n}}\right) \quad t_{n-1} = \frac{\bar{x} - \mu}{\frac{s_x}{\sqrt{n}}} \quad \text{donde } s_x \equiv \text{cuasidesviación típica muestral}$$

$$P[\bar{x} \geq 13] = P\left(\frac{\bar{x} - 12}{3/\sqrt{25}} \geq \frac{13 - 12}{3/\sqrt{25}}\right) = P[t_{24} \geq 1,666] = 0,056$$

Interpolando:

$$\frac{1,318 - 1,711}{0,10 - 0,05} = \frac{1,666 - 1,711}{x - 0,05} \rightarrow \frac{-0,393}{0,05} = \frac{-0,045}{x - 0,05} \rightarrow x = 0,05 + \frac{0,05 \cdot 0,045}{0,393} = 0,056$$

20. El error cometido en expedir tiques por una maquina en el aeropuerto sigue una normal $N(0, \sigma)$.

- a) ¿Cuál es la probabilidad de que el error cometido en valor absoluto de una medida cualquiera sea al menos σ ?
- b) ¿Cuánto valdría la probabilidad si se toma como medida la media aritmética de 10 medidas independientes?

Solución:

a) Sea la variable aleatoria $X =$ "Expedir tiques por la maquina", $X \sim N(0, \sigma)$

$$P[|X| \geq \sigma] = P\left[\left|\frac{X - 0}{\sigma}\right| \geq \frac{\sigma - 0}{\sigma}\right] = P[|z| \geq 1] = P[(z \leq -1) \cup (z \geq 1)] = P(z \leq -1) + P(z \geq 1) =$$

$$= P(z \geq 1) + P(z \geq 1) = 2P(z \geq 1) = 2 \times 0,1587 = 0,3174$$

b) Sea la variable aleatoria $Y = X_1 + X_2 + \dots + X_{10} \mapsto \bar{y} = \frac{1}{10} \sum_{i=1}^{10} X_i$

$$E(\bar{y}) = E\left[\frac{1}{10} \sum_{i=1}^{10} X_i\right] = \frac{1}{10} E\left[\sum_{i=1}^{10} X_i\right] = \frac{1}{10} 10 \mu = \mu = 0$$

$$\sigma_{\bar{y}}^2 = \text{Var}(\bar{y}) = \text{Var}\left[\frac{1}{10} \sum_{i=1}^{10} X_i\right] = \frac{1}{100} \text{Var}\left[\sum_{i=1}^{10} X_i\right] = \frac{1}{100} 10 \sigma^2 = \frac{\sigma^2}{10} \quad \bar{y} \sim N\left[0, \frac{\sigma}{\sqrt{10}}\right]$$

$$P[|\bar{y}| \geq \sigma] = P\left[\left|\frac{\bar{y} - 0}{\sigma/\sqrt{10}}\right| \geq \frac{\sigma - 0}{\sigma/\sqrt{10}}\right] = P[|z| \geq \sqrt{10}] = P[(z \leq -\sqrt{10}) \cup (z \geq \sqrt{10})] =$$

$$= P(z \leq -\sqrt{10}) + P(z \geq \sqrt{10}) = P(z \geq \sqrt{10}) + P(z \geq \sqrt{10}) = 2P(z \geq 3,16) = 0,0016$$

21. Las variables aleatorias en estudio son independientes.

Sean: $X_1 \sim N(10, 2)$, $X_2 \sim N(15, 2)$ y $X_3 \sim N(20, 2)$ donde $Y = 5X_1 + 4X_2 - 3X_3$ y

$$U = \left(\frac{X_1 - 10}{2} \right)^2 + \left(\frac{X_2 - 15}{2} \right)^2 + \left(\frac{X_3 - 20}{2} \right)^2.$$

Calcular: a) $P[20 \leq Y \leq 30]$ b) $P[U \leq 11]$

Solución:

a) $Y = 5X_1 + 4X_2 - 3X_3$ siendo $Y \sim N\left(\sum_{i=1}^3 \lambda_i \mu_i, \sqrt{\sum_{i=1}^3 \lambda_i^2 \sigma_i^2}\right)$

$$Y \sim N\left[5 \times 10 + 4 \times 15 - 3 \times 20, \sqrt{5^2 \times 2^2 + 4^2 \times 2^2 + (-3)^2 \times 2^2}\right] \rightarrow Y \sim N[50, 2\sqrt{50}]$$

$$P[20 \leq Y \leq 30] = P\left[\frac{20 - 50}{2\sqrt{50}} \leq \frac{Y - 50}{2\sqrt{50}} \leq \frac{30 - 50}{2\sqrt{50}}\right] = P(-2,12 \leq z \leq -1,41) =$$

$$= P(1,41 \leq z \leq 2,12) = P(z \geq 1,41) - P(z \geq 2,12) = 0,0793 - 0,0170 = 0,0623$$

b)

$$X_1 \sim N(10, 2) \rightarrow \frac{X_1 - 10}{2} \sim N(0, 1) \quad X_2 \sim N(15, 2) \rightarrow \frac{X_2 - 15}{2} \sim N(0, 1)$$

$$X_3 \sim N(20, 2) \rightarrow \frac{X_3 - 20}{2} \sim N(0, 1)$$

$$U = \left(\frac{X_1 - 10}{2} \right)^2 + \left(\frac{X_2 - 15}{2} \right)^2 + \left(\frac{X_3 - 20}{2} \right)^2 \mapsto U = [N(0, 1)]^2 + [N(0, 1)]^2 + [N(0, 1)]^2 = \chi_3^2$$

$$P[U \leq 11] = P[\chi_3^2 \leq 11] = 1 - P[\chi_3^2 \geq 11] = 1 - 0,05 = 0,95$$

22. El número de millones de metros cúbicos que tiene un embalse sigue una distribución normal $N(980, 50)$. El consumo diario de las poblaciones que sirve es una normal $N(85, 30)$. Si se sabe que durante una tormenta la cantidad de agua que se embalsa es una normal $N(50, 25)$. Un día han caído dos tormentas, calcular la probabilidad de que al final del día el agua embalsada sea menor o igual que 980 metros cúbicos.

Solución:

Agua del embalse: $X_1 \sim N(\mu_1, \sigma_1) \equiv N(980, 50)$

Consumo diario: $X_2 \sim N(\mu_2, \sigma_2) \equiv N(85, 30)$

Agua una tormenta: $X_3 \sim N(\mu_3, \sigma_3) \equiv N(50, 25)$

Agua dos tormentas: $\begin{cases} E(2X_3) = 2E[X_3] = 2 \times 50 = 100 \\ \text{Var}(2X_3) = 4 \text{Var}[X_3] = 4 \times 25^2 \end{cases}$

Agua embalse un día con dos tormentas: $Y = X_1 - X_2 + 2X_3$

$Y \sim N(980 - 85 + 100, \sqrt{50^2 + 30^2 + 4 \times 25^2}) \equiv N(995, 76,81)$

$$P(Y \leq 980) = P\left[\frac{Y - 995}{76,81} \leq \frac{980 - 995}{76,81}\right] = P(z \leq -0,19) = P(z \geq 0,19) = 2P(z \geq 0,19) = 0,4247$$

23. Sean las variables aleatorias: $X \sim B(8, 0,4)$, $Y \sim B(7, 0,4)$, $U \sim B(55, 0,03)$, $V \sim B(65, 0,03)$
¿Qué distribución siguen las variables $X + Y$, $U + V$?

Solución:

$$X + Y \sim B(8 + 7, 0,4) \equiv B(15, 0,4)$$

$$U + V \sim B(55 + 65, 0,03) \equiv B(120, 0,03) \xrightarrow{p < 0,1 \text{ y } np = 3,6 < 5} P(\lambda = 3,6)$$

24. Se tienen dos variables aleatorias cuya distribución es normal: $X \sim N(5, 1)$, $Y \sim N(3, 2)$.
Obtener la esperanza matemática y la varianza de la variable $W = 2X - 3Y$ y de la variable $Z = (X - Y) / 2$

Solución:

$$W = 2X - 3Y \mapsto \begin{cases} E(W) = E(2X - 3Y) = 2E(X) - 3E(Y) = 2 \cdot 5 - 3 \cdot 3 = 1 \\ V(W) = V(2X - 3Y) = 4V(X) + 9V(Y) = 4 \cdot 1^2 + 9 \cdot 2^2 = 40 \end{cases} \quad W \sim N(1, \sqrt{40})$$

$$Z = \frac{X - Y}{2} \mapsto \begin{cases} E(Z) = E\left(\frac{X - Y}{2}\right) = \frac{1}{2} E(X - Y) = \frac{1}{2} E(X) - \frac{1}{2} E(Y) = \frac{5}{2} - \frac{3}{2} = 1 \\ V(Z) = V\left(\frac{X - Y}{2}\right) = \frac{1}{4} V(X - Y) = \frac{1}{4} V(X) + \frac{1}{4} V(Y) = \frac{1^2}{4} + \frac{2^2}{4} = \frac{5}{4} \end{cases} \quad Z \sim N\left(1, \sqrt{\frac{5}{4}}\right)$$

25. En cierta población, en una muestra de tamaño 100, la media muestral $\bar{x}$ de una característica sigue una distribución normal. Se conoce que $P(\bar{x} \leq 75) = 0,58$ y $P(\bar{x} > 80) = 0,04$. Hallar la media y la desviación típica poblacional.

Solución:

La media muestral $\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \equiv N\left(\mu, \frac{\sigma}{\sqrt{100}}\right) \equiv N\left(\mu, \frac{\sigma}{10}\right)$

Se desconocen los parámetros poblacionales μ y σ , por lo que se necesitan dos ecuaciones:

$$P(\bar{x} \leq 75) \stackrel{\text{TCL}}{=} P\left(\frac{\bar{x} - \mu}{\frac{\sigma}{10}} \leq \frac{75 - \mu}{\frac{\sigma}{10}}\right) = 0,58 \rightarrow P\left(\frac{\bar{x} - \mu}{\frac{\sigma}{10}} \geq \frac{75 - \mu}{\frac{\sigma}{10}}\right) = 0,42 \rightarrow \frac{75 - \mu}{\frac{\sigma}{10}} = 0,20$$

$$P(\bar{x} > 80) \stackrel{\text{TCL}}{=} P\left(\frac{\bar{x} - \mu}{\frac{\sigma}{10}} > \frac{80 - \mu}{\frac{\sigma}{10}}\right) = 0,04 \mapsto \frac{80 - \mu}{\frac{\sigma}{10}} = 1,75$$

De las ecuaciones $\begin{cases} 750 - 10\mu = 0,20 \cdot \sigma \\ 800 - 10\mu = 1,75 \cdot \sigma \end{cases} \rightarrow 50 = 1,55 \cdot \sigma \rightarrow \sigma = 32,26 \quad \mu = 74,35$

26. Para analizar el peso promedio de niños y niñas, siguiendo ambos pesos una distribución normal, se utiliza una muestra aleatoria de 20 niños y 25 niñas. El promedio de los pesos de los niños es 45 kg. con una desviación típica de 6,4 kg., mientras que el promedio del peso de las niñas es 38 kg. y una desviación típica de 5,6 kg. ¿Cuál es la probabilidad de que en la muestra el peso promedio de los niños sea al menos 10 kg. mayor que el de las niñas?

Solución:

Sean las variables aleatorias X = "Peso de los niños" e Y = "Peso de las niñas", $X \sim N(\mu_x, \sigma_x)$ e $Y \sim N(\mu_y, \sigma_y)$, independientes entre sí.

En las muestras respectivas:

$$\bar{x} \sim N\left(\mu_x, \frac{\sigma_x}{\sqrt{n}}\right) \equiv N\left(45, \frac{6,4}{\sqrt{20}}\right) \text{ e } \bar{y} \sim N\left(\mu_y, \frac{\sigma_y}{\sqrt{m}}\right) \equiv N\left(38, \frac{5,6}{\sqrt{25}}\right).$$

$$\xi = (\bar{x} - \bar{y}) \sim N\left(\mu_x - \mu_y, \sqrt{\left(\frac{\sigma_x}{\sqrt{n}}\right)^2 + \left(\frac{\sigma_y}{\sqrt{m}}\right)^2}\right) \equiv N\left(\mu_x - \mu_y, \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}\right)$$

La variable $\xi = (\bar{x} - \bar{y}) \sim N\left(45 - 38, \sqrt{\frac{6,4^2}{20} + \frac{5,6^2}{25}}\right) = N(7, 1,82)$

$$P(\xi \geq 10) = P\left(\frac{\xi - 7}{1,82} \geq \frac{10 - 7}{1,82}\right) = P(z \geq 1,648) = 0,05$$

27. Sea X una variable aleatoria que se distribuye según una normal $N(\mu, 4)$. Suponiendo que el tamaño muestral $n = 10$. Se pide:

- Determinar el número k tal que, exista una probabilidad de 0,99 de que el estadístico desviación típica muestral, sea mayor que él.
- Determinar el número k tal que, exista una probabilidad de 0,99 de que el estadístico cuasidesviación típica muestral, sea mayor que él.

Solución:

a) $P(\sigma_x > k) = 0,99 \rightarrow P(\sigma_x^2 > k^2) = 0,99 \quad n = 10 \quad X \sim N(\mu, 4)$

La distribución $\frac{n\sigma_x^2}{\sigma^2} = \frac{(n-1)s_x^2}{\sigma^2} \sim \chi_{n-1}^2$, donde $\begin{cases} \sigma_x^2 \equiv \text{varianza muestral} \\ s_x^2 \equiv \text{cuasivarianza muestral} \end{cases}$

$$P(\sigma_x > k) = 0,99 \rightarrow P(\sigma_x^2 > k^2) = 0,99 \rightarrow P\left(\frac{n\sigma_x^2}{\sigma^2} > \frac{n k^2}{\sigma^2}\right) = P\left(\chi_{n-1}^2 > \frac{n k^2}{\sigma^2}\right) = 0,99$$

por tanto,

$$P\left(\chi_{n-1}^2 > \frac{n k^2}{\sigma^2}\right) = P\left(\chi_9^2 > \frac{10 k^2}{16}\right) = 0,99 \rightarrow \frac{10 k^2}{16} = 2,09 \rightarrow k = \pm 1,82$$

La solución es $k = 1,82$. El resultado es independiente del valor de la media poblacional μ

b) $P(s_x > k) = 0,99 \rightarrow P(s_x^2 > k^2) = 0,99 \rightarrow P\left(\frac{(n-1)s_x^2}{\sigma^2} > \frac{(n-1) k^2}{\sigma^2}\right) = 0,99$

$$P\left(\chi_{n-1}^2 > \frac{(n-1)k^2}{\sigma^2}\right) = P\left(\chi_9^2 > \frac{9k^2}{16}\right) = 0,99 \rightarrow \frac{9k^2}{16} = 2,09 \rightarrow k = \pm 1,928$$

La solución es $k = 1,928$. El resultado es independiente del valor de la media poblacional μ

28. Una encuesta del periódico constó de 320 trabajadores de una multinacional que fueron despedidos entre 2017-2018,, encontrando que el 20% habían estado sin trabajo por lo menos durante dos años. Si el periódico tuviera que seleccionar otra muestra aleatoria de 320 empleados despedidos entre 2017-2018.

¿Cuál sería la probabilidad de que el porcentaje muestral de trabajadores sin empleo durante por lo menos dos años, difiera del porcentaje obtenido en la primera encuesta, en un 5% o más?.

Solución:

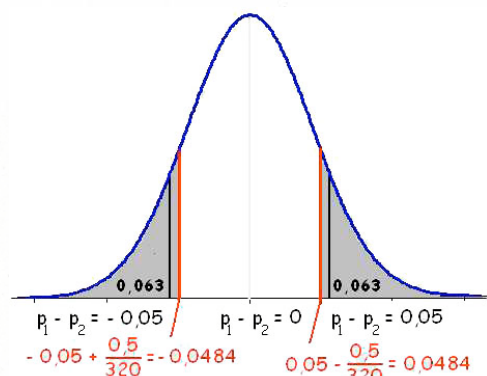
En la situación descrita hay una sola población, de la que se extraen dos muestras, queriendo conocer la diferencia de porcentajes de esas dos muestras, por lo que se debe utilizar la distribución muestral de diferencias de proporciones con $p_1 = p_2$ al tratarse de una misma población.

Otra cuestión a plantearse es que se desconoce la proporción poblacional $p_1 = p_2$, sólo se conoce que la proporción muestral en 320 trabajadores es $\hat{p}_1 = 0,20$.

Para ello, se utiliza la proporción muestral $\hat{p}_1 = 0,20$ como una estimación puntual de la proporción poblacional p_1 , es decir, $p_1 = E(\hat{p}_1) = 0,20$

Analizando la pregunta: ¿Cuál sería la probabilidad de que el porcentaje muestral de trabajadores sin empleo durante por lo menos dos años, difiera del porcentaje obtenido en la primera encuesta, en un 5% o más?.

Con diferir debe entenderse que la diferencia puede ser a favor de la muestra uno, o a favor de la muestra dos, por lo que se deben calcular dos áreas en la distribución y sumarlas.



$$n = m = 320 \quad p_1 - p_2 = 0$$

$$(\hat{p}_1 - \hat{p}_2) \sim N\left[(p_1 - p_2), \sqrt{\frac{p_1 \cdot q_1}{n} + \frac{p_2 \cdot q_2}{m}}\right]$$

$$\sqrt{\frac{p_1 \cdot q_1}{n} + \frac{p_2 \cdot q_2}{m}} = \sqrt{\frac{0,2 \cdot 0,8}{320} + \frac{0,2 \cdot 0,8}{320}} = 0,0316$$

$$(\hat{p}_1 - \hat{p}_2) \sim N[0, 0,316]$$

$$\begin{aligned}
 P[|\hat{p}_1 - \hat{p}_2| \geq 0,05] &= P[(\hat{p}_1 - \hat{p}_2 \geq 0,05) \cup (\hat{p}_1 - \hat{p}_2 \leq -0,05)] = \\
 &= P[(\hat{p}_1 - \hat{p}_2) \geq 0,05] + P[(\hat{p}_1 - \hat{p}_2) \leq -0,05] = (\text{factor corrección}) = \\
 &= P\left[(\hat{p}_1 - \hat{p}_2) \geq 0,05 - \frac{0,5}{320}\right] + P\left[(\hat{p}_1 - \hat{p}_2) \leq -0,05 + \frac{0,5}{320}\right] = \\
 &= P[(\hat{p}_1 - \hat{p}_2) \leq -0,0484] + P[(\hat{p}_1 - \hat{p}_2) \geq 0,0484] = \\
 &= P\left[\frac{(\hat{p}_1 - \hat{p}_2) - 0}{0,0316} \leq \frac{-0,0484 - 0}{0,0316}\right] + P\left[\frac{(\hat{p}_1 - \hat{p}_2) - 0}{0,0316} \geq \frac{0,0484 - 0}{0,0316}\right] = \\
 &= P\left[\frac{0,0484 - 0}{0,0316} \leq \frac{(\hat{p}_1 - \hat{p}_2) - 0}{0,0316} \leq \frac{-0,0484 - 0}{0,0316}\right] = P[1,53 \leq z] + P[z \leq -1,53] = \\
 &= 2 P[z \geq 1,53] = 2 \cdot 0,0630 = 0,1260
 \end{aligned}$$

29. Un individuo juega con probabilidad de ganar igual a $1/2$ en cada partida. Si gana en una partida obtiene 5 euros y si pierde paga 5 euros. En una tarde juega 400 partidas. ¿Qué cantidad debe llevar si quiere tener una probabilidad de 0,95 de hacer frente a posibles pérdidas?

Solución:

Sea la v.a. X = "número de partidas ganadas de 400", donde $X \sim B(400, 1/2)$

El tamaño de la muestra es grande $n > 30$, se tiene $X \sim N[200, 10]$, donde la media

$\mu = n \cdot p = 400 \cdot 1/2 = 200$ y la desviación típica $\sigma = \sqrt{n \cdot p \cdot q} = \sqrt{400 \cdot 1/2 \cdot 1/2} = 10$

El beneficio B que obtiene: $B = 5 \cdot X - 5 \cdot (400 - X) = 10 \cdot X - 2000$

La cantidad C que lleva: $P[B + C \geq 0] = 0,95 \rightarrow P[10 \cdot X - 2000 + C \geq 0] = 0,95$

$$P[10 \cdot X - 2000 + C \geq 0] = P\left[X \geq \frac{2000 - C}{10}\right] = P\left[\frac{X - 200}{10} \geq \frac{(2000 - C)/10 - 200}{10}\right] =$$

$$= P\left[z \geq \frac{2000 - C - 2000}{100}\right] = P\left[z \geq \frac{-C}{100}\right] = P\left[z \leq \frac{C}{100}\right] = 0,95$$

$$P\left[z \geq \frac{C}{100}\right] = 0,05 \mapsto \frac{C}{100} = 1,65 \Rightarrow C = 165 \text{ euros}$$

30. Un estudio de la DGT estima que el número de horas prácticas necesarias para la obtención del permiso de conducir sigue una distribución normal $N(24, 3)$. Se sabe que la autoescuela Fuenterrebollo ingresa por alumno una parte fija de 250 euros, más 23 euros por hora de práctica.

- ¿Qué probabilidad hay de obtener el permiso de conducir con 20 horas de prácticas o menos?
- ¿Cuántas horas de prácticas ha necesitado un conductor para obtener el permiso si el 68% de los conductores ha necesitado más horas de prácticas que él?
- Calcular el ingreso por alumno esperado.
- Calcular la desviación típica del ingreso por alumno.

Solución:

a) X = "número de horas de prácticas necesarias para obtener el permiso de conducir"

$$P(X \leq 20) = P\left(\frac{X - 24}{3} \leq \frac{20 - 24}{3}\right) = P(z \leq -1,33) = P(z \geq 1,33) = 0,0918$$

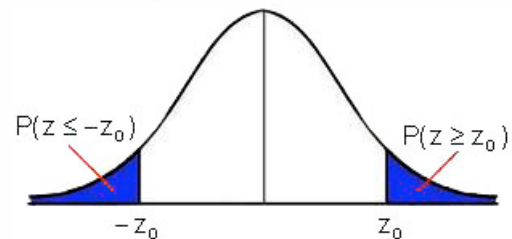
b) Sea h = "número de horas necesarias"

$$P(X \geq h) = P\left(\frac{X - 24}{3} \geq \frac{h - 24}{3}\right) = P\left(z \geq \frac{h - 24}{3}\right) = 0,68$$

$$P\left(z \geq \frac{h - 24}{3}\right) = 0,68 \rightarrow P\left(z \leq -\frac{h - 24}{3}\right) = 0,68$$

$$P\left(z \leq -\frac{h - 24}{3}\right) = 0,68 \Rightarrow P\left(z \geq -\frac{h - 24}{3}\right) = 0,32 \mapsto -\frac{h - 24}{3} = 0,46$$

$$-\frac{h - 24}{3} = 0,46 \mapsto h = 24 - 3 \times 0,46 = 22,62$$



c) Para un alumno cualquiera, el ingreso de la autoescuela Fuenterrebollo por las prácticas es $I = 23X + 250$

El ingreso por alumno esperado será:

$$E(I) = E(23X + 250) = 23E(X) + 250 = 23 \times 24 + 250 = 802 \text{ euros}$$

d) La varianza del ingreso del alumno:

$$\sigma_I^2 = V(I) = V(23X + 250) = 23^2 V(X) = 529 \times 9 = 4761 \text{ euros}^2$$

La desviación típica del ingreso $\sigma_I = \sqrt{4761} = 69 \text{ euros}$

31. La renta disponible mensualmente de una familia rural es una variable aleatoria que sigue una distribución normal $N(905, 450)$, expresados en euros. Para realizar un estudio sobre los hábitos de esta comunidad rural, se ha tomado una muestra aleatoria simple de 1000 familias. Calcular la probabilidad de que la cuasidesviación típica de la muestra sea superior a 430 euros.

Solución:

Denotando por $s_x^2 \equiv$ cuasivarianza muestral, se pide $P(s_x > 430)$

Por otra parte, siendo la muestra aleatoria simple, y procediendo de una población normal, se

$$\text{tiene: } \frac{(n-1)s_x^2}{\sigma^2} \sim \chi_{n-1}^2$$

$$\text{con lo cual, } \frac{999 s_x^2}{450^2} \sim \chi_{999}^2$$

$$P(s_x > 430) \rightarrow P(s_x^2 > 430^2) = P\left(\frac{999 \times s_x^2}{450^2} > \frac{999 \times 430^2}{450^2}\right) = P(\chi_{999}^2 > 912,17)$$

En las tablas de la χ^2 no se encuentran tantos grados de libertad, teniendo que recurrir al comportamiento asintótico de la normal proporcionado por Fisher:

$$\text{Para } n > 30 \text{ la variable } \sqrt{2\chi_n^2} \xrightarrow{n \rightarrow \infty} N(\sqrt{2n-1}, 1)$$

$$P(\chi_{999}^2 > 912,17) = P\left(\frac{\sqrt{2\chi_{999}^2} - \sqrt{2 \times 999 - 1}}{1} > \frac{\sqrt{2 \times 912,17} - \sqrt{2 \times 999 - 1}}{1}\right) =$$

$$= P(z > -1,97) = P(z < 1,97) = 1 - P(z > 1,97) = 1 - 0,0244 = 0,9756$$

32. En una muestra aleatoria de personas se analizan algunos hábitos de la vida, habiendo recogido datos de las siguientes variables:

$X_1 \equiv$ Estado general de salud: muy bueno (3), bueno (2), regular (1), malo (0)

$X_2 \equiv$ Sexo: mujer (1), hombre (0)

$X_3 \equiv$ Nivel del ejercicio diario: intenso (2), moderado (1), ninguno (0)

Realizadas las tablas de contingencia correspondientes, se calcularon los siguientes estadísticos para contrastar la asociación:

a) $\chi^2(X_1, X_2) = 8$ b) $\chi^2(X_2, X_3) = 4,5$ c) $\chi^2(X_1, X_3) = 6,1$

Con la información facilitada, a un nivel de significación del 5%, elaborar un diagnóstico para cada una de las parejas de variables.

Solución:

Calculando los p – valor (α_p) de cada estadístico se obtiene:

a) H_0 : X_1 e X_2 son independientes

En $\chi^2(X_1, X_2) = 8$ el número de grados de libertad es $(4 - 1) \times (2 - 1) = 3$

$\alpha_p = P(\chi_{p,3}^2 \geq 8)$. Interpolando en la tabla de la Chi-cuadrado:

0,05	α_p	0,025
7,815	8	9,348

$$0,05 - 0,025 \longrightarrow 7,815 - 9,348$$

$$\alpha_p - 0,025 \longrightarrow 8 - 9,348$$

$$(\alpha_p - 0,025) \times (7,815 - 9,348) = (0,05 - 0,025) \times (8 - 9,348) \longrightarrow \alpha_p = 0,0469$$

Siendo $\alpha_p = 0,0469 < 0,05$ se rechaza la hipótesis nula, concluyendo que el estado general de salud está asociado al sexo.

b) H_0 : X_2 e X_3 son independientes

En $\chi^2(X_2, X_3) = 4,5$ el número de grados de libertad es $(2 - 1) \times (3 - 1) = 2$

$\alpha_p = P(\chi_{p,2}^2 \geq 4,5)$. Interpolando en la tabla de la Chi-cuadrado:

0,90	α_p	0,10
0,211	4,5	4,605

$$0,90 - 0,10 \longrightarrow 0,211 - 4,605$$

$$\alpha_p - 0,10 \longrightarrow 4,5 - 4,605$$

$$(\alpha_p - 0,10) \times (0,211 - 4,605) = (0,90 - 0,10) \times (4,5 - 4,605) \longrightarrow \alpha_p = 0,119$$

Siendo $\alpha_p = 0,119 > 0,05$ se acepta la hipótesis nula, concluyendo que el sexo es independiente del nivel del ejercicio diario.

c) $H_0: X_1$ e X_3 son independientes

En $\chi^2(X_1, X_3) = 6,1$ el número de grados de libertad es $(4 - 1) \times (3 - 1) = 6$

$\alpha_p = P(\chi^2_{p,6} \geq 6,1)$. Interpolando en la tabla de la Chi-cuadrado:

0,90	α_p	0,10	$0,90 - 0,10 \longrightarrow 2,204 - 10,645$
2,204	6,1	10,645	$\alpha_p - 0,10 \longrightarrow 6,1 - 10,645$

$$(\alpha_p - 0,10) \times (2,204 - 10,645) = (0,90 - 0,10) \times (6,1 - 10,645) \rightarrow \alpha_p = 0,530$$

Siendo $\alpha_p = 0,530 > 0,05$ se acepta la hipótesis nula, concluyendo que el estado general de salud es independiente del nivel del ejercicio diario.

33. La probabilidad de que un banco reciba un cheque sin fondos es 0.01

- Si en una hora reciben 20 cheques, ¿cuál es la probabilidad de que tenga algún cheque sin fondos?
- El banco dispone de 12 sucursales en la ciudad, ¿cuál es la probabilidad de que al menos cuatro sucursales reciban algún cheque sin fondos?
- La media del valor de los cheques sin fondos es de 600 euros. Sabiendo que el banco trabaja 6 horas diarias, ¿qué cantidad no se espera pagar?
- Si se computasen los 500 primeros cheques, ¿cuál es la probabilidad de recibir entre 3 y 6 (inclusive) cheques sin fondos?

Solución:

a) $X =$ "Número de cheques sin fondos" con $X \sim B(20, 0,01)$

$$P[X \geq 1] = 1 - P[X < 1] = 1 - P[X = 0] = 1 - \binom{20}{0} \cdot 0,01^0 \cdot 0,99^{20} = 1 - 0,980 = 0,182$$

b) $Y =$ "Número de sucursales que reciben al menos 1 cheque sin fondos"

$$Y \sim B(12, 0,182) \quad P[Y \geq 4] = 1 - P[Y < 4] = 1 - [P[X = 0] + P[X = 1] + P[X = 2] + P[X = 3]] =$$

$$= 1 - \left[\binom{12}{0} \cdot 0,182^0 \cdot 0,818^{12} + \binom{12}{1} \cdot 0,182^1 \cdot 0,818^{11} + \binom{12}{2} \cdot 0,182^2 \cdot 0,818^{10} + \binom{12}{3} \cdot 0,182^3 \cdot 0,818^9 \right] = 1 - [0,0897 + 0,2396 + 0,2932 + 0,2174] = 0,16$$

c) $\frac{1 \text{ hora}}{20 \text{ cheques}} = \frac{6 \text{ horas}}{n \text{ cheques}} \mapsto n = 120 \text{ cheques}$

Los cheques sin fondos esperados: $\mu = E(X) = n \cdot p = 120 \cdot 0,01 = 1,2 \text{ cheques}$

En consecuencia, se espera no pagar $1,2 \times 600 = 720 \text{ euros}$

d) $U =$ "Número de cheques sin fondos computados" donde $U \sim B(500, 0,01)$, que al ser $n \cdot p = 500 \cdot 0,01 = 5$ se aproxima a una distribución de Poisson de parámetro $P[\lambda = 5]$

$$P[3 \leq U \leq 6] = P[U=3] + P[U=4] + P[U=5] + P[U=6] =$$

$$= \left[\frac{5^3}{3!} + \frac{5^4}{4!} + \frac{5^5}{5!} + \frac{5^6}{6!} \right] \cdot e^{-5} = [20,833 + 26,042 + 26,042 + 21,701] \cdot e^{-5} = 0,6375$$

34. Una compañía de seguros garantiza pólizas de seguros individuales contra retrasos aéreos de más de doce horas. Una encuesta ha permitido estimar a lo largo de un año que cada persona tiene una probabilidad de una de cada de mil de ser víctima de un retraso aéreo que esté cubierto por este tipo de póliza y que la compañía aseguradora podrá vender una media de cuatro mil pólizas al año.

Se pide hallar las siguientes probabilidades:

- a) Que el número de retrasos cubiertos por la póliza no pase de cuatro por año
- b) Número de retrasos esperados por año
- c) Que el número de retrasos sea superior a dos por año
- d) Que ocurran doce retrasos por año

Solución:

Sea $X =$ "Número de retrasos por año", la variable sigue una distribución binomial

$$n = 4000, p = \frac{1}{1000} = 0,001, B(4000, 0,001)$$

$$\text{con lo que, } P(X=k) = \binom{4000}{k} \cdot 0,001^k \cdot 0,999^{4000-k} \quad k = 0,1,\dots,4000$$

Es necesario buscar una distribución que sea una buena aproximación de ésta. La distribución de Poisson es una buena aproximación de la binomial $B(4000, 0,001)$, ya que $p = 0,001$ es muy pequeña y $n.p = 4000.0,001 = 4 < 5$.

Por tanto, $X \sim B(4000, 0,001) \approx X \sim P(\lambda = n.p = 4)$ $P(X = 4) = \frac{4^k}{k!} \cdot e^{-4}$

a) $P(X \leq 4) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) =$

$$= \left[\frac{4^0}{0!} + \frac{4^1}{1!} + \frac{4^2}{2!} + \frac{4^3}{3!} + \frac{4^4}{4!} \right] \cdot e^{-4} = [1 + 4 + 8 + 10,667 + 10,667] \cdot e^{-4} = 0,6289$$

b) El número de retrasos esperado por año es la media $\mu_x = \lambda = 4$

c) $P(X > 2) = 1 - P(X \leq 2) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)] =$

$$= 1 - \left[\frac{4^0}{0!} + \frac{4^1}{1!} + \frac{4^2}{2!} \right] \cdot e^{-4} = 1 - [1 + 4 + 8] \cdot e^{-4} = 1 - 0,381 = 0,7619$$

d) $P(X = 12) = \frac{4^{12}}{12!} \cdot e^{-4} = 0,035 \cdot e^{-4} = 0,00064$

35. El jefe de recursos humanos de una empresa realiza un test de diez ítems a los aspirantes a un puesto, teniendo en cada ítems cuatro posibles respuestas, de las que sólo una es correcta. Suponiendo que los aspirantes teniendo la misma probabilidad de responder. Se pide hallar las probabilidades para el aspirante:

- Conteste todos los ítems mal
- Conteste al menos cuatro ítems bien
- Conteste entre cuatro y seis ítems bien
- Conteste todos los ítems bien
- Conteste menos de tres ítems bien

Solución:

Sea $X =$ "Contestar ítems bien en el test", la variable sigue una distribución binomial

$$n = 10, p = \frac{1}{4} = 0,25, B(10, 0,25), P(X = k) = \binom{10}{k} \cdot 0,25^k \cdot 0,75^{10-k} \quad k = 0, 1, \dots, 10$$

a) $P(X = 0) = \binom{10}{0} \cdot 0,25^0 \cdot 0,75^{10} = 0,25^0 \cdot 0,75^{10} = 0,0563$

$$\begin{aligned} \text{b) } P(X \geq 4) &= 1 - P(X < 4) = 1 - (P(X=0) + P(X=1) + P(X=2) + P(X=3)) = \\ &= 1 - \left[\binom{10}{0} \cdot 0,25^0 \cdot 0,75^{10} + \binom{10}{1} \cdot 0,25^1 \cdot 0,75^9 + \binom{10}{2} \cdot 0,25^2 \cdot 0,75^8 + \binom{10}{3} \cdot 0,25^3 \cdot 0,75^7 \right] = \\ &= 1 - [0,0563 + 0,1877 + 0,2816 + 0,2503] = 0,2241 \end{aligned}$$

$$\begin{aligned} \text{c) } P(4 \leq X \leq 6) &= P(X=4) + P(X=5) + P(X=6) = \\ &= \binom{10}{4} \cdot 0,25^4 \cdot 0,75^6 + \binom{10}{5} \cdot 0,25^5 \cdot 0,75^5 + \binom{10}{6} \cdot 0,25^6 \cdot 0,75^4 = 0,1460 + 0,0584 + 0,0162 = 0,2206 \end{aligned}$$

$$\text{d) } P(X=10) = \binom{10}{10} \cdot 0,25^{10} \cdot 0,75^0 = 0$$

$$\begin{aligned} \text{e) } P(X < 3) &= P(X=0) + P(X=1) + P(X=2) = \\ &= \binom{10}{0} \cdot 0,25^0 \cdot 0,75^{10} + \binom{10}{1} \cdot 0,25^1 \cdot 0,75^9 + \binom{10}{2} \cdot 0,25^2 \cdot 0,75^8 = 0,0563 + 0,1877 + 0,2816 = 0,5256 \end{aligned}$$

36. Una fábrica aeronáutica produce en cada turno 100000 bolas para rodamientos, siendo la probabilidad de bola defectuosa 0,04. Las bolas se supervisan todas, depositando las defectuosas en un recipiente que se vacía al final de cada turno. ¿Cuántas bolas ha de contener el recipiente para que la probabilidad de que su capacidad no sea rebasada sea 0,95?

Solución:

Sea X = "número de bolas defectuosas entre 100000", donde $X \sim B(100000, 0,04)$, que como el número de bolas es muy grande se puede aproximar por una normal de media

$\mu = 100000 \cdot 0,04 = 4000$ y desviación típica $\sigma = \sqrt{100000 \cdot 0,04 \cdot 0,96} = 61,96$, es decir,

$$X \sim N[4000, 61,96]$$

La capacidad del recipiente C verifica $P[X < C] = 0,95$, con lo cual:

$$P[X < C] = P\left[\frac{X - 4000}{61,96} < \frac{C - 4000}{61,96}\right] = P\left[z < \frac{C - 4000}{61,96}\right] = 0,95 \quad \mapsto \quad P\left[z \geq \frac{C - 4000}{61,96}\right] = 0,05$$

$$\frac{C - 4000}{61,96} = 1,645 \quad \Rightarrow \quad C = 4000 + 1,645 \cdot 61,96 = 4101,92 \approx 4102 \text{ bolas}$$

37. El 7% de los pantalones de una determinada marca salen con algún defecto. Se empaquetan en caja de 80 pantalones para diferentes tiendas. ¿Cuál es la probabilidad de que en una caja haya entre 8 y 10 pantalones defectuosos?

Solución:

Sea $X =$ " Número de pantalones defectuosos en una caja "

Se trata de una distribución binomial (los pantalones son o no son defectuosos), es decir, una binomial con $n = 80$, $p = 0,07$: $B(80, 0,07)$, donde:

$$\mu = n \cdot p = 80 \cdot 0,07 = 5,6 \quad \sigma = \sqrt{n \cdot p \cdot q} = \sqrt{80 \cdot 0,07 \cdot 0,93} = 2,28$$

Adviértase que se dan las condiciones para aproximar la distribución discreta binomial a una distribución continua normal: $p = 0,07 \leq 0,5$ y $n \cdot p = 80 \cdot 0,07 = 5,6 > 5$

$$\text{con lo que, } B(n, p) \sim N\left[\mu = n \cdot p, \sigma = \sqrt{n \cdot p \cdot q}\right] \rightarrow B(80, 0,07) \sim N[5,6, 2,28]$$

Para utilizar correctamente la transformación de una variable aleatoria discreta X (distribución binomial) en una variable aleatoria continua z (con distribución normal) es necesario hacer una corrección de continuidad:

$$\begin{aligned} P[8 \leq X \leq 10] &\stackrel{\text{TRANSFORMACIÓN}}{=} P[7,5 \leq X^* \leq 10,5] \stackrel{\text{TIPIFICANDO}}{=} N(5,6, 2,28) \\ &= P\left[\frac{7,5 - 5,6}{2,28} \leq \frac{X^* - 5,6}{2,28} \leq \frac{10,5 - 5,6}{2,28}\right] = P[0,83 \leq z \leq 2,15] = \\ &= P[z > 0,83] - P[z > 2,15] = 0,2033 - 0,0158 = 0,1875 \end{aligned}$$

38. Un servicio dedicado a la reparación de electrodomésticos recibe por término medio 15 llamadas diarias. Determinar la probabilidad de que reciba un día más de 20 llamadas.

Solución:

Sea $X =$ " Número de llamadas recibidas al día "

$$\text{La variable aleatoria: } X \sim P[\lambda = 15] \rightarrow P[X = k] = \frac{15^k}{k!} \cdot e^{-15}$$

$$P[\lambda = 15] \xrightarrow{\lambda = 15 > 10} N[15, \sqrt{15}]$$

$$P[X > 20] = P\left[\frac{X - 15}{\sqrt{15}} > \frac{(20 - 0,5) - 15}{\sqrt{15}}\right] = P[z > 1,16] = 0,1230$$

39. En una fábrica se sabe que la probabilidad de que r artículos sean defectuosos es

$P[X = k] = \frac{4^k \cdot e^{-4}}{k!}$. Determinar la probabilidad de que en 100 días el número de artículos defectuosos esté comprendido entre (400, 600)

Solución:

Se trata de una distribución de Poisson: $P[X = k] = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \rightarrow \lambda = 4, \sigma = \sqrt{4} = 2$

En 100 días: $[X_1, X_2, \dots, X_{100}] \sim P(n \cdot \lambda, \sqrt{n \cdot \lambda}) = P(100 \cdot 4, \sqrt{100 \cdot 4}) = P(400, 20)$

$$P[n \cdot \lambda] \xrightarrow{n \cdot \lambda = 400 > 10} N[400, \sqrt{400}] \equiv N(400, 20)$$

$$P[400 \leq X \leq 600] = P\left[\frac{400 - 400}{20} \leq \frac{X - 400}{20} \leq \frac{600 - 400}{20}\right] = P[0 \leq z \leq 10] =$$

$$= P[z \geq 0] - P[z \geq 10] = 0,5$$

40. Se ha realizado una muestra aleatoria simple (m.a.s.) de tamaño 10 de una población considerada normal, llegando a que la varianza muestral es 4.

Calcular la probabilidad $P[|\bar{x} - \mu| \leq 1,51]$

Solución:

$$X \sim N(\mu, \sigma) \quad n\sigma_x^2 = (n-1)s_x^2 \rightarrow \frac{\sigma_x}{\sqrt{n-1}} = \frac{s_x}{\sqrt{n}}$$

$n = 10 < 30$
 $\sigma_x = 2$

$$t_{n-1} = \frac{\bar{x} - \mu}{\sigma_x / \sqrt{n-1}}$$

$$t_{10-1} = \frac{\bar{x} - \mu}{2 / \sqrt{10-1}} = \frac{\bar{x} - \mu}{2/3}$$

$$P[|\bar{x} - \mu| \leq 1,51] = P\left[\frac{|\bar{x} - \mu|}{2/3} \leq \frac{1,51}{2/3}\right] = P[|t_9| \leq 2,265] = P[-2,265 \leq t_9 \leq 2,265] =$$

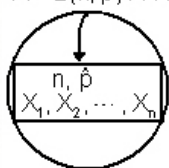
$$= P[t_9 \geq -2,265] - P[t_9 \geq 2,265] = P[t_9 \leq 2,265] - P[t_9 \geq 2,265] = 1 - 2P[t_9 \geq 2,265] = 0,95$$

41. Se realiza una encuesta para conocer la proporción de españoles a los que no le gusta el fútbol, tomando una muestra aleatoria simple (m.a.s.) de tamaño 100.

Por análisis anteriores se conoce que dicha proporción es del 45%. Calcular la probabilidad de que la proporción muestral sea superior al 52%.

Solución:

$$X \sim B(n, p) \mapsto N(np, \sqrt{npq})$$



Tamaño muestral $n = 100$

$p = 0,45$ y $q = 0,55$

$$\hat{p} \sim N\left(0,45, \sqrt{\frac{0,45 \times 0,55}{100}}\right) = N(0,45, 0,05)$$

$$P[\hat{p} > 0,52] = P\left[\frac{\hat{p} - 0,45}{0,05} > \frac{0,52 - 0,45}{0,05}\right] = P(z > 1,4) = 0,0808$$

42. Calcular la media, varianza y coeficiente de variación de la variable aleatoria discreta que tiene como función de distribución:

$$F(x) = \begin{cases} 0 & x < 2 \\ 0,2 & 2 \leq x < 4 \\ 0,55 & 4 \leq x < 6 \\ 0,85 & 6 \leq x < 8 \\ 1 & x \geq 8 \end{cases}$$

Solución:

La ley de probabilidad o función de cuantía:

$X = x_i$	2	4	6	8
$P(X = x_i)$	0,2	0,35	0,30	0,15

Adviértase que la función de distribución $F(x)$ es una función acumulativa, por tanto:

$$P(X = 2) = F(2) - F(0) = 0,2$$

$$P(X = 4) = F(4) - F(2) = 0,55 - 0,2 = 0,35$$

$$P(X = 6) = F(6) - F(4) = 0,85 - 0,55 = 0,30 \quad P(X = 8) = F(8) - F(6) = 1 - 0,85 = 0,15$$

Cálculo de la esperanza matemática y varianza

$X = x_i$	2	4	6	8	
$P(X = x_i)$	0,2	0,35	0,30	0,15	
$x_i \cdot P(X = x_i)$	0,4	1,4	1,8	1,2	$\sum_{i=1}^4 x_i \cdot P(X = x_i) = 4,8$
$x_i^2 \cdot P(X = x_i)$	0,8	5,6	10,8	9,6	$\sum_{i=1}^4 x_i^2 \cdot P(X = x_i) = 26,8$

Media: $\alpha_1 = \mu_x = E(X) = \sum_{i=1}^4 x_i \cdot P(X = x_i) = 4,8$ $\alpha_2 = E(X^2) = \sum_{i=1}^4 x_i^2 \cdot P(X = x_i) = 26,8$

Varianza: $\sigma_x^2 = \text{Var}(X) = \alpha_2 - \alpha_1^2 = 26,8 - 4,8^2 = 3,76$

Desviación típica: $\sigma_x = \sqrt{3,76} = 1,94$

Coefficiente variación: $CV = \frac{\sigma_x}{\mu_x} = \frac{1,94}{4,8} = 0,40$

43. La variable aleatoria discreta X tiene como distribución de probabilidad

$X = x_i$	1	2	3	4
$P(X = x_i)$	0,30	0,25	0,10	0,35

Se realiza un cambio de origen hacia la izquierda de dos unidades y un cambio de escala de tres unidades. Calcular y razonar propiedades:

- Media y varianza de la X .
- Media, varianza y coeficiente de variación de la variable transformada por el cambio de origen.
- Media, varianza y coeficiente de variación de la variable transformada por el cambio de escala.
- Media, varianza y coeficiente de variación de la variable transformada por el cambio de origen y escala.

Solución:

$X = x_i$	$P(X = x_i) = p_i$	$x_i \cdot p_i$	x_i^2	$x_i^2 \cdot p_i$
$x_1 = 1$	0,30	0,30	1	0,30
$x_2 = 2$	0,25	0,50	4	1,00
$x_3 = 3$	0,10	0,30	9	0,90
$x_4 = 4$	0,35	1,40	16	5,60
	1	2,5		7,8

$$\text{Media: } \alpha_1 = \mu_X = E(X) = \sum_{i=1}^4 x_i \cdot P(X = x_i) = \sum_{i=1}^4 x_i \cdot p_i = 2,5$$

$$\alpha_2 = E(X^2) = \sum_{i=1}^4 x_i^2 \cdot P(X = x_i) = \sum_{i=1}^4 x_i^2 \cdot p_i = 7,8$$

$$\text{Varianza: } \sigma_X^2 = \alpha_2 - \alpha_1^2 = 7,8 - 2,5^2 = 1,55$$

$$\text{Desviación típica: } \sigma_X = \sqrt{1,55} = 1,245$$

$$\text{Coeficiente de variación: } CV_X = \frac{\sigma_X}{\mu_X} = \frac{1,245}{2,5} = 0,498$$

b) Sea Y la variable transformada, al realizar un cambio de origen hacia la izquierda de dos unidades hay que restar 2, quedando: $Y = X - 0' = X - (-2) = X + 2$.

$$\text{Media: } \mu_Y = E(Y) = E[X + 2] = E(X + 2) = E(X) + 2 \mapsto \mu_Y = E(Y) = 2,5 + 2 = 4,5$$

Varianza: $\sigma_Y^2 = \text{Var}[X + 2] = \text{Var}(X) + \text{Var}(2) = \sigma_X^2 + 0 = \sigma_X^2 \mapsto \sigma_Y^2 = 1,55$

Desviación típica: $\sigma_Y = \sqrt{1,55} = 1,245$

Coefficiente de variación: $CV_Y = \frac{\sigma_Y}{\mu_Y} = \frac{\sigma_X}{\mu_X + 2} = \frac{1,245}{4,5} = 0,28 \neq CV_X$

El cambio de origen afecta a la media y, en consecuencia, al coeficiente de variación.

c) Al realizar un cambio de escala de 3 unidades, la variable transformada es $Y = \frac{X}{3}$

Media: $\mu_Y = E(Y) = E\left[\frac{X}{3}\right] = \frac{1}{3} \cdot E(X) \mapsto \mu_Y = \frac{1}{3} \cdot \mu_X = \frac{2,5}{3}$

Varianza: $\sigma_Y^2 = \text{Var}\left[\frac{X}{3}\right] = \frac{1}{9} \cdot \text{Var}(X) = \frac{1}{9} \cdot \sigma_X^2 \mapsto \sigma_Y^2 = \frac{1}{9} \cdot 1,55 = \frac{1,55}{9}$

Desviación típica: $\sigma_Y = \sqrt{\frac{1,55}{9}} = \frac{1}{3} \cdot \sqrt{1,55} = \frac{1}{3} \cdot \sigma_X$

Coefficiente de variación: $CV_Y = \frac{\sigma_Y}{\mu_Y} = \frac{\frac{1}{3} \cdot \sigma_X}{\frac{1}{3} \cdot \mu_X} = \frac{\sigma_X}{\mu_X} = CV_X = 0,498$

El cambio de escala afecta a la media y a la desviación típica de la misma forma, en consecuencia deja invariante al coeficiente de variación.

Resultados que se observan en la tabla, donde $Y = \frac{X}{3}$

$Y = y_j$	$P(Y = y_j) = p_j$	$y_j \cdot p_j$	y_j^2	$y_j^2 \cdot p_j$
$x_1 = 1/3$	0,30	0,1	1/9	0,3/9
$x_2 = 2/3$	0,25	0,5/3	4/9	1/9
$x_3 = 1$	0,10	0,1	1	0,1
$x_4 = 4/3$	0,35	1,4/3	16/9	5,6/9
	1	2,5/3		7,8/9

Media: $\alpha_1 = \mu_Y = E(Y) = \sum_{j=1}^4 y_j \cdot P(Y = y_j) = \sum_{j=1}^4 y_j \cdot p_j = \frac{2,5}{3} = \frac{1}{3} \cdot \mu_X$

$$\alpha_2 = E(Y^2) = \sum_{j=1}^4 y_j^2 \cdot P(Y = y_j) = \sum_{j=1}^4 y_j^2 \cdot p_j = \frac{7,8}{9} = \frac{1}{9} \cdot E(Y^2)$$

$$\text{Varianza: } \sigma_Y^2 = \alpha_2 - \alpha_1^2 = \frac{7,8}{9} - \left(\frac{2,5}{3}\right)^2 = \frac{1}{9} \cdot \sigma_X^2 = \frac{1,55}{9}$$

$$\text{Desviación típica: } \sigma_Y = \sqrt{\frac{1,55}{9}} = \frac{1}{3} \cdot \sqrt{1,55} = \frac{1}{3} \cdot \sigma_X$$

$$\text{Coeficiente de variación: } CV_Y = \frac{\sigma_Y}{\mu_Y} = \frac{\frac{1}{3} \cdot \sigma_X}{\frac{1}{3} \cdot \mu_X} = \frac{\sigma_X}{\mu_X} = CV_X = 0,498$$

d) Al realizar simultáneamente un cambio de origen de 2 unidades a la izquierda y un cambio de escala de 3 unidades, la variable transformada es $Y = \frac{X+2}{3}$

$$\text{Media: } \mu_Y = E(Y) = E\left[\frac{X+2}{3}\right] = \frac{1}{3} \cdot E(X+2) = \frac{1}{3} \cdot E(X) + \frac{2}{3}$$

$$\text{con lo que, } \mu_Y = E(Y) = \frac{1}{3} \cdot E(X) + \frac{2}{3} = \frac{1}{3} \cdot 2,5 + \frac{2}{3} = \frac{4,5}{3} = 1,5$$

$$\text{Varianza: } \sigma_Y^2 = \text{Var}(Y) = \text{Var}\left[\frac{X+2}{3}\right] = \frac{1}{9} \cdot \text{Var}(X+2) = \frac{1}{9} \cdot \text{Var}(X) = \frac{1}{9} \cdot \sigma_X^2$$

$$\text{Desviación típica: } \sigma_Y = \sqrt{\frac{1,55}{9}} = \frac{1}{3} \cdot \sqrt{1,55} = \frac{1}{3} \cdot \sigma_X$$

$$\text{Coeficiente de variación: } CV_Y = \frac{\sigma_Y}{\mu_Y} = \frac{\frac{1}{3} \cdot \sigma_X}{\frac{1}{3} \cdot \mu_X + \frac{2}{3}} = \frac{\sigma_X}{\mu_X + 2} = \frac{1,245}{4,5} = 0,28 \neq CV_X$$

El cambio de origen y de escala afecta a la media y desviación típica de distinta forma, en consecuencia también queda afectado el coeficiente de variación.

Resultados que se observan en la tabla, donde $Y = \frac{X+2}{3}$

$Y = y_j$	$P(Y = y_j) = p_j$	$y_j \cdot p_j$	y_j^2	$y_j^2 \cdot p_j$
$x_1 = 1$	0,30	0,30	1	0,30
$x_2 = 4/3$	0,25	1/3	16/9	4/9
$x_3 = 5/3$	0,10	0,5/3	25/9	2,5/9
$x_4 = 2$	0,35	0,70	4	1,4
	1	4,5/3		21,8/9

Media: $\alpha_1 = \mu_Y = E(Y) = \sum_{j=1}^4 y_j \cdot P(Y = y_j) = \sum_{j=1}^4 y_j \cdot p_j = \frac{4,5}{3} = 1,5$

$\alpha_2 = E(Y^2) = \sum_{j=1}^4 y_j^2 \cdot P(Y = y_j) = \sum_{j=1}^4 y_j^2 \cdot p_j = \frac{21,8}{9}$

Varianza: $\sigma_Y^2 = \alpha_2 - \alpha_1^2 = \frac{21,8}{9} - \left(\frac{4,5}{3}\right)^2 = \frac{1}{9} \cdot \sigma_x^2 = \frac{1,55}{9}$

Desviación típica: $\sigma_Y = \sqrt{\frac{1,55}{9}} = \frac{1}{3} \cdot \sqrt{1,55} = \frac{1}{3} \cdot \sigma_x$

Coeficiente de variación: $CV_Y = \frac{\sigma_Y}{\mu_Y} = \frac{\frac{1}{3} \cdot \sigma_x}{\frac{1}{3} \cdot 4,5} = \frac{\sigma_x}{4,5} = \frac{1,245}{4,5} = 0,28 \neq CV_x$

44. Un dispositivo está formado por muchos elementos que trabajaban independientemente, siendo la probabilidad de fallo durante la primera hora de trabajo muy pequeña e iguales en todos los elementos.

Si la probabilidad de que en ese tiempo falle por lo menos un elemento es 0,98.

- Hallar la media y desviación típica del número de elementos que fallen en la primera hora.
- Calcular la probabilidad de que fallen a lo sumo dos elementos en ese tiempo.

Solución:

a) Habiendo muchos elementos independientes, el tamaño (n) es muy grande, con una probabilidad de fallo muy pequeña, condiciones para que el dispositivo siga una distribución de Poisson.

Sea la variable aleatoria $X = \text{"Fallo en la primera hora de un elemento"}$, $X \sim P(\lambda)$ donde

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

$$P(X \geq 1) = 0,98 \rightarrow P(X < 1) = P(X = 0) = \frac{\lambda^0}{0!} e^{-\lambda} = 0,02 \rightarrow e^{-\lambda} = 0,02$$

$$-\lambda = \ln 0,02 \mapsto \lambda = 3,91$$

$$\mu = \lambda = 3,91 \quad \sigma = \sqrt{3,91} = 1,98$$

$$\begin{aligned} \text{b) } P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) = \frac{3,9^0}{0!} e^{-3,9} + \frac{3,9}{1!} e^{-3,9} + \frac{3,9^2}{2!} e^{-3,9} = \\ &= \left[1 + 3,9 + \frac{3,9^2}{2} \right] e^{-3,9} = 12,505 \times e^{-3,9} = 0,2531 \end{aligned}$$

45. Para aprobar la asignatura de estadística teórica se realiza un test con veinte ítems. Sabiendo que una persona determinada tiene una probabilidad de 0,8 de contestar bien cada ítem. Se pide:

- Probabilidad de que la primera pregunta que contesta bien sea la tercera que hace.
- Para aprobar el test es necesario contestar diez ítems bien. ¿Cuál es la probabilidad de que apruebe al contestar el doceavo ítem?

Solución:

- La variable $X =$ "Número de ítems que tiene que hacer hasta que responda uno bien" sigue una distribución de Pascal o geométrica, siendo $p = 0,8$, $q = 0,2$. $X \sim G(0,8)$

$$P(X = 3) = q^2 \cdot p = 0,2^2 \cdot 0,8 = 0,032$$

- La variable $X =$ "Número de ítems que tiene que realizar hasta contestar 10 bien" sigue una distribución binomial negativa

$$X \sim \text{Bn}(12, 0,8) \quad k = 10 \quad p = 0,8 \quad q = 0,2$$

$$P(X = 12) = \binom{12-1}{10-1} 0,8^{10} \cdot 0,2^2 = \binom{11}{9} 0,8^{10} \cdot 0,2^2 = \frac{11!}{2!9!} 0,8^{10} \cdot 0,2^2 = \frac{11 \cdot 10}{2} 0,8^{10} \cdot 0,2^2 = 0,24$$

46. En una caja hay 5 triángulos, 3 círculos y 2 rectángulos. Realizando extracciones con reemplazamiento, se piden las siguientes probabilidades:

- Al realizar 8 extracciones, se obtengan en 4 ocasiones un círculo.
- Se necesiten 8 extracciones para obtener 4 círculos.
- Que aparezca el primer círculo en la 8 extracción.
- Al realizar 8 extracciones aparezcan 3 triángulos, 3 círculos y 2 rectángulos.
- Al realizar 6 extracciones sin reemplazamiento aparezcan en 2 ocasiones un círculo.

Solución:

- a) Hay dos situaciones (círculo, no-círculo), se trata de una distribución binomial,

$$B(8, 0,3) \quad n=8 \quad p=3/10=0,3$$

$$P(X=4)=\binom{8}{4} 0,3^4 0,7^4 = \frac{8!}{4!4!} 0,21^4 = \frac{8 \cdot 7 \cdot 6 \cdot 5}{4 \cdot 3 \cdot 2 \cdot 1} 0,21^4 = 70 \cdot 0,21^4 = 0,136$$

- b) $X =$ " Número de extracciones hasta que aparece la cuarta extracción de círculo"

Es una binomial negativa: $X \sim Bn(8, 0,3) \quad n=8 \quad k=4 \quad p=0,3 \quad q=0,7$

$$P(X=8)=\binom{8-1}{4-1} 0,3^4 \cdot 0,7^4 = \binom{7}{3} 0,21^4 = \frac{7!}{4!3!} 0,21^4 = \frac{7 \cdot 6 \cdot 5}{3 \cdot 2} 0,21^4 = 0,0681$$

- c) $X =$ " Aparece el primer círculo en la octava extracción"

Es una distribución geométrica o de Pascal $X \sim G(0,3)$

$$P(X=8)=q^{8-1} \cdot p=0,7^7 \cdot 0,3=0,0247$$

- d) $X =$ " Número de veces que se extrae triángulo, círculo o rectángulo en ocho extracciones".

Se trata de una distribución polinomial, es decir, en cada prueba se consideran k sucesos independientes.

$$P(T=3; C=3; R=2)=\frac{8!}{3!3!2!} \left(\frac{5}{10}\right)^3 \left(\frac{3}{10}\right)^3 \left(\frac{2}{10}\right)^2 = 560 \cdot 0,5^3 \cdot 0,3^3 \cdot 0,2^2 = 0,0756$$

- e) Hay dos situaciones excluyentes (círculo, no-círculo).

$X =$ " Número de veces que se extrae círculo en una muestra de tamaño ocho"

Distribución hipergeométrica: $N=10 \quad n=6 \quad k=2 \quad p=0,3 \quad q=0,7 \quad Np=3 \quad Nq=7$

$$P(X=k) = \frac{\binom{N \cdot p}{k} \binom{N \cdot q}{n-k}}{\binom{N}{n}} \mapsto P(X=2) = \frac{\binom{3}{2} \binom{7}{4}}{\binom{10}{6}} = \frac{3 \cdot 35}{210} = 0,5$$

47. Por prescripción médica, un enfermo debe hacer una toma de tres píldoras de un determinado medicamento. De las doce píldoras que contiene el envase hay cuatro en malas condiciones. Se pide:

- Probabilidad de que tome sólo una buena.
- Probabilidad de que de las tres píldoras de la toma al menos una esté en malas condiciones.
- ¿Cuál es el número de píldoras que se espera tome el enfermo en buenas condiciones en cada toma?
- Si existe otro envase que contenga cuarenta píldoras, de las que diez se encuentran en malas condiciones. ¿qué envase sería más beneficiosos para el enfermo?

Solución:

- Hay dos situaciones excluyentes (buena, no-buena).

La variable $X = \text{"número de píldoras buenas al tomar tres"}$ sigue una distribución hipergeométrica $X \sim H(3, 12, 8)$, en donde

$$N = 12 \text{ píldoras : } 8 \text{ buenas} \rightarrow p = 8/12 = 2/3 \quad 4 \text{ malas} \rightarrow q = 4/12 = 1/3$$

$$N \cdot p = 12 \cdot \frac{2}{3} = 8 \quad N \cdot q = 12 \cdot \frac{1}{3} = 4 \quad n = 3$$

$$P(X=k) = \frac{\binom{N \cdot p}{k} \binom{N \cdot q}{n-k}}{\binom{N}{n}} \rightarrow P(X=1) = \frac{\binom{8}{1} \binom{4}{2}}{\binom{12}{3}} = \frac{8 \cdot \frac{4!}{2!2!}}{\frac{12!}{3!9!}} = \frac{24}{110} = 0,22$$

- La probabilidad de que al menos una esté en malas condiciones equivale a la probabilidad de que a lo sumo dos píldoras sean buenas. Por tanto:

$$P(X \leq 2) = P(X=0) + P(X=1) + P(X=2) = \frac{2}{110} + \frac{24}{110} + \frac{56}{110} = \frac{82}{110} = 0,75$$

$$P(X=0) = \frac{\binom{8}{0} \binom{4}{3}}{\binom{12}{3}} = \frac{1 \cdot 4}{\frac{12!}{3!9!}} = \frac{4 \cdot 3!9!}{12!} = \frac{4 \cdot 3 \cdot 2}{12 \cdot 11 \cdot 10} = \frac{2}{110} = 0,02$$

$$P(X=2) = \frac{\binom{8}{2} \binom{4}{1}}{\binom{12}{3}} = \frac{\frac{8!}{2!6!} \cdot 4}{\frac{12!}{3!9!}} = \frac{\frac{8 \cdot 7}{2} \cdot 4 \cdot 3!9!}{12!} = \frac{56}{110} = 0,51$$

c) El número de píldoras que se espera que tome es la media de la distribución:

$$\mu_x = n \cdot p = 3 \cdot \frac{2}{3} = 2 \text{ píldoras}$$

d) Para tomar una decisión, hay que calcular el número de píldoras esperado en buenas condiciones al tomar tres del segundo envase.

$$N = 40 \text{ píldoras: } 30 \text{ buenas} \rightarrow p = 30/40 = 3/4 = 0,75 \quad 10 \text{ malas} \rightarrow q = 10/40 = 1/4 = 0,25$$

$$\mu_y = n \cdot p = 3 \cdot 0,75 = 2,25 \text{ píldoras}$$

El segundo envase es más beneficioso para el enfermo.

48. El consumo familiar de cierto artículo se distribuye uniformemente con esperanza 10 y varianza unidad. Determinar la probabilidad de que el consumo de dicho artículo se encuentre comprendido entre 8 y 12 unidades.

Solución:

Sea X = "número de unidades consumidas del artículo", donde $X \sim U(10, 1)$ en el intervalo $[a, b]$

$$\text{Se tiene: } \mu = E(X) = \frac{a+b}{2} = 10 \quad \sigma^2 = \frac{(b-a)^2}{12} = 1$$

$$\begin{cases} a+b=20 \\ (b-a)^2=12 \end{cases} \mapsto [b-(20-b)]^2=12 \mapsto (2b-20)=\sqrt{12} \Rightarrow \begin{cases} b=10+\sqrt{3}=11,73 \\ a=10-\sqrt{3}=8,27 \end{cases}$$

La distribución es uniforme entre 8,27 y 11,73, en consecuencia la probabilidad de que el consumo del artículo se encuentre comprendido entre 8 y 12 es la unidad.

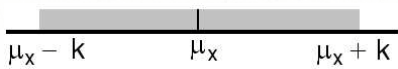
49. En un cine de verano hay instaladas 800 sillas, sabiendo que el número de asistentes es una variable aleatoria de media 600 y desviación típica 100.

¿Qué probabilidad existe de que el número de personas que vaya al cine un día cualquiera sea superior al número de sillas instaladas?

Solución:

Sea la variable aleatoria $X =$ "Número de sillas del cine", donde $\mu = 600$, $\sigma = 100$

Desigualdad de Chebyshev:

$$P[X > 800] < P[|X - \mu_x| > k] \leq \frac{\sigma^2}{k^2}$$


$$\mu_x + k = 800 \quad \mapsto \quad k = 800 - 600 = 200 \quad \rightarrow \quad P[X > 800] \leq \frac{100^2}{200^2} = \frac{1}{4} = 0,25$$

50. Se desea conocer el número de automóviles que se deben poner a la venta durante un periodo determinado para que se satisfaga una demanda media de 300 unidades con una desviación típica de 100 unidades, con una probabilidad no inferior al 75%.

Solución:

Sea la variable aleatoria $X =$ "Número de automóviles a la venta" $\mu = 300$, $\sigma = 100$

Según Chebyshev:

$$P[|X - \mu_x| \leq k] \geq 1 - \frac{\sigma^2}{k^2} \quad \longrightarrow \quad P[\mu_x - k \leq X \leq \mu_x + k] \geq 1 - \frac{\sigma^2}{k^2}$$

$$P[300 - k \leq X \leq 300 + k] \geq \overbrace{1 - \frac{100^2}{k^2}}^{0,75}$$

$$0,75 = 1 - \frac{100^2}{k^2} \quad \rightarrow \quad \frac{100^2}{k^2} = 0,25 \quad \rightarrow \quad k^2 = \frac{100^2}{0,25} \quad \rightarrow \quad k = \sqrt{\frac{100^2}{0,25}} = 200$$

$$300 + k = 300 + 200 = 500 \text{ automóviles}$$

51. La demanda media de un producto es de 100 unidades con una desviación típica de 40 unidades. Calcular la cantidad del producto que se debe tener a la venta para satisfacer la demanda de forma que puedan ser atendidos al menos el 80% de los clientes.

Solución:

$\mu = 100$, $\sigma = 40$ según Chebyshev:

$$P[|X - \mu_x| \leq k] \geq 1 - \frac{\sigma^2}{k^2} \longrightarrow P[\mu_x - k \leq X \leq \mu_x + k] \geq 1 - \frac{\sigma^2}{k^2}$$

$$P[100 - k \leq X \leq 100 + k] \geq 1 - \frac{40^2}{k^2} \quad \text{con } 0,80$$

$$0,80 = 1 - \frac{40^2}{k^2} \rightarrow \frac{40^2}{k^2} = 0,20 \rightarrow k^2 = \frac{40^2}{0,20} \rightarrow k = \sqrt{\frac{40^2}{0,20}} = 89,44$$

Se deben poner a la venta 90 unidades.

52. La variable discreta X tiene como distribución de probabilidad

$$P(X = k) = \frac{1}{10} \quad \text{siendo } k = 2, 3, \dots, 11$$

Se pide:

- | | |
|----------------------------|----------------------|
| a) Función de distribución | b) $P(X > 7)$ |
| c) $P(X < 5)$ | d) $P(3 \leq X < 7)$ |

Solución:

$$a) F(x) = P(X \leq x) = \frac{x-1}{10} \quad \text{siendo } x = 2, 3, \dots, 11$$

Adviértase que entre dos valores consecutivos de la variable, la función de distribución toma el valor menor.

$$b) P(X > 7) = 1 - P(X \leq 7) = 1 - F(7) = 1 - \frac{6}{10} = \frac{4}{10} = 0,4$$

$$\text{O bien, } P(X > 7) = P(X = 8) + P(X = 9) + P(X = 10) + P(X = 11) = \frac{4}{10} = 0,4$$

$$c) \quad P(X < 5) = F(5) = \frac{4}{10} = 0,4$$

$$\text{O bien, } P(X < 5) = P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) = \frac{4}{10} = 0,4$$

$$d) \quad P(3 \leq X < 7) = F(7) - F(3) = \frac{6}{10} - \frac{2}{10} = \frac{4}{10} = 0,4$$

$$\text{O bien, } P(3 \leq X < 7) = P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6) = \frac{4}{10} = 0,4$$

53. Al lanzar cuatro monedas se considera el número de escudos obtenidos. De la variable aleatoria X así obtenida, se pide:

- Ley de probabilidad. Representación gráfica
- Función de distribución. Representación gráfica
- Esperanza matemática y varianza
- Mediana y moda de la distribución
- Probabilidad de obtener más de uno y menos de tres escudos

Solución:

- a) Sea $X =$ 'Número de escudos en la tirada de cuatro monedas'

$$\Omega = \left\{ (c, c, c, c), (c, c, c, e), (c, c, e, c), (c, c, e, e), (c, e, c, c), (c, e, c, e), (e, c, c, c), (e, c, c, e), \right. \\ \left. (e, e, e, e), (e, e, e, c), (e, e, c, e), (e, e, c, c), (e, c, e, e), (e, c, e, c), (c, e, e, e), (c, e, e, c) \right\}$$

$$X(c, c, c, c) = 0$$

$$P(X = 0) = 1/16$$

$$X(c, c, c, e) = X(c, c, e, c) = X(c, e, c, c) = X(e, c, c, c) = 1$$

$$P(X = 1) = 4/16$$

$$X(c, c, e, e) = X(c, e, c, e) = X(e, c, e, c) =$$

$$P(X = 2) = 6/16$$

$$X(e, e, c, c) = X(e, c, e, c) = X(c, e, c, e) = 2$$

$$X(e, e, e, c) = X(e, e, c, e) = X(e, c, e, e) = X(c, e, e, e) = 3$$

$$P(X = 3) = 4/16$$

$$X(e, e, e, e) = 4$$

$$P(X = 4) = 1/16$$

La ley de probabilidad o función de cuantía:

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

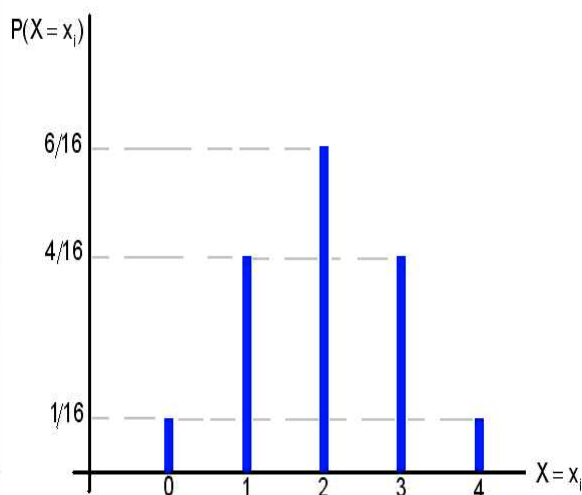
$X = x_i$	0	1	2	3	4
$P(X = x_i)$	1/16	4/16	6/16	4/16	1/16

b) Función de distribución:

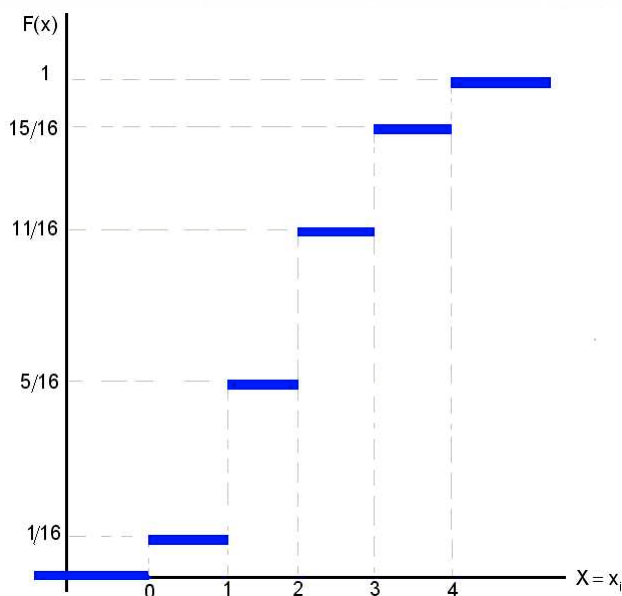
$X = x_i$	0	1	2	3	4
$P(X = x_i)$	1/16	4/16	6/16	4/16	1/16
$F(x) = P(X \leq x)$	1/16	5/16	11/16	15/16	1

$$F(x) = \begin{cases} 0 & x < 0 \\ 1/16 & 0 \leq x < 1 \\ 5/16 & 1 \leq x < 2 \\ 11/16 & 2 \leq x < 3 \\ 15/16 & 3 \leq x < 4 \\ 1 & x \geq 4 \end{cases}$$

Ley de Probabilidad



Función de distribución



c) Cálculo de la esperanza matemática y varianza

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

$X = x_i$	0	1	2	3	4	
$P(X = x_i)$	1/16	4/16	6/16	4/16	1/16	
$x_i \cdot P(X = x_i)$	0	4/16	12/16	12/16	4/16	$\sum_{i=1}^5 x_i \cdot P(X = x_i) = 2$
$x_i^2 \cdot P(X = x_i)$	0	4/16	24/16	36/16	16/16	$\sum_{i=1}^5 x_i^2 \cdot P(X = x_i) = 5$

Media: $\alpha_1 = \mu_x = E(X) = \sum_{i=1}^5 x_i \cdot P(X = x_i) = 2$ $\alpha_2 = E(X^2) = \sum_{i=1}^5 x_i^2 \cdot P(X = x_i) = 5$

Varianza: $\sigma_x^2 = \text{Var}(X) = \alpha_2 - \alpha_1^2 = 5 - 2^2 = 1$

d) Observando la ley de probabilidad la moda $M_d = 2$

Observando la función de distribución la mediana $M_e = 2$ por ser $F(x=2) = 11/16$ el primer valor que iguala o deja por debajo a 0,5

e) $P(1 < X < 3) = P(X = 2) = \frac{6}{16} = 0,375$ O bien $P(1 < X < 3) = F(2) - F(1) = \frac{11}{16} - \frac{5}{16} = \frac{6}{16}$

54. Sea (X, Y) una variable aleatoria bidimensional con función de densidad:

$$f(x, y) = \begin{cases} k & 0 < y < x < 1 \\ 0 & \text{restantes valores} \end{cases}$$

- Hallar k para que sea función de densidad.
- Hallar las funciones de densidad marginales. ¿Son X e Y independientes?
- Hallar las funciones de distribución marginales.
- Hallar las funciones de densidad condicionadas.

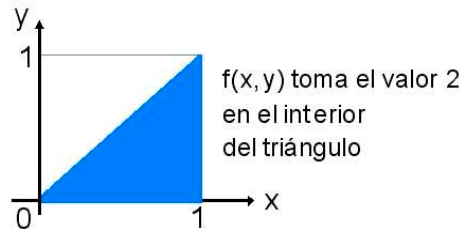
Solución:

a) Para que $f(x, y)$ sea función de densidad tiene que verificarse: $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$

$$\int_0^1 \int_0^x k \, dx \, dy = \int_0^1 k \left(\int_0^x dy \right) dx = \int_0^1 k [y]_0^x dx = k \int_0^1 x \, dx = k \left[\frac{x^2}{2} \right]_0^1 = \frac{k}{2} = 1 \rightarrow k = 2$$

Por tanto,

$$f(x, y) = \begin{cases} 2 & 0 < y < x < 1 \\ 0 & \text{restantes valores} \end{cases}$$



b) Funciones de densidad marginales:

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) \, dy = \int_0^x 2 \, dy = 2[y]_0^x = 2x \quad 0 < x < 1$$

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y) \, dx = \int_y^1 2 \, dx = 2[x]_y^1 = 2 - 2y \quad 0 < y < 1$$

X e Y son independientes cuando $f(x, y) = f_1(x) \cdot f_2(y)$

$$f_1(x) \cdot f_2(y) = 2x \cdot (2 - 2y) = 4x - 4xy \neq 2 = f(x, y) \quad \text{luego no son independientes}$$

b) Funciones de distribución marginales:

$$F_1(x) = \int_{-\infty}^x f_1(t) \, dt = \int_0^x f_1(t) \, dt = \int_0^x 2t \, dt = [t^2]_0^x = x^2 \quad 0 < x < 1$$

$$F_2(y) = \int_{-\infty}^y f_2(t) \, dt = \int_0^y f_2(t) \, dt = \int_0^y (2 - 2t) \, dt = [2t - t^2]_0^y = 2y - y^2 \quad 0 < y < 1$$

c) Funciones de densidad condicionadas:

$$f(x/y) = \frac{f(x, y)}{f_2(y)} = \frac{2}{2 - 2y} \quad 0 < y < 1 \quad 0 < x < 1$$

$$f(y/x) = \frac{f(x, y)}{f_1(x)} = \frac{2}{2x} \quad 0 < x < 1 \quad 0 < y < 1$$

55. Una variable aleatoria continua X tiene por función de distribución:

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{2} & 0 \leq x \leq 1 \\ 2x - \frac{x^2}{2} - 1 & 1 < x \leq 2 \\ 1 & x > 2 \end{cases}$$

Se pide:

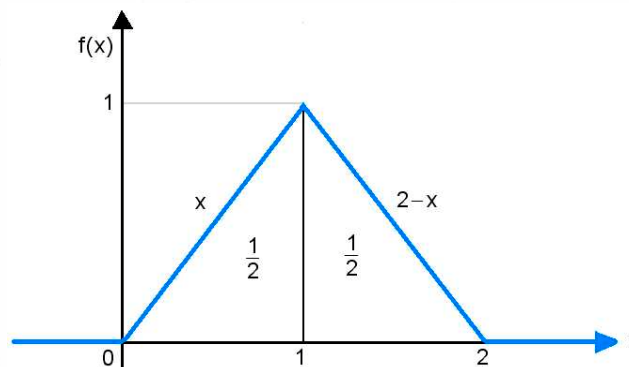
- Hallar la función de distribución y representarla
- Media, varianza, desviación típica y coeficiente de variación
- $P\left(\frac{1}{2} < X \leq \frac{3}{2}\right)$

Solución:

a) La función de densidad es la derivada de la función de distribución en los puntos donde exista la derivada, entonces:

$$f(x) = \frac{dF(x)}{dx} = \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 2 - x & 1 < x \leq 2 \\ 0 & x > 2 \end{cases}$$

$$f(x) = \begin{cases} x & 0 \leq x \leq 1 \\ 2 - x & 1 < x \leq 2 \\ 0 & \text{otros valores} \end{cases}$$



b) Media

$$\begin{aligned} \alpha_1 = \mu_x = E(X) &= \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x \cdot x \cdot dx + \int_1^2 x \cdot (2-x) \cdot dx = \int_0^1 x^2 dx + \int_1^2 (2x - x^2) \cdot dx = \\ &= \left[\frac{x^3}{3} \right]_0^1 + \left[x^2 - \frac{x^3}{3} \right]_1^2 = \frac{1}{3} + \left(4 - \frac{8}{3} \right) - \left(1 - \frac{1}{3} \right) = 1 \end{aligned}$$

Varianza: $\sigma_x^2 = \alpha_2 - \alpha_1^2$

$$\alpha_2 = E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^1 x^2 \cdot x \cdot dx + \int_1^2 x^2 \cdot (2-x) \cdot dx = \int_0^1 x^3 dx + \int_1^2 (2x^2 - x^3) \cdot dx =$$

$$= \left[\frac{x^4}{4} \right]_0^1 + \left[\frac{2x^3}{3} - \frac{x^4}{4} \right]_1^2 = \frac{1}{4} + \left(\frac{16}{3} - \frac{16}{4} \right) - \left(\frac{2}{3} - \frac{1}{4} \right) = \frac{14}{12} = \frac{7}{6}$$

$$\sigma_x^2 = \alpha_2 - \alpha_1^2 = \frac{7}{6} - 1^2 = \frac{1}{6}$$

Desviación típica: $\sigma_x = \sqrt{\frac{1}{6}} = 0,41$

Coeficiente variación: $CV_x = \frac{\sigma_x}{\mu_x} = \frac{0,41}{1} = 0,41$

c) $P\left(\frac{1}{2} < X \leq \frac{3}{2}\right) = F\left(\frac{3}{2}\right) - F\left(\frac{1}{2}\right) = \left(2 \cdot \frac{3}{2} - \frac{(3/2)^2}{2} - 1\right) - \left(\frac{(1/2)^2}{2}\right) = 3 - \frac{9}{8} - 1 - \frac{1}{8} = \frac{3}{4} = 0,75$

56. El flujo de demanda de teléfonos móviles (en miles a la hora) de una determinada fábrica se ajusta a una variable aleatoria con la siguiente función de densidad:

$$f(x) = \begin{cases} k - x & \text{si } 0 \leq x \leq 1 \\ 0 & \text{en el resto} \end{cases}$$

- a) Hallar k para que f(x) sea función de densidad
b) Hallar la función de distribución F(x)
c) Hallar la esperanza y la varianza

d) $P\left[\frac{1}{2} \leq x \leq \frac{3}{4}\right]$

Solución:

a) $\int_0^1 (k - x) dx = \left[kx - \frac{x^2}{2} \right]_0^1 = k - \frac{1}{2} = 1 \quad \mapsto \quad k = \frac{3}{2}$

Función de densidad: $f(x) = \begin{cases} \frac{3}{2} - x & \text{si } 0 \leq x \leq 1 \\ 0 & \text{en el resto} \end{cases}$

b) $0 \leq x < 1 \quad F(x) = \int_0^x f(t) dt = \int_0^x \left[\frac{3}{2} - t \right] dt = \left[\frac{3}{2}t - \frac{1}{2}t^2 \right]_0^x = \frac{3}{2}x - \frac{1}{2}x^2$

$$x \geq 1 \quad F(x) = \int_0^1 \left[\frac{3}{2} - t \right] dt + \int_1^x \left[\frac{3}{2} - t \right] dt = \left[\frac{3}{2}t - \frac{1}{2}t^2 \right]_0^1 = \frac{3}{2} - \frac{1}{2} = 1$$

$$\text{Función de distribución } F(x) = \begin{cases} 0 & \text{si } x < 0 \\ \frac{3}{2}x - \frac{x^2}{2} & \text{si } 0 \leq x < 1 \\ 1 & \text{si } x \geq 1 \end{cases}$$

$$c) \quad \mu = E(x) = \int_0^1 x \cdot f(x) dx = \int_0^1 x \cdot \left[\frac{3}{2} - x \right] dx = \int_0^1 \left[\frac{3}{2}x - x^2 \right] dx = \left[\frac{3}{4}x^2 - \frac{x^3}{3} \right]_0^1 = \frac{3}{4} - \frac{1}{3} = \frac{5}{12}$$

$$E(x^2) = \int_0^1 x^2 \cdot f(x) dx = \int_0^1 x^2 \cdot \left[\frac{3}{2} - x \right] dx = \int_0^1 \left[\frac{3}{2}x^2 - x^3 \right] dx = \left[\frac{3}{6}x^3 - \frac{x^4}{4} \right]_0^1 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

$$\sigma^2 = \text{Var}(x) = E(x^2) - [E(x)]^2 = \frac{1}{4} - \left(\frac{5}{12} \right)^2 = \frac{36 - 25}{144} = \frac{11}{144}$$

$$d) \quad P\left[\frac{1}{2} \leq x \leq \frac{3}{4}\right] = \int_{\frac{1}{2}}^{\frac{3}{4}} f(x) dx = \int_{\frac{1}{2}}^{\frac{3}{4}} \left[\frac{3}{2} - x \right] dx = \left[\frac{3}{2}x - \frac{1}{2}x^2 \right]_{\frac{1}{2}}^{\frac{3}{4}} = \frac{17}{32}$$

57. Una variable aleatoria continua X tiene por función de distribución:

$$F(x) = \begin{cases} 0 & x < 1 \\ x - 1 & 1 \leq x < 2 \\ 1 & x \geq 2 \end{cases}$$

- a) Calcular la función de densidad o función de cuantía
b) Calcular la media, mediana y coeficiente de variación

Solución:

- a) La función de densidad o función de cuantía es la derivada de la función de distribución en los puntos donde exista la derivada, entonces:

$$f(x) = \frac{dF(x)}{dx} = \begin{cases} 0 & x < 1 \\ 1 & 1 \leq x < 2 \\ 0 & x \geq 2 \end{cases} \longrightarrow f(x) = \begin{cases} 1 & 1 \leq x < 2 \\ 0 & \text{en otro caso} \end{cases}$$

$$b) \quad \text{Media: } \alpha_1 = \mu_x = E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_1^2 x dx = \left[\frac{x^2}{2} \right]_1^2 = 2 - \frac{1}{2} = \frac{3}{2} = 1,5$$

♦ La Mediana de una distribución es el valor que deja el 50% de la distribución a la derecha y el otro 50% a la izquierda, por lo que:

$$F(M_e) = 0,5 \rightarrow M_e - 1 = 0,5 \rightarrow M_e = 1,5$$

$$\int_1^{M_e} f(x) dx = 0,5 \Rightarrow \int_1^{M_e} dx = 0,5 \Rightarrow [x]_1^{M_e} = 0,5 \rightarrow M_e - 1 = 0,5 \rightarrow M_e = 1,5$$

♦ Coeficiente de variación: $CV_x = \frac{\sigma_x}{\mu_x}$

$$\alpha_2 = E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_1^2 x^2 dx = \left[\frac{x^3}{3} \right]_1^2 = \frac{8}{3} - \frac{1}{3} = \frac{7}{3}$$

$$\sigma_x^2 = \alpha_2 - \alpha_1^2 = \frac{7}{3} - \left(\frac{3}{2} \right)^2 = \frac{7}{3} - \frac{9}{4} = \frac{1}{12} \Rightarrow \sigma_x = \sqrt{\frac{1}{12}} = 0,08$$

$$CV_x = \frac{\sigma_x}{\mu_x} = \frac{0,08}{1,5} = 0,05$$

58. Una variable aleatoria continua X tiene por función de densidad

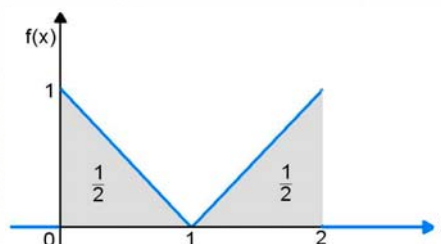
$$f(x) = \begin{cases} 1-x & 0 \leq x < 1 \\ x-1 & 1 \leq x \leq 2 \\ 0 & \text{otros casos} \end{cases}$$

Se pide:

- Representa la función de densidad
- Hallar la función de distribución y su gráfica
- $P(0 \leq X \leq 1)$ $P(-2 \leq X \leq 2)$ $P\left(\frac{1}{2} \leq X < \infty\right)$

Solución:

- Se observa que el área encerrada es igual a la unidad.



- La función de distribución se define $F(x) = \int_{-\infty}^x f(t) dt$

$$x < 0 \quad F(x) = \int_{-\infty}^x f(t) dt = \int_{-\infty}^x 0 dt = 0$$

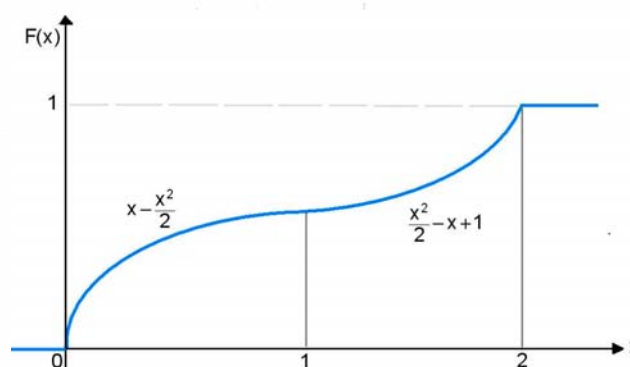
$$0 \leq x < 1 \quad F(x) = \int_{-\infty}^0 f(t) dt + \int_0^x f(t) dt = \int_0^x (1-t) dt = \left[t - \frac{t^2}{2} \right]_0^x = x - \frac{x^2}{2}$$

$$1 \leq x < 2 \quad F(x) = \int_{-\infty}^0 f(t) dt + \int_0^1 f(t) dt + \int_1^x f(t) dt = \int_0^1 (1-t) dt + \int_1^x (t-1) dt =$$

$$= \left[t - \frac{t^2}{2} \right]_0^1 + \left[\frac{t^2}{2} - t \right]_1^x = \left(1 - \frac{1}{2} \right) + \left[\left(\frac{x^2}{2} - x \right) - \left(\frac{1}{2} - 1 \right) \right] = \frac{x^2}{2} - x + 1$$

$$x \geq 2 \quad F(x) = \int_0^1 (1-t) dt + \int_1^2 (t-1) dt + \int_2^x 0 dt = \left[t - \frac{t^2}{2} \right]_0^1 + \left[\frac{t^2}{2} - t \right]_1^2 = 1$$

$$F(x) = \begin{cases} 0 & x < 0 \\ x - \frac{x^2}{2} & 0 \leq x < 1 \\ \frac{x^2}{2} - x + 1 & 1 \leq x \leq 2 \\ 1 & x > 2 \end{cases}$$



c) $P(0 \leq X \leq 1)$ $P(-2 \leq X \leq 2)$ $P\left(\frac{1}{2} \leq X < \infty\right)$

$$P(0 \leq X \leq 1) = F(1) - F(0) = \left(\frac{1}{2} - 1 + 1 \right) - 0 = \frac{1}{2}$$

$$P(-2 \leq X \leq 2) = F(2) - F(-2) = \left(\frac{4}{2} - 2 + 1 \right) - 0 = 1$$

$$P\left(\frac{1}{2} \leq X < \infty\right) = F(\infty) - F\left(\frac{1}{2}\right) = 1 - \left(\frac{1}{2} - \frac{1/4}{2} \right) = \frac{5}{8}$$

59. Sea (X, Y) una variable aleatoria bidimensional con función de densidad:

$$f(x, y) = \begin{cases} 1 & |y| < x, \quad 0 < x < 1 \\ 0 & \text{restantes valores} \end{cases}$$

a) Comprobar que $f(x, y)$ es función de densidad.

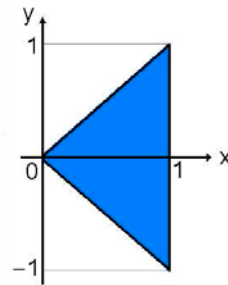
b) Hallar las medias de X e Y .

c) Hallar las probabilidades: $P\left[X < \frac{1}{2}; Y < 0\right]$ y $P\left[X > \frac{1}{2}; -\frac{1}{2} < Y < \frac{1}{2}\right]$

Solución:

a) $f(x, y)$ es función de densidad si se verifica $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy &= \int_0^1 \int_{-x}^x dx dy = \int_0^1 \left(\int_{-x}^x dy \right) dx = \int_0^1 [y]_{-x}^x dx = \\ &= \int_0^1 2x dx = [x^2]_0^1 = 1 \end{aligned}$$



en consecuencia, $f(x, y)$ es función de densidad.

b) Para hallar las medias de X e Y hay que calcular primero las funciones de densidad marginales:

$$\begin{aligned} \bullet \quad f_1(x) &= \int_{-\infty}^{\infty} f(x, y) dy = \int_{-x}^x dy = [y]_{-x}^x = 2x \quad 0 < x < 1 \\ \bullet \quad f_2(y) &= \int_{-\infty}^{\infty} f(x, y) dx = \begin{cases} \int_{-y}^1 dx = [x]_{-y}^1 = 1 + y & -1 < y < 0 \\ \int_y^1 dx = [x]_y^1 = 1 - y & 0 < y < 1 \end{cases} \end{aligned}$$

$$\alpha_{10} = \mu_x = E(X) = \int_{-\infty}^{\infty} x f_1(x) dx = \int_0^1 x \cdot 2x dx = 2 \left[\frac{x^3}{3} \right]_0^1 = \frac{2}{3}$$

$$\begin{aligned} \alpha_{01} = \mu_y = E(Y) &= \int_{-\infty}^{\infty} y f_2(y) dy = \int_{-1}^0 y f_2(y) dy + \int_0^1 y f_2(y) dy = \int_{-1}^0 y(1+y) dy + \int_0^1 y(1-y) dy = \\ &= \int_{-1}^0 (y + y^2) dy + \int_0^1 (y - y^2) dy = \left[\frac{y^2}{2} + \frac{y^3}{3} \right]_{-1}^0 + \left[\frac{y^2}{2} - \frac{y^3}{3} \right]_0^1 = -\frac{1}{2} + \frac{1}{3} + \frac{1}{2} - \frac{1}{3} = 0 \end{aligned}$$

c) Probabilidades: $P\left[X < \frac{1}{2} ; Y < 0\right]$ y $P\left[X > \frac{1}{2} ; -\frac{1}{2} < Y < \frac{1}{2}\right]$

$$P\left[X < \frac{1}{2} ; Y < 0\right] = \int_0^{1/2} \int_{-x}^0 f(x, y) dx dy = \int_0^{1/2} \left(\int_{-x}^0 dy \right) dx = \int_0^{1/2} [y]_{-x}^0 dx = \int_0^{1/2} x dx = \left[\frac{x^2}{2} \right]_0^{1/2} = \frac{1}{8}$$

$$P\left[X > \frac{1}{2} ; -\frac{1}{2} < Y < \frac{1}{2}\right] = \int_{1/2}^1 \int_{-1/2}^{1/2} f(x, y) dx dy = \int_{1/2}^1 \left(\int_{-1/2}^{1/2} dy \right) dx = \int_{1/2}^1 [y]_{-1/2}^{1/2} dx = \int_{1/2}^1 dx = [x]_{1/2}^1 = \frac{1}{2}$$

60. La función de densidad asociada a la emisión de billetes de una compañía área es:

$$f(x, y) = \begin{cases} x + y & 0 < x < 1 \quad 0 < y < 1 \\ 0 & \text{en el resto} \end{cases}$$

- a) Hallar la función de distribución.
b) Hallar las funciones de densidad marginales de X e Y.
c) ¿Son X e Y independientes?

Solución:

$$\begin{aligned} \text{a) } F(x, y) &= \int_{-\infty}^x \int_{-\infty}^y f(u, v) dv du = \int_0^x \int_0^y (u + v) du dv = \int_0^x \left(\int_0^y (u + v) dv \right) du = \\ &= \int_0^x \left(u[v]_0^y + \left[\frac{v^2}{2} \right]_0^y \right) du = \int_0^x \left[uy + \frac{y^2}{2} \right] du = y \left[\frac{u^2}{2} \right]_0^x + \frac{y^2}{2} [u]_0^x = \\ &= \frac{yx^2}{2} + \frac{y^2x}{2} = \frac{1}{2}(yx^2 + y^2x) \quad 0 \leq x < 1 \quad 0 \leq y < 1 \end{aligned}$$

$$\text{En consecuencia, } F(x, y) = \begin{cases} 0 & x < 0 \text{ ó } y < 0 \\ \frac{1}{2}(yx^2 + y^2x) & 0 \leq x < 1 \quad 0 \leq y < 1 \\ F_1(x) = \frac{1}{2}(x^2 + x) & 0 \leq x < 1, y \geq 1 \\ F_2(y) = \frac{1}{2}(y + y^2) & 0 \leq y < 1, x \geq 1 \\ 1 & x \geq 1, y \geq 1 \end{cases}$$

b) Funciones de densidad marginales de X e Y

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_0^1 (x + y) dy = x \left[y \right]_0^1 + \left[\frac{y^2}{2} \right]_0^1 = x + \frac{1}{2} \quad 0 < x < 1$$

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_0^1 (x + y) dx = \left[\frac{x^2}{2} \right]_0^1 + y \left[x \right]_0^1 = \frac{1}{2} + y \quad 0 < y < 1$$

Adviértase que:

$$f_1(x) = \frac{\partial F_1(x)}{\partial x} = \frac{\partial}{\partial x} \left[\frac{1}{2}(x^2 + x) \right] = x + \frac{1}{2} \quad f_2(y) = \frac{\partial F_2(y)}{\partial y} = \frac{\partial}{\partial y} \left[\frac{1}{2}(y + y^2) \right] = \frac{1}{2} + y$$

c) X e Y son independientes cuando se verifica $f(x, y) = f_1(x) \cdot f_2(y)$

$$f_1(x) \cdot f_2(y) = \left(x + \frac{1}{2} \right) \cdot \left(\frac{1}{2} + y \right) \neq x + y = f(x, y) \quad \text{luego no son independientes.}$$

61. La función de distribución asociada a un fenómeno de la naturaleza es:

$$F(x, y) = \begin{cases} (1 - e^{-x}) \cdot (1 - e^{-y}) & x > 0, y > 0 \\ 0 & \text{en el resto} \end{cases}$$

- Hallar la función de densidad.
- Hallar las funciones de densidad marginales de X e Y.
- Hallar las funciones de densidad condicionadas.
- Calcular el coeficiente de correlación.

Solución:

$$\begin{aligned} \text{a) } f(x, y) &= \frac{\partial^2 F(x, y)}{\partial x \partial y} = \frac{\partial}{\partial y} \left[\frac{\partial F(x, y)}{\partial x} \right] = \frac{\partial}{\partial y} \left[\frac{\partial ((1 - e^{-x}) \cdot (1 - e^{-y}))}{\partial x} \right] = \frac{\partial}{\partial y} \left[(1 - e^{-y}) \frac{\partial (1 - e^{-x})}{\partial x} \right] = \\ &= \frac{\partial}{\partial y} \left[(1 - e^{-y}) \cdot e^{-x} \right] = e^{-x} \cdot \frac{\partial (1 - e^{-y})}{\partial y} = e^{-x} \cdot e^{-y} = e^{-(x+y)} = \frac{1}{e^{x+y}} \quad x > 0, y > 0 \end{aligned}$$

$$\text{Función de densidad } f(x, y) = \begin{cases} e^{-(x+y)} & x > 0, y > 0 \\ 0 & \text{en el resto} \end{cases}$$

b) Funciones de densidad marginales

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_0^{\infty} e^{-(x+y)} dy = \int_0^{\infty} e^{-x} \cdot e^{-y} dy = e^{-x} \int_0^{\infty} e^{-y} dy = -e^{-x} \left[e^{-y} \right]_0^{\infty} = -e^{-x} \cdot (-1) = e^{-x}$$

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_0^{\infty} e^{-(x+y)} dx = \int_0^{\infty} e^{-y} \cdot e^{-x} dx = e^{-y} \int_0^{\infty} e^{-x} dx = -e^{-y} [e^{-x}]_0^{\infty} = -e^{-y} \cdot (-1) = e^{-y}$$

Adviértase que X e Y son independientes al verificarse $f(x, y) = f_1(x) \cdot f_2(y)$

$$f(x, y) = e^{-(x+y)} = f_1(x) \cdot f_2(y) = e^{-x} \cdot e^{-y}$$

X e Y independientes $\rightarrow$ La covarianza $\mu_{11} = \sigma_{xy} = 0 \rightarrow \rho = 0$

c) Funciones de densidad condicionadas

$$f(x / y) = \frac{f(x, y)}{f_2(y)} = \frac{e^{-(x+y)}}{e^{-y}} = e^{-x} = f_1(x) \text{ al ser X e Y independientes}$$

$$f(y / x) = \frac{f(x, y)}{f_1(x)} = \frac{e^{-(x+y)}}{e^{-x}} = e^{-y} = f_2(y) \text{ al ser X e Y independientes}$$

d) El coeficiente de correlación $\rho = \sigma_{xy} = \frac{\mu_{11}}{\sigma_x \cdot \sigma_y}$

$$\alpha_{10} = \mu_x = E(X) = \int_{-\infty}^{\infty} x \cdot f_1(x) dx = \int_0^{\infty} x \cdot e^{-x} dx \doteq [-x \cdot e^{-x}]_0^{\infty} + \int_0^{\infty} e^{-x} dx = [-x \cdot e^{-x} - e^{-x}]_0^{\infty} = 1$$

$$\alpha_{01} = \mu_y = E(Y) = \int_{-\infty}^{\infty} y \cdot f_2(y) dy = \int_0^{\infty} y \cdot e^{-y} dy \doteq [-y \cdot e^{-y}]_0^{\infty} + \int_0^{\infty} e^{-y} dy = [-y \cdot e^{-y} - e^{-y}]_0^{\infty} = 1$$

Nota: $\int_0^{\infty} x \cdot e^{-x} dx = [x \cdot e^{-x}]_0^{\infty} + \int_0^{\infty} e^{-x} dx = [x \cdot e^{-x} - e^{-x}]_0^{\infty}$

donde se ha realizado el cambio $\begin{cases} u = x \rightarrow du = dx \\ dv = e^{-x} dx \rightarrow v = \int_0^{\infty} e^{-x} dx = [-e^{-x}]_0^{\infty} \end{cases}$

$$\alpha_{11} = E(X \cdot Y) = \int_0^{\infty} \int_0^{\infty} x \cdot y \cdot f(x, y) dx dy = \int_0^{\infty} \int_0^{\infty} x \cdot y \cdot e^{-(x+y)} dx dy = \int_0^{\infty} x \cdot e^{-x} dx \cdot \int_0^{\infty} y \cdot e^{-y} dy = 1$$

$$\alpha_{20} = E(X^2) = \int_{-\infty}^{\infty} x^2 \cdot f_1(x) dx = \int_0^{\infty} x^2 \cdot e^{-x} dx \doteq [-x^2 \cdot e^{-x}]_0^{\infty} + \int_0^{\infty} 2 \cdot x \cdot e^{-x} dx =$$

$$= [-x^2 \cdot e^{-x}]_0^{\infty} + 2[-x \cdot e^{-x} - e^{-x}]_0^{\infty} = [-x^2 \cdot e^{-x} - 2 \cdot x \cdot e^{-x} - 2 \cdot e^{-x}]_0^{\infty} = 2$$

Análogamente, $\alpha_{02} = E(Y^2) = 2$

$$\sigma_x^2 = \alpha_{20} - \alpha_{10}^2 = 2 - 1 = 1 \quad \sigma_x = \sqrt{1} = 1$$

$$\sigma_y^2 = \alpha_{02} - \alpha_{01}^2 = 2 - 1 = 1 \quad \sigma_y = \sqrt{1} = 1$$

$$\text{Covarianza: } \mu_{11} = \sigma_{xy} = \alpha_{11} - \alpha_{10} \cdot \alpha_{01} = 1 - 1 \cdot 1 = 0$$

$$\text{Coeficiente de correlación } \rho = \sigma_{xy} = \frac{\mu_{11}}{\sigma_x \cdot \sigma_y} = \frac{0}{1 \cdot 1} = 0 \Rightarrow \text{Las variables son incorreladas}$$

62. La venta en un mercado de abastos lleva asociada la función:

$$f(x, y) = \begin{cases} k \left[\frac{xy}{2} + 1 \right] & 0 < x < 1 \quad -1 < y < 1 \\ 0 & \text{en el resto} \end{cases}$$

- Hallar k para que sea función de densidad
- Hallar la función de distribución
- Funciones de densidad marginales y condicionadas
- Se considera la transformación $Z = X - Y$ y $T = X + 2Y$, hallar la función de densidad de la variable (Z, T)

Solución:

$$a) \quad f(x, y) \geq 0 \Leftrightarrow k \geq 0$$

Para que $f(x, y)$ sea función de densidad debe verificarse que $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$

$$\begin{aligned} \int_0^1 \int_{-1}^1 k \left[\frac{xy}{2} + 1 \right] dx dy &= \int_0^1 k \left(\int_{-1}^1 \left[\frac{xy}{2} + 1 \right] dy \right) dx = \int_0^1 k \left(x \left[\frac{y^2}{4} \right]_{-1}^1 + [y]_{-1}^1 \right) dx = \\ &= \int_0^1 2k dx = 2k [x]_0^1 = 2k = 1 \quad \mapsto \quad k = \frac{1}{2} \end{aligned}$$

$$\text{La función de densidad } f(x, y) = \begin{cases} \frac{xy + 2}{4} & 0 < x < 1 \quad -1 < y < 1 \\ 0 & \text{en el resto} \end{cases}$$

$$b) \quad \text{Función de distribución } F(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^x \int_{-\infty}^y f(u, v) dv du$$

$$\begin{aligned}
 F(x, y) &= \int_0^x \int_{-1}^y \left[\frac{uv+2}{4} \right] dv du = \int_0^x \left(\int_{-1}^y \left[\frac{uv+2}{4} \right] dv \right) du = \frac{1}{4} \int_0^x \int_{-1}^y ([uv+2] dv) du = \\
 &= \frac{1}{4} \int_0^x \left(u \left[\frac{v^2}{2} \right]_{-1}^y + 2[v]_{-1}^y \right) du = \frac{1}{4} \int_0^x \left(\frac{uy^2}{2} - \frac{u}{2} + 2y + 2 \right) du = \\
 &= \frac{1}{4} \left[\frac{y^2}{2} \left(\frac{u^2}{2} \right)_0^x - \frac{1}{2} \left(\frac{u^2}{2} \right)_0^x + 2y(u)_0^x + 2(u)_0^x \right] = \frac{x^2(y^2-1)}{16} + \frac{x(y+1)}{2}
 \end{aligned}$$

En consecuencia, $F(x, y) = \frac{x^2(y^2-1)}{16} + \frac{x(y+1)}{2} \quad 0 \leq x < 1 \quad -1 \leq y < 1$

Las funciones de distribución marginales, resultan:

$$\begin{aligned}
 F_1(x) = P(X \leq x) &= \int_{-\infty}^x \int_{-\infty}^y f(u, v) dv du = \int_0^x \int_{-1}^1 \frac{uv+2}{4} dv du = \int_0^x \left(\int_{-1}^1 \frac{uv+2}{4} dv \right) du = \\
 &= \int_0^x \left(u \left[\frac{v^2}{8} \right]_{-1}^1 + \frac{2}{4} [v]_{-1}^1 \right) du = \int_0^x du = [u]_0^x = x \quad 0 \leq x < 1
 \end{aligned}$$

$$\begin{aligned}
 F_2(y) = P(Y \leq y) &= \int_{-\infty}^y \int_{-\infty}^x f(u, v) dv du = \int_0^1 \int_{-1}^y \frac{uv+2}{4} dv du = \int_0^1 \left(\int_{-1}^y \frac{uv+2}{4} dv \right) du = \\
 &= \int_0^1 \left(u \left[\frac{v^2}{8} \right]_{-1}^y + \frac{2}{4} [v]_{-1}^y \right) du = \int_0^1 \left[\left(\frac{y^2-1}{8} \right) u + \frac{2}{4} (y+1) \right] du = \\
 &= \left(\frac{y^2-1}{8} \right) \left[\frac{u^2}{2} \right]_0^1 + \frac{2}{4} (y+1) [u]_0^1 = \frac{y^2-1}{16} + \frac{y+1}{2} = \frac{y^2+8y+7}{16} \quad -1 \leq y < 1
 \end{aligned}$$

c) También se podrían haber hallado a través de las funciones de densidad marginales:

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_{-1}^1 \left[\frac{xy}{4} + \frac{1}{2} \right] dy = \frac{x}{4} \left[\frac{y^2}{2} \right]_{-1}^1 + \frac{1}{2} [y]_{-1}^1 = 1$$

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_0^1 \left[\frac{xy}{4} + \frac{1}{2} \right] dx = \frac{y}{4} \left[\frac{x^2}{2} \right]_0^1 + \frac{1}{2} [x]_0^1 = \frac{y+4}{8}$$

$$F_1(x) = P(X \leq x) = \int_{-\infty}^x f_1(u) du = \int_0^x du = [u]_0^x = x \quad 0 \leq x < 1$$

$$F_2(y) = P(Y \leq y) = \int_{-\infty}^y f_2(v) dv = \int_{-1}^y \left[\frac{v+4}{8} \right] dv = \frac{1}{8} \left[\frac{v^2}{2} + 4v \right]_{-1}^y = \frac{y^2 + 8y + 7}{16} \quad -1 \leq y < 1$$

La función de distribución conjunta:

$$F(x, y) = \begin{cases} 0 & x < 0 \quad \text{ó} \quad y < -1 \\ \frac{x^2(y^2 - 1)}{16} + \frac{x(y+1)}{2} & 0 \leq x < 1 \quad -1 \leq y < 1 \\ x & 0 \leq x < 1, \quad y \geq 1 \\ \frac{y^2 + 8y + 7}{16} & -1 \leq y < 1, \quad x \geq 1 \\ 1 & x \geq 1, \quad y \geq 1 \end{cases}$$

Las funciones de densidad marginales se pueden hallar a partir de la función de distribución conjunta:

$$f_1(x) = \frac{\partial F_1(x)}{\partial x} = \frac{\partial}{\partial x}(x) = 1 \quad f_2(y) = \frac{\partial F_2(y)}{\partial y} = \frac{\partial}{\partial y} \left[\frac{y^2 + 8y + 7}{16} \right] = \frac{y+4}{8}$$

o bien, a partir de la función de densidad:

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_{-1}^1 \left[\frac{xy}{4} + \frac{1}{2} \right] dy = \frac{x}{4} \left[\frac{y^2}{2} \right]_{-1}^1 + \frac{1}{2} [y]_{-1}^1 = 1$$

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_0^1 \left[\frac{xy}{4} + \frac{1}{2} \right] dx = \frac{y}{4} \left[\frac{x^2}{2} \right]_0^1 + \frac{1}{2} [x]_0^1 = \frac{y+4}{8}$$

Funciones de densidad condicionadas:

$$f(y/x) = \frac{f(x, y)}{f_1(x)} = \frac{(xy+2)/4}{1} = \frac{xy+2}{4} \neq f_2(y)$$

$$f(x/y) = \frac{f(x, y)}{f_2(y)} = \frac{(xy+2)/4}{(y+4)/8} = \frac{2xy+4}{y+4} \neq f_1(x)$$

d) Las variables X e Y no son independientes al ser $f(x, y) \neq f_1(x) \cdot f_2(y)$

$$\text{En la transformación } \begin{cases} Z = X - Y \\ T = X + 2Y \end{cases} \quad J = \frac{\partial(z, t)}{\partial(x, y)} = \begin{vmatrix} \frac{\partial z}{\partial x} & \frac{\partial z}{\partial y} \\ \frac{\partial t}{\partial x} & \frac{\partial t}{\partial y} \end{vmatrix} = \begin{vmatrix} 1 & -1 \\ 1 & 2 \end{vmatrix} = 3 \neq 0$$

por lo que existe la función de densidad $g(z, t)$

Despejando (X,Y) en la función de (Z,T) se calcula el jacobiano $J_1 = \frac{\partial(x,y)}{\partial(z,t)} = \begin{vmatrix} \frac{\partial x}{\partial z} & \frac{\partial x}{\partial t} \\ \frac{\partial y}{\partial z} & \frac{\partial y}{\partial t} \end{vmatrix}$

$$\begin{cases} Z = X - Y \\ T = X + 2Y \end{cases} \mapsto \begin{cases} 2Z = 2X - 2Y \\ T = X + 2Y \end{cases} \mapsto \begin{cases} X = \frac{2Z + T}{3} \\ Y = \frac{-Z + T}{3} \end{cases} \quad J_1 = \frac{\partial(x,y)}{\partial(z,t)} = \begin{vmatrix} 2/3 & 1/3 \\ -1/3 & 1/3 \end{vmatrix} = \frac{1}{3}$$

La función de densidad $g(z,t) = f[h_1(z,t), h_2(z,t)] \cdot |J_1|$

$$h_1(z,t) = \frac{2z+t}{3}, \quad h_2(z,t) = \frac{-z+t}{3}, \quad \begin{cases} 0 < x < 1 & \rightarrow 0 < 2z+t < 3 \\ -1 < y < 1 & \rightarrow -3 < -z+t < 3 \end{cases}$$

$$g(z,t) = \frac{\left[\frac{2z+t}{3} \right] \left[\frac{-z+t}{3} \right]}{4} \cdot \left| \frac{1}{3} \right| = \frac{(2z+t)(-z+t)}{108}$$

$$f(x,y) = \begin{cases} \frac{xy+2}{4} & 0 < x < 1 \\ 0 & -1 < y < 1 \\ 0 & \text{en el resto} \end{cases} \quad g(z,t) = \begin{cases} \frac{(2z+t)(-z+t)}{108} & 0 < 2z+t < 3 \\ 0 & -3 < -z+t < 3 \\ 0 & \text{en el resto} \end{cases}$$

63. Gran número de fenómenos aeronáuticos tienen asociada una variable aleatoria con ley de probabilidad:

$$f(x) = \begin{cases} ke^{-kx} & 0 < x < \infty \quad k > 0 \\ 0 & \text{en otros casos} \end{cases}$$

Se pide:

- ¿Puede tomar k cualquier valor?
- Para $k = 0,1$ representar la función de densidad, la función de distribución y su gráfica
- Siendo $k = 0,1$ hallar $P(X > 10)$
- Para $k = 0,1$ calcular $P(50 < X \leq 100)$

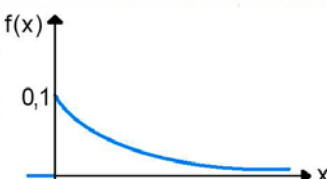
Solución:

- Para que $f(x)$ sea función de densidad debe verificar:

$$1 = \int_{-\infty}^{\infty} f(x) dx = \int_{-\infty}^0 f(x) dx + \int_0^{\infty} f(x) dx = \int_0^{\infty} k e^{-kx} dx = - \int_0^{\infty} -k e^{-kx} dx = -[e^{-kx}]_0^{\infty} = -\left[\frac{1}{e^{kx}}\right]_0^{\infty} = 1$$

La función de densidad no depende del valor del parámetro k , pudiendo tomar éste cualquier valor positivo.

b) La función de densidad para $k = 0,1$ será:

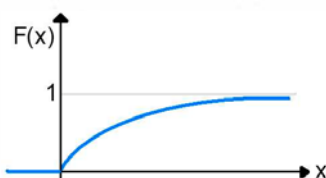
$$f(x) = \begin{cases} 0,1 \cdot e^{-0,1x} & x > 0 \\ 0 & \text{otros casos} \end{cases}$$


La función de distribución se define $F(x) = \int_{-\infty}^x f(t) dt$

$$x \leq 0 \quad F(x) = \int_{-\infty}^x f(t) dt = 0$$

$$x > 0 \quad F(x) = \int_{-\infty}^0 f(t) dt + \int_0^x f(t) dt = \int_0^x 0,1 \cdot e^{-0,1t} dt = - \int_0^x -0,1 \cdot e^{-0,1t} dt =$$

$$= -[e^{-0,1t}]_0^x = -\left[\frac{1}{e^{0,1t}}\right]_0^x = -\left(\frac{1}{e^{0,1x}} - 1\right) = 1 - e^{-0,1x}$$

$$F(x) = \begin{cases} 0 & x \leq 0 \\ 1 - e^{-0,1x} & x > 0 \end{cases}$$


$$c) P(X > 10) = 1 - P(X \leq 10) = 1 - F(10) = 1 - (1 - e^{-0,1 \cdot 10}) = e^{-1} = \frac{1}{e}$$

$$d) P(50 < X \leq 100) = F(100) - F(50) = (1 - e^{-0,1 \cdot 100}) - (1 - e^{-0,1 \cdot 50}) = -e^{-10} + e^{-5} = \frac{1}{e^5} - \frac{1}{e^{10}}$$

64. La función de distribución asociada a la producción de una máquina, en miles de unidades, es del tipo:

$$F(x) = \begin{cases} 0 & x < 0 \\ x(2-x) & 0 \leq x \leq k \\ 1 & x > k \end{cases}$$

- Determinar k para que sea función de distribución
- Hallar la función de densidad
- Calcular la media, mediana, moda y varianza de la producción
- Hallar $P(X < 0,5)$ y $P(X > 0,25)$

Solución:

- a) Para que sea función de distribución se debe verificar:

$$1 = \lim_{x \rightarrow k^+} F(x) = \lim_{x \rightarrow k^-} F(x) \rightarrow \lim_{x \rightarrow k^-} x(x-2) = k(k-2) = 1 \rightarrow k^2 - 2k + 1 = 0 \Rightarrow k = 1$$

En consecuencia, la función de distribución es: $F(x) = \begin{cases} 0 & x < 0 \\ x(2-x) & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$

- b) La función de densidad o función de cuantía es la derivada de la función de distribución en los puntos donde exista la derivada.

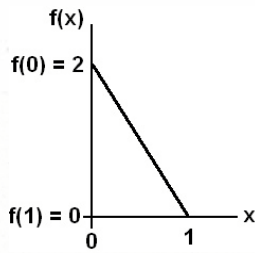
$$f(x) = \frac{dF(x)}{dx} = \begin{cases} 0 & x < 0 \\ 2-2x & 0 \leq x \leq 1 \\ 0 & x > 1 \end{cases} \longrightarrow f(x) = \begin{cases} 2-2x & 0 \leq x \leq 1 \\ 0 & \text{en otro caso} \end{cases}$$

- c) Media:

$$\alpha_1 = \mu_x = E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_0^1 x(2-2x) dx = \int_0^1 (2x-2x^2) dx = \left[x^2 - \frac{2x^3}{3} \right]_0^1 = 1 - \frac{2}{3} = \frac{1}{3}$$

- Para calcular la Moda hay que ver el valor que hace mínima la función de densidad o de cuantía, es decir:

$$f(x) = \begin{cases} 2-2x & 0 \leq x \leq 1 \\ 0 & \text{en otro caso} \end{cases} \rightarrow f'(x) = \begin{cases} -2 & 0 \leq x \leq 1 \\ 0 & \text{en otro caso} \end{cases}$$



La derivada de la función de cuantía $f'(x) = -2 < 0$, por lo que se trata de una función decreciente y toma el valor máximo en el extremo del intervalo $[0, 1]$, por tanto la moda $M_d = 0$

$$f(1) = 0 \leq f(x) \leq f(0) = 1 \rightarrow M_d = 0$$

• La Mediana de una distribución es el valor que deja el 50% de la distribución a la derecha y el otro 50% a la izquierda, por lo que:

$$F(M_e) = 0,5 \rightarrow M_e(2 - M_e) = 0,5 \Rightarrow M_e^2 - 2M_e + 0,5 = 0 \rightarrow 2M_e^2 - 4M_e + 1 = 0$$

$$2M_e^2 - 4M_e + 1 = 0 \rightarrow M_e = \frac{4 \pm \sqrt{16 - 8}}{4} = \frac{4 \pm 2\sqrt{2}}{4} = 1 \pm \frac{\sqrt{2}}{2}$$

De las dos soluciones se rechaza aquella que es mayor que 1: $M_e = 1 - \frac{\sqrt{2}}{2}$

• La Varianza de la producción: $\sigma_x^2 = \alpha_2 - \alpha_1^2$

$$\alpha_2 = E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_0^1 x^2 (2 - 2x) dx = \left[\frac{2x^3}{3} - \frac{x^4}{2} \right]_0^1 = \frac{2}{3} - \frac{1}{2} = \frac{1}{6}$$

$$\sigma_x^2 = \alpha_2 - \alpha_1^2 = \frac{1}{6} - \left(\frac{1}{3} \right)^2 = \frac{1}{18}$$

$$c) \text{ Función de distribución: } F(x) = \begin{cases} 0 & x < 0 \\ x(2 - x) & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases}$$

$$P(X < 0,5) = P(X \leq 0,5) = F(0,5) = 0,5(2 - 0,5) = 0,75$$

$$P(X > 0,25) = 1 - P(X \leq 0,25) = 1 - F(0,25) = 1 - 0,25(2 - 0,25) = 0,5625$$

También mediante la función de cuantía: $f(x) = \begin{cases} 2 - 2x & 0 \leq x \leq 1 \\ 0 & \text{en otro caso} \end{cases}$

$$P(X < 0,5) = \int_0^{0,5} f(x) dx = \int_0^{0,5} (2 - 2x) dx = \left[2x - x^2 \right]_0^{0,5} = 1 - 0,25 = 0,75$$

$$P(X > 0,25) = \int_{0,25}^1 f(x) dx = \int_{0,25}^1 (2 - 2x) dx = \left[2x - x^2 \right]_{0,25}^1 = 1 - (0,5 - 0,0625) = 0,5625$$

65. Dada la variable aleatoria continua X con función de densidad:

$$f(x) = \begin{cases} k(x+2) & 0 \leq x \leq 4 \\ 0 & \text{en el resto} \end{cases}$$

Hallar:

- El valor de k para que sea realmente una función de densidad
- La función de distribución
- La varianza
- $P(2 \leq X \leq 3)$

Solución:

$$a) \int_0^4 f(x) dx = \int_0^4 k(x+2) dx = k \int_0^4 (x+2) dx = k \left[\frac{x^2}{2} + 2x \right]_0^4 = 16k = 1 \Rightarrow k = \frac{1}{16}$$

$$f(x) = \begin{cases} \frac{1}{16}(x+2) & 0 \leq x \leq 4 \\ 0 & \text{en el resto} \end{cases}$$

b) Función de distribución: $F(x) = \int_{-\infty}^x f(t) dt$, en este caso:

$$x < 0 \quad F(x) = \int_{-\infty}^x f(t) dt = \int_{-\infty}^x 0 dt = 0$$

$$0 \leq x < 4 \quad F(x) = \int_{-\infty}^x f(t) dt = \int_{-\infty}^x 0 dt + \int_0^x \frac{1}{16}(t+2) dt = \frac{1}{16} \int_0^x (t+2) dt = \frac{1}{16} \left[\frac{t^2}{2} + 2t \right]_0^x = \frac{x^2 + 4x}{32}$$

$$x \geq 4 \quad F(x) = \int_{-\infty}^x f(t) dt = \int_{-\infty}^x 0 dt + \int_0^4 \frac{1}{16}(t+2) dt + \int_4^x 0 dt = \frac{1}{16} \left[\frac{t^2}{2} + 2t \right]_0^4 = 1$$

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2 + 4x}{32} & 0 \leq x < 4 \\ 1 & x \geq 4 \end{cases}$$

c) Para calcular la varianza: $f(x) = \begin{cases} \frac{1}{16}(x+2) & 0 \leq x \leq 4 \\ 0 & \text{en el resto} \end{cases}$

$$\alpha_1 = \mu = E[X] = \int_0^4 x f(x) dx = \int_0^4 x \left[\frac{1}{16}(x+2) \right] dx = \frac{1}{16} \int_0^4 (x^2 + 2x) dx = \frac{1}{16} \left[\frac{x^3}{3} + x^2 \right]_0^4 =$$

$$= \frac{1}{16} \left[\frac{112}{3} \right] = \frac{7}{3}$$

$$\alpha_2 = E[X^2] = \int_0^4 x^2 f(x) dx = \int_0^4 x^2 \left[\frac{1}{16}(x+2) \right] dx = \frac{1}{16} \int_0^4 (x^3 + 2x^2) dx = \frac{1}{16} \left[\frac{x^4}{4} + \frac{2x^3}{3} \right]_0^4 =$$

$$= \frac{1}{16} \left[64 + \frac{128}{3} \right] = \frac{20}{3}$$

$$\sigma_x^2 = \text{Var}(X) = \alpha_2 - (\alpha_1)^2 \rightarrow \sigma_x^2 = \frac{20}{3} - \left(\frac{7}{3} \right)^2 = \frac{11}{9}$$

$$d) \quad F(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2 + 4x}{32} & 0 \leq x < 4 \\ 1 & x \geq 4 \end{cases} \quad f(x) = \begin{cases} \frac{1}{16}(x+2) & 0 \leq x \leq 4 \\ 0 & \text{en el resto} \end{cases}$$

$$P(2 \leq X \leq 3) = F(3) - F(2) = \left(\frac{9+12}{32} \right) - \left(\frac{4+8}{32} \right) = \frac{21-12}{32} = \frac{9}{32}$$

$$P(2 \leq X \leq 3) = \int_2^3 \frac{1}{16}(x+2) dx = \frac{1}{16} \left[\frac{x^2}{2} + 2x \right]_2^3 = \frac{1}{16} \times \frac{9}{2} = \frac{9}{32}$$

66. Dada la función $f(x) = e^{-2x}$

- a) Comprobar si puede ser función de densidad de una variable aleatoria X cuando su campo de variación es el intervalo $x \geq 0$
- b) En caso de que no lo pueda ser, qué modificaciones habría que introducir para que lo fuera.

Solución:

a) Para que sea función de densidad, debe cumplir dos condiciones en el campo de variación de la variable aleatoria:

- $f(x)$ no puede ser negativa
- La integral de $f(x)$ en el campo de variación es 1
- $f(x) = e^{-2x} \geq 0 \rightarrow L e^{-2x} \geq L 0 \Rightarrow -2x > -\infty \rightarrow x < \infty$ es positiva

$$\bullet \int_0^{\infty} e^{-2x} dx = \left[-\frac{1}{2} e^{-2x} \right]_0^{\infty} = \left[0 + \frac{1}{2} \right] = \frac{1}{2} \neq 1.$$

No se cumple, luego la función dada no es de densidad en el intervalo.

b) Para que sea función de densidad, se define $f(x) = k e^{-2x}$

$$\int_0^{\infty} k e^{-2x} dx = k \int_0^{\infty} e^{-2x} dx = k \left[-\frac{1}{2} e^{-2x} \right]_0^{\infty} = \frac{k}{2} = 1 \rightarrow k = 2$$

En consecuencia, $f(x) = 2 e^{-2x}$

67. Sea X una variable aleatoria continua con función de densidad:

$$f(x) = \begin{cases} \frac{8}{7x^2} & 1 \leq x \leq 8 \\ 0 & \text{otro caso} \end{cases}$$

a) Calcular el primer y tercer cuartil, el decil 7 y el percentil 85

b) Calcular la mediana y moda

Solución:

a) Función de distribución:

$$F(x) = P[X \leq x] = \int_{-\infty}^x f(t) dt = \int_1^x \frac{8}{7t^2} dt = -\frac{8}{7} \left[\frac{1}{t} \right]_1^x = \frac{8(x-1)}{7x} \quad 1 \leq x \leq 8$$

sustituyendo, queda:

$$F(Q_1) = \frac{1}{4} = \frac{8(Q_1 - 1)}{7Q_1} \rightarrow 7Q_1 = 32(Q_1 - 1) \rightarrow Q_1 = \frac{32}{25} = 1,28 \mapsto Q_1 = P_{25} = 1,28$$

$$F(Q_3) = \frac{3}{4} = \frac{8(Q_3 - 1)}{7Q_3} \rightarrow 21Q_3 = 32(Q_3 - 1) \rightarrow Q_3 = \frac{32}{11} = 2,91 \mapsto Q_3 = D_5 = P_{75} = 2,91$$

$$F(D_7) = \frac{7}{10} = \frac{8(D_7 - 1)}{7D_7} \rightarrow 49D_7 = 80(D_7 - 1) \rightarrow D_7 = \frac{80}{31} = 2,58$$

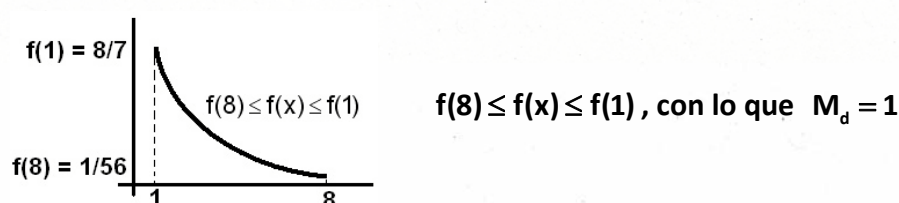
$$F(P_{85}) = \frac{85}{100} = \frac{8(P_{85} - 1)}{7P_{85}} \rightarrow 595P_{85} = 800(P_{85} - 1) \rightarrow P_{85} = \frac{800}{205} = 3,90$$

b) $M_e = Q_2 = D_5 = P_{50}$

$$F(M_e) = \frac{1}{2} = \frac{8(M_e - 1)}{7M_e} \rightarrow 7M_e = 16(M_e - 1) \rightarrow M_e = \frac{16}{9} = 1,78$$

La Moda M_d se obtiene calculando el máximo de la función de densidad:

$$f(x) = \frac{8}{7x^2} \rightarrow f'(x) = -\frac{16}{7x^3} < 0 \mapsto \text{La función es decreciente}$$



68. La demanda diaria de un determinado artículo es una variable aleatoria con función de densidad:

$$f(x) = \begin{cases} \frac{1}{8} & 0 < x \leq 4 \\ \frac{12-x}{64} & 4 < x \leq 12 \\ 0 & \text{otro caso} \end{cases}$$

Los beneficios diarios dependen de la demanda según la siguiente función:

$$B = \begin{cases} -5 & \text{si } x < 2 \\ 5 & \text{si } 2 < x \leq 4 \\ 10 & \text{si } 4 < x \leq 8 \\ 15 & \text{si } 8 < x \leq 12 \end{cases}$$

Calcular:

- Probabilidad de que en un día cualquiera la demanda sea superior a 10
- Probabilidad de que la demanda sea inferior a 3
- La esperanza y la varianza de la demanda
- Función de distribución de la demanda
- Función de cuantía y función de distribución de la variable aleatoria beneficios diarios
- Esperanza y varianza de la variable beneficios

Solución:

$$a) P(X > 10) = \int_{10}^{12} f(x) dx = \int_{10}^{12} \frac{12-x}{64} dx = \frac{1}{64} \left[12x - \frac{x^2}{2} \right]_{10}^{12} = \frac{1}{32} = 0,03125$$

$$b) P(X < 3) = \int_0^3 f(x) dx = \int_0^3 \frac{1}{8} dx = \frac{1}{8} [x]_0^3 = \frac{3}{8} = 0,375$$

c) Media o Esperanza

$$\begin{aligned} \alpha_1 = \mu_x = E(X) &= \int_{-\infty}^{\infty} x \cdot f(x) dx = \int_0^4 x \cdot f(x) dx + \int_4^{12} x \cdot f(x) dx = \int_0^4 x \cdot \frac{1}{8} dx + \int_4^{12} x \cdot \frac{12-x}{64} dx = \\ &= \frac{1}{8} \int_0^4 x dx + \frac{1}{64} \int_4^{12} (12x - x^2) dx = \frac{1}{16} [x^2]_0^4 + \frac{1}{64} \left[6x^2 - \frac{x^3}{3} \right]_4^{12} = \\ &= 1 + \frac{1}{64} \left(864 - \frac{1728}{3} - 96 + \frac{64}{3} \right) = \frac{13}{3} = 4,33 \end{aligned}$$

Varianza:

$$\begin{aligned} \alpha_2 = E(X^2) &= \int_{-\infty}^{\infty} x^2 \cdot f(x) dx = \int_0^4 x^2 \cdot f(x) dx + \int_4^{12} x^2 \cdot f(x) dx = \int_0^4 x^2 \cdot \frac{1}{8} dx + \int_4^{12} x^2 \cdot \frac{12-x}{64} dx = \\ &= \frac{1}{8} \int_0^4 x^2 dx + \frac{1}{64} \int_4^{12} (12x^2 - x^3) dx = \frac{1}{24} [x^3]_0^4 + \frac{1}{64} \left[4x^3 - \frac{x^4}{4} \right]_4^{12} = \\ &= \frac{64}{24} + \frac{1}{64} (6912 - 5184 - 256 + 64) = \frac{5120}{192} = \frac{80}{3} = 26,67 \end{aligned}$$

$$\sigma_x^2 = \text{Var}(X) = \alpha_2 - \alpha_1^2 = \frac{80}{3} - \left(\frac{13}{3} \right)^2 = \frac{71}{9} = 7,89$$

d) La función de distribución de la demanda: $F(x) = \int_{-\infty}^x f(t) dt$

$$F(x) = \begin{cases} x < 0 & \int_{-\infty}^x f(x) dx = \int_{-\infty}^x 0 dx = 0 \\ 0 \leq x < 4 & \int_{-\infty}^x f(x) dx = \int_{-\infty}^0 0 dx + \int_0^x \frac{1}{8} dx = \frac{x}{8} \\ 4 \leq x < 12 & \int_{-\infty}^x f(x) dx = \int_0^4 \frac{1}{8} dx + \int_4^x \frac{12-x}{64} dx = \frac{1}{2} + \frac{1}{64} \left(-\frac{x^2}{2} + 12x - 40 \right) \\ x \geq 12 & \int_{-\infty}^x f(x) dx = \int_{-\infty}^0 0 dx + \int_0^4 \frac{1}{8} dx + \int_4^{12} \frac{12-x}{64} dx + \int_{12}^x 0 dx = 1 \end{cases}$$

En resumen:
$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x}{8} & 0 \leq x < 4 \\ \frac{1}{2} + \frac{1}{64} \left(-\frac{x^2}{2} + 12x - 40 \right) & 4 \leq x < 12 \\ 1 & x \geq 12 \end{cases}$$

e) La función de cuantía y la función de distribución de la variable aleatoria beneficios diarios se hallan considerando:

$$B = \begin{cases} -5 & \text{si } x < 2 \\ 5 & \text{si } 2 < x \leq 4 \\ 10 & \text{si } 4 < x \leq 8 \\ 15 & \text{si } 8 < x \leq 12 \end{cases} \quad f(x) = \begin{cases} \frac{1}{8} & 0 < x \leq 4 \\ \frac{12-x}{64} & 4 < x \leq 12 \\ 0 & \text{otro caso} \end{cases}$$

Función de cuantía o probabilidad:

b_i	$P[B = b_i]$
-5	$\int_0^2 f(x) dx = \int_0^2 \frac{1}{8} dx = \frac{1}{4} = 0,25$
5	$\int_2^4 f(x) dx = \int_2^4 \frac{1}{8} dx = \frac{1}{4} = 0,25$
10	$\int_4^8 f(x) dx = \int_4^8 \frac{12-x}{64} dx = \frac{1}{64} \left(12x - \frac{x^2}{2} \right)_4^8 = 0,375$
15	$\int_8^{12} f(x) dx = \int_8^{12} \frac{12-x}{64} dx = \frac{1}{64} \left(12x - \frac{x^2}{2} \right)_8^{12} = 0,125$

Función de distribución $F(B) = P(B \leq b_i)$

b_i	$P[B = b_i]$	$F(B) = P(B \leq b_i)$	$b_i \cdot P[B = b_i]$	$b_i^2 \cdot P[B = b_i]$
-5	0,25	0,25	-1,25	6,25
5	0,25	0,50	1,25	6,25
10	0,375	0,875	3,75	37,5
15	0,125	1	1,875	28,125
			$\sum_{i=1}^4 b_i \cdot P[B = b_i] = 5,625$	$\sum_{i=1}^4 b_i^2 \cdot P[B = b_i] = 78,125$

f) Media o Esperanza beneficios: $\mu_B = E(B) = \sum_{i=1}^4 b_i \cdot P[B = b_i] = 5,625$

Varianza beneficios:

$$E[B^2] = \sum_{i=1}^4 b_i^2 \cdot P[B = b_i] = 78,125$$

$$\sigma_B^2 = \text{Var}(B) = E(B^2) - \mu_B^2 = 78,125 - (5,625)^2 = 46,48$$

$$\text{Desviación típica de los beneficios: } \sigma_B = \sqrt{46,48} = 6,817$$

69. Se ha verificado que la variable $X =$ "Peso en kilos de los niños al nacer" es una variable aleatoria continua con función de densidad

$$f(x) = \begin{cases} kx & 2 \leq x \leq 4 \\ 0 & \text{en otros casos} \end{cases}$$

Se pide:

- Hallar k para que $f(x)$ sea función de densidad. Representarla
- Hallar la función de distribución. Representarla
- Media, varianza y desviación típica
- Probabilidad de que un niño elegido al azar pese más de 3 kilos
- Probabilidad de que pese entre 2 y 3,5 kilos
- Qué debe pesar un niño para tener un peso igual o inferior al 90% de los niños

Solución:

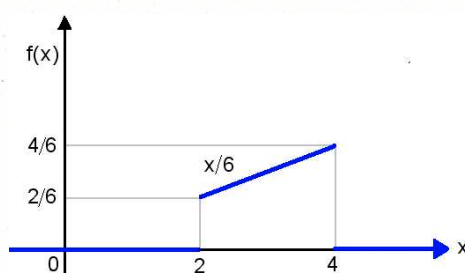
- Para que $f(x)$ sea función de densidad debe verificar:

$$1 = \int_{-\infty}^{\infty} f(x) dx = \int_{-\infty}^2 f(x) dx + \int_2^4 f(x) dx + \int_4^{\infty} f(x) dx = \int_2^4 f(x) dx$$

La primera y tercera integral son cero al ser $f(x) = 0$ en esos intervalos.

$$1 = \int_2^4 f(x) dx = \int_2^4 kx dx = k \int_2^4 x dx = k \left[\frac{x^2}{2} \right]_2^4 = k \left[\frac{16}{2} - \frac{4}{2} \right] = 6k \rightarrow k = \frac{1}{6}$$

$$f(x) = \begin{cases} \frac{x}{6} & 2 \leq x \leq 4 \\ 0 & \text{en otros casos} \end{cases}$$



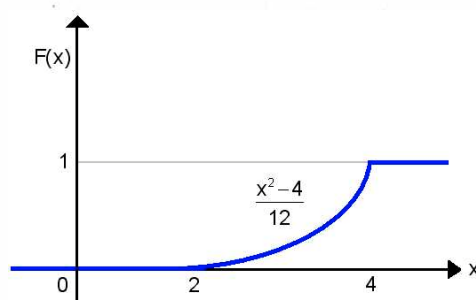
b) La función de distribución se define $F(x) = \int_{-\infty}^x f(t) dt$

$$x < 2 \quad F(x) = \int_{-\infty}^x f(t) dt = 0$$

$$2 \leq x \leq 4 \quad F(x) = \int_{-\infty}^x f(t) dt = \int_2^x f(t) dt = \int_2^x \frac{t}{6} dt = \frac{1}{6} \left[\frac{t^2}{2} \right]_2^x = \frac{1}{6} \left[\frac{x^2 - 4}{2} \right] = \frac{x^2 - 4}{12}$$

$$x > 4 \quad F(x) = \int_{-\infty}^x f(t) dt = \int_2^4 f(t) dt + \int_4^x f(t) dt = \int_2^4 \frac{t}{6} dt = \frac{1}{6} \left[\frac{t^2}{2} \right]_2^4 = \frac{1}{6} \left[\frac{16 - 4}{2} \right] = 1$$

$$F(x) = \begin{cases} 0 & x < 2 \\ \frac{x^2 - 4}{12} & 2 \leq x \leq 4 \\ 1 & x > 4 \end{cases}$$



c) Media

$$\alpha_1 = \mu_x = E(X) = \int_{-\infty}^{\infty} x f(x) dx = \int_2^4 x \cdot \frac{x}{6} dx = \frac{1}{6} \int_2^4 x^2 dx = \frac{1}{6} \left[\frac{x^3}{3} \right]_2^4 = \frac{1}{6} \left[\frac{64}{3} - \frac{8}{3} \right] = \frac{56}{18} = 3,1 \text{ kilos}$$

Varianza: $\sigma_x^2 = \alpha_2 - \alpha_1^2$

$$\alpha_2 = E(X^2) = \int_{-\infty}^{\infty} x^2 f(x) dx = \int_2^4 x^2 \cdot \frac{x}{6} dx = \frac{1}{6} \int_2^4 x^3 dx = \frac{1}{6} \left[\frac{x^4}{4} \right]_2^4 = \frac{1}{6} \left[\frac{256}{4} - \frac{16}{4} \right] = 10 \text{ kilos}^2$$

$$\sigma_x^2 = \alpha_2 - \alpha_1^2 = 10 - 3,1^2 = 0,39 \text{ kilos}^2$$

Desviación típica: $\sigma_x = \sqrt{0,39} = 0,62 \text{ kilos}$

$$c) P(X > 3) = 1 - P(X \leq 3) = 1 - F(3) = 1 - \frac{3^2 - 4}{12} = 1 - \frac{5}{12} = \frac{7}{12} = 0,58$$

O también, $P(X > 3) = \int_3^4 f(x) dx = \int_3^4 \frac{x}{6} dx = \frac{1}{6} \left[\frac{x^2}{2} \right]_3^4 = \frac{1}{6} \left(8 - \frac{9}{2} \right) = \frac{7}{12} = 0,58$

$$d) P(2 \leq X \leq 3,5) = F(3,5) - F(2) = \frac{3,5^2 - 4}{12} - 0 = 0,6875$$

$$P(2 \leq X \leq 3,5) = \int_2^{3,5} f(x) dx = \int_2^{3,5} \frac{x}{6} dx = \frac{1}{6} \left[\frac{x^2}{2} \right]_2^{3,5} = \frac{1}{6} \left(\frac{12,25}{2} - \frac{4}{2} \right) = \frac{8,25}{12} = 0,6875$$

e) Sea k el peso del niño, se tiene:

$$F(k) = P(X \leq k) = 0,9 \longrightarrow \frac{k^2 - 4}{12} = 0,9 \Rightarrow k^2 - 4 = 10,8 \Rightarrow k^2 = 14,8$$

$k = \sqrt{14,8} = 3,85$, es decir, el niño debe pesar 3,85 kilos para tener para tener al 90% de los niños con un peso igual o inferior.

70. La función de densidad de una variable aleatoria es:

$$f(x) = \begin{cases} ax^2 + b & 0 < x < 2 \\ 0 & \text{en el resto} \end{cases} \quad \text{Sabido que } P\left[\frac{1}{2} < x < 1\right] = 0,1666$$

Determinar a y b .

Solución:

Hay que calcular dos parámetros (a y b), por lo que se necesitan dos ecuaciones:

- Por ser función de densidad:

$$1 = \int_0^2 f(x) dx = \int_0^2 (ax^2 + b) dx = \left[a \frac{x^3}{3} + bx \right]_0^2 = \frac{8a}{3} + 2b \rightarrow 8a + 6b = 3$$

- $P\left[\frac{1}{2} \leq x \leq 1\right] = \int_{1/2}^1 f(x) dx = \int_{1/2}^1 (ax^2 + b) dx = \left[a \frac{x^3}{3} + bx \right]_{1/2}^1 = 0,1666$

con lo que:

$$\left[a \frac{x^3}{3} + bx \right]_{1/2}^1 = \left[\frac{a}{3} + b \right] - \left[\frac{a}{24} + \frac{b}{2} \right] = \frac{7a}{24} + \frac{b}{2} = 0,1666 \rightarrow 7a + 12b \approx 4$$

En consecuencia,

$$\left. \begin{array}{l} 8a + 6b = 3 \\ 7a + 12b = 4 \end{array} \right\} \rightarrow \left. \begin{array}{l} 16a + 12b = 6 \\ 7a + 12b = 4 \end{array} \right\} \rightarrow a = \frac{2}{9} = 0,22 \quad 6b = 3 - \frac{16}{9} = \frac{11}{9} \rightarrow b = \frac{11}{54} = 0,20$$

71. Sea X una variable aleatoria continua, cuya función de densidad es

$$f_x(x) = \begin{cases} 3x^2 & 0 < x < 1 \\ 0 & \text{en otro caso} \end{cases}$$

Sea $Y = 1 - X^2$ una transformación de la variable aleatoria X . Se pide:

- Calcular la función de densidad de la variable aleatoria Y
- Calcular la función de distribución de la variable aleatoria Y

Solución:

a) La transformación asociada a la v.a. Y es derivable y estrictamente monótona cuando X toma valores en el intervalo $(0, 1)$. En consecuencia, se puede aplicar la transformación, quedando la función de densidad:

$$Y = 1 - X^2 \rightarrow x = \sqrt{1 - y} \rightarrow \begin{cases} \frac{dx}{dy} = \frac{-1}{2\sqrt{1-y}} \\ g^{-1}(y) = \sqrt{1-y} \end{cases}$$

La función de densidad de la variable continua Y se obtiene:

$$f_y(y) = f_x[g^{-1}(y)] \cdot \left| \frac{dx}{dy} \right| = 3(\sqrt{1-y})^2 \left| \frac{-1}{2\sqrt{1-y}} \right| = \frac{3}{2}\sqrt{1-y}$$

La función de densidad de la v.a. Y :
$$f_y(y) = \begin{cases} \frac{3}{2}\sqrt{1-y} & 0 < y < 1 \\ 0 & \text{en otro caso} \end{cases}$$

b) Función de distribución:

$$y \leq 0 \quad F_Y(y) = \int_{-\infty}^y f(t) dt = 0$$

$$0 < y < 1 \quad F_Y(y) = \int_{-\infty}^0 f(t) dt + \int_0^y f(t) dt = \int_0^y \frac{3}{2} \sqrt{1-t} dt = \left[-\sqrt{(1-t)^3} \right]_0^y = 1 - \sqrt{(1-y)^3}$$

$$y \geq 1 \quad F_Y(y) = \int_{-\infty}^0 f(t) dt + \int_0^1 f(t) dt + \int_1^y f(t) dt = \int_0^1 f(t) dt = \int_0^1 \frac{3}{2} \sqrt{1-t} dt = 1$$

La función de distribución de la v.a. Y será:
$$F_Y(y) = \begin{cases} 0 & y \leq 0 \\ 1 - \sqrt{(1-y)^3} & 0 < y < 1 \\ 1 & y \geq 1 \end{cases}$$

72. Sea X una variable aleatoria continua, cuya función de densidad es

$$f_X(x) = \begin{cases} \frac{1}{2} & -1 < x < 1 \\ 0 & \text{en otro caso} \end{cases}$$

Sea $Y = X^2$ una transformación de la variable aleatoria X. Se pide:

- Calcular la función de densidad de la variable aleatoria Y
- Calcular la función de distribución de la variable aleatoria Y

Solución:

a) La transformación $Y = X^2$ es derivable, pero no es estrictamente monótona, puesto que en el intervalo $(-1, 0)$ la transformación es decreciente y en el intervalo $[0, 1)$ es creciente.

En este caso, hay que determinar la función de distribución de la variable aleatoria Y para el caso general de las transformaciones de una variable aleatoria, ya que no se puede aplicar el método descrito en el ejercicio anterior.

b) Cálculo de la función de distribución

$$F_Y(y) = P[Y \leq y] = P[X^2 \leq y] = P[|X| \leq \sqrt{y}] = P[-\sqrt{y} \leq X \leq \sqrt{y}] = \int_{-\sqrt{y}}^{\sqrt{y}} f(x) dx =$$

$$= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{2} dx = \frac{1}{2} [x]_{-\sqrt{y}}^{\sqrt{y}} = \sqrt{y}$$

La función de distribución de la v.a. Y es: $F_Y(y) = \begin{cases} 0 & y < 0 \\ \sqrt{y} & 0 \leq y < 1 \\ 1 & y \geq 1 \end{cases}$

a) La función de densidad $f_Y(y) = \frac{dF_Y(y)}{dy} = \begin{cases} \frac{1}{2\sqrt{y}} & 0 \leq y < 1 \\ 0 & \text{en otro caso} \end{cases}$

73. Sea X una variable aleatoria continua con función de densidad tal que

$$f(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & \text{otro caso} \end{cases} \quad \text{Se pide:}$$

- a) Función generatriz de los momentos (f.g.m.)
- b) Esperanza y varianza a partir de la f.g.m.
- c) Función característica

Solución:

$$a) M(t) = E[e^{tx}] = \int_{-\infty}^{\infty} e^{tx} \cdot f(x) \cdot dx = \int_0^{\infty} e^{x(t-1)} dx = \frac{1}{t-1} [e^{x(t-1)}]_0^{\infty} = \frac{1}{1-t} \quad \text{si } t < 1$$

$$b) \alpha_1 = E(X) = M^{(1)}(0) = \left. \frac{dM(t)}{dt} \right|_{t=0} = \left. \frac{d}{dt} \left[\frac{1}{1-t} \right] \right|_{t=0} = \left. \frac{1}{(1-t)^2} \right|_{t=0} = 1$$

$$\alpha_2 = E(X^2) = M^{(2)}(0) = \left. \frac{d^2 M(t)}{dt^2} \right|_{t=0} = \left. \frac{d}{dt} \left[\frac{d}{dt} \left(\frac{1}{1-t} \right) \right] \right|_{t=0} = \left. \frac{d}{dt} \left(\frac{1}{(1-t)^2} \right) \right|_{t=0} = \left. \frac{2}{(1-t)^3} \right|_{t=0} = 2$$

$$\text{Var}(X) = \alpha_2 - \alpha_1^2 = 2 - 1 = 1$$

c) La función característica se puede calcular utilizando la relación entre función característica y los momentos:

$$\varphi(t) = 1 + (it)\alpha_1 + \frac{(it)^2}{2!}\alpha_2 + \frac{(it)^3}{3!}\alpha_3 + \dots + \frac{(it)^k}{k!}\alpha_k + \dots = \sum_{j=0}^{\infty} \frac{(it)^j}{j!}\alpha_j \quad \text{si } t < 1$$

74. El tiempo de revisión del motor de un avión sigue aproximadamente una distribución exponencial, con media 22 minutos.

- Hallar la probabilidad de que el tiempo de la revisión sea menor de 10 minutos.
- El costo de la revisión es de 200 euros por cada media hora o fracción. ¿Cuál es la probabilidad de que una revisión cueste 400 euros?
- Para efectuar una programación sobre las revisiones del motor, ¿cuánto tiempo se debe asignar a cada revisión para que la probabilidad de que cualquier tiempo de revisión mayor que el tiempo asignado sea solo de 0,1?

Solución:

- a) Sea X = "tiempo de revisión del motor de un avión en minutos"

$$\mu_x = E(X) = \frac{1}{\lambda} = 22 \text{ minutos} \rightarrow \lambda = \frac{1}{22} \quad X \sim \text{Exp}\left[\frac{1}{22}\right]$$

Función de densidad:

$$f(x) = \begin{cases} \frac{1}{22} e^{-x/22} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Función distribución:

$$F(x) = P(X \leq x) = \begin{cases} 1 - e^{-x/22} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$P[X < 10] = F(10) = 1 - e^{-10/22} = 1 - e^{-5/11} = 0,365$$

$$\text{O bien, } P[X < 10] = \int_0^{10} f(x) dx = \int_0^{10} \left[\frac{1}{22} e^{-x/22} \right] dx = -\left[e^{-x/22} \right]_0^{10} = -e^{-10/22} + 1 = 1 - e^{-5/11} = 0,365$$

- b) Como el costo de la revisión del motor es de 200 euros por cada media hora o fracción, para que la revisión cueste 400 euros la duración de la revisión debe de ser inferior o igual a 60 minutos. Es decir, se tendrá que calcular $P[30 < X \leq 60]$

$$P[30 < X \leq 60] = F(60) - F(30) = \left[1 - e^{-60/22} \right] - \left[1 - e^{-30/22} \right] = e^{-30/22} - e^{-60/22} = e^{-15/11} - e^{-30/11} = 0,19$$

$$\begin{aligned} \text{O bien, } P[30 < X \leq 60] &= \int_{30}^{60} f(x) dx = \int_{30}^{60} \left[\frac{1}{22} e^{-x/22} \right] dx = -\left[e^{-x/22} \right]_{30}^{60} = -e^{-60/22} + e^{-30/22} = \\ &= e^{-15/11} - e^{-30/11} = 0,19 \end{aligned}$$

- c) Sea t = "tiempo que se debe asignar a la revisión", verificando $P[X > t] = 0,1$

$$P[X > t] = \int_t^{\infty} f(x) dx = \int_t^{\infty} \left[\frac{1}{22} e^{-x/22} \right] dx = -\left[e^{-x/22} \right]_t^{\infty} = 0 + e^{-t/22} = 0,1$$

$$e^{-t/22} = 0,1 \mapsto -t/22 = \ln(0,1) \mapsto -t/22 = -2,30 \rightarrow t = 50,6 \approx 51 \text{ minutos}$$

75. Una variable aleatoria X se distribuye uniformemente en $(2, 4)$. Se pide:

- a) $P(X < 2,5)$ b) $P(X > 3,2)$ c) $P(2,2 < X < 3,5)$ d) Esperanza y varianza

Solución:

Función de densidad:

$$f(x) = \begin{cases} \frac{1}{4-2} = \frac{1}{2} & 2 \leq x \leq 4 \\ 0 & \text{otros valores} \end{cases}$$

Función de distribución:

$$F(x) = P(X \leq x) = \begin{cases} 0 & x < 2 \\ \frac{x-2}{4-2} & 2 \leq x \leq 4 \\ 1 & x > 4 \end{cases}$$

a) $P(X < 2,5) = F(2,5) = \frac{2,5-2}{2} = 0,25$

O también, $P(X < 2,5) = \int_2^{2,5} f(x) dx = \int_2^{2,5} \frac{1}{2} dx = \frac{1}{2} [x]_2^{2,5} = \frac{2,5-2}{2} = 0,25$

b) $P(X > 3,2) = 1 - P(X \leq 3,2) = 1 - F(3,2) = 1 - \frac{3,2-2}{2} = \frac{4-3,2}{2} = 0,4$

O también, $P(X > 3,2) = \int_{3,2}^4 f(x) dx = \int_{3,2}^4 \frac{1}{2} dx = \frac{1}{2} [x]_{3,2}^4 = \frac{4-3,2}{2} = 0,4$

c) $P(2,2 < X < 3,5) = F(3,5) - F(2,2) = \frac{3,5-2}{2} - \frac{2,2-2}{2} = \frac{3,5-2,2}{2} = 0,65$

O también, $P(2,2 < X < 3,5) = \int_{2,2}^{3,5} f(x) dx = \int_{2,2}^{3,5} \frac{1}{2} dx = \frac{1}{2} [x]_{2,2}^{3,5} = \frac{3,5-2,2}{2} = 0,65$

d) $E(X) = \frac{2+4}{2} = 3$ $\sigma^2 = \frac{(4-2)^2}{12} = \frac{4}{12} = \frac{1}{3}$

76. Una empresa produce un artículo que sigue una distribución uniforme entre 25000 y 30000 unidades. Sabiendo que vende cada unidad a 10 euros y la función de costes viene dada por $C = 100000 + 2X$, ¿cuál será el beneficio esperado?

Solución:

$X \sim U(25000, 30000)$

Función de densidad $f(x) = \begin{cases} \frac{1}{30000-25000} = \frac{1}{5000} & 25000 \leq x \leq 30000 \\ 0 & \text{otros valores} \end{cases}$

Los beneficios $B = \text{Ventas} - \text{Costes}$

$$E(B) = E(V - C) = E[10 \cdot X - (100000 + 2 \cdot X)] = E(8 \cdot X - 100000) = 8 \cdot E(X) - 100000 =$$

$$= 8 \cdot 27500 - 100000 = 120000 \text{ euros}$$

siendo, $\mu = E(X) = \frac{25000 + 30000}{2} = 27500$

77. La duración de vida de una pieza de un motor sigue una distribución exponencial, sabiendo que la probabilidad de que sobrepase las 100 horas de uso es 0,9. Se pide:

- Probabilidad de que sobrepase las 200 horas de uso.
- ¿Cuántas horas se mantiene funcionando con probabilidad 0,95?

Solución:

a) Sea v.a. $X =$ "Tiempo de vida de la pieza del motor", donde $X \sim \text{Exp}(\lambda)$

Respectivamente, la función de densidad y la función de distribución:

$$f(x) = \lambda \cdot e^{-\lambda x} \quad F(x) = 1 - e^{-\lambda x} \quad \forall x > 0$$

$$\text{Siendo } P[X > 100] = 1 - P[X \leq 100] = 1 - F(100) = 1 - [1 - e^{-100\lambda}] = e^{-100\lambda} = 0,9$$

$$e^{-100\lambda} = 0,9 \rightarrow -100\lambda = \ln 0,9 \rightarrow -100\lambda = -0,105 \rightarrow \lambda = 0,00105$$

$$\text{Por tanto, } X \sim \text{Exp}(0,00105) \quad f(x) = 0,00105 \cdot e^{-0,00105 \cdot x} \quad F(x) = 1 - e^{-0,00105 \cdot x} \quad \forall x > 0$$

$$P[X > 200] = 1 - P[X \leq 200] = 1 - F(200) = 1 - [1 - e^{-0,00105 \cdot 200}] = 0,81$$

$$\text{Oo bien, } P[X > 200] = \int_{200}^{\infty} f(x) dx = \int_{200}^{\infty} 0,00105 \cdot e^{-0,00105 \cdot x} dx = -[e^{-0,00105 \cdot x}]_{200}^{\infty} = 0,81$$

$$\text{b) } P[X > t] = 0,95 \mapsto P[X > t] = 1 - P[X \leq t] = 1 - F(t) = e^{-0,00105 \cdot t} = 0,95$$

$$e^{-0,00105 \cdot t} = 0,95 \rightarrow -0,00105 t = \ln 0,95 \rightarrow -0,00105 t = -0,05129 \rightarrow t = 48,85$$

78. La cabina de un avión tiene 30 dispositivos electrónicos ($E_1, E_2, \dots, E_{30}$), tan pronto como falla E_1 se activa E_2 , y así sucesivamente. El tiempo de fallo de cualquier dispositivo electrónico E_i es de tipo exponencial con parámetro $\lambda = 0,1$ hora. Hallar la probabilidad de que el tiempo total de funcionamiento de todos los dispositivos supere las 350 horas.

Solución:

Sea X_i = " Fallo de cualquier dispositivo electrónico", $X_i \sim \text{Exp}(0,1)$ $i = 1, 2, \dots, 30$

$$\mu_i = E[X_i] = \frac{1}{\lambda} = \frac{1}{0,1} = 10 \quad \sigma_i = \frac{1}{\lambda} = 10$$

Por el Teorema Central del Límite, $X = \sum_{i=1}^{30} X_i$ sigue una distribución normal de media

$$\mu = 30 \cdot \mu_i = 300 \text{ y desviación típica } \sigma = \sqrt{\sum_{i=1}^{30} \sigma_i^2} = \sqrt{30 \cdot 10^2} = \sqrt{3000}$$

$$X = \sum_{i=1}^{30} X_i \sim N[300, \sqrt{3000}]$$

$$P[X \geq 350] = P\left[\frac{X - 300}{\sqrt{3000}} \geq \frac{350 - 300}{\sqrt{3000}}\right] = P[z \geq 0,91] = 0,1814$$

79. Un corredor de bolsa adquiere 50 acciones diferentes, concertando con sus clientes una ganancia de 1200 euros por acción. Por experiencias anteriores, se sabe que los beneficios de cada acción son independientes y se distribuyen uniformemente en el intervalo $[1000, 2000]$. ¿Qué probabilidad tiene el corredor de no perder dinero?.

Solución:

Denotando por X = " Ganancia por acción" y G = " Ganancia total del corredor de bolsa"

$$G = 50 \cdot (X - 1200) = 50 \cdot X - 60000$$

$$X \sim U(1000, 2000) \rightarrow f(x) = \frac{1}{2000 - 1000} = \frac{1}{1000} \quad F(x) = \frac{x - 1000}{2000 - 1000} \quad 1000 \leq x \leq 2000$$

$$P[G \geq 0] = P[50 \cdot X - 60000 \geq 0] = P[50 \cdot X \geq 60000] = P[X \geq 1200] =$$

$$= \int_{1200}^{2000} f(x) dx = \int_{1200}^{2000} \frac{1}{1000} dx = \frac{1}{1000} [x]_{1200}^{2000} = \frac{800}{1000} = 0,8$$

$$\text{O también, } P[X \geq 1200] = 1 - P[X \leq 1200] = 1 - F(1200) = 1 - \frac{1200 - 1000}{2000 - 1000} = 1 - \frac{200}{1000} = 0,8$$

80. El número promedio de recepción de solicitudes en una ventanilla de atención al cliente es de tres al día.

- ¿Cuál es la probabilidad de que el tiempo antes de recibir una solicitud exceda cinco días?
- ¿Cuál es la probabilidad de que el tiempo antes de recibir una solicitud sea menor de diez días?
- ¿Cuál es la probabilidad de que el tiempo antes de recibir una solicitud sea menor de diez días, si ya han pasado cinco días sin recibir solicitudes?

Solución:

X = "Días antes de recibir una solicitud", es una distribución exponencial con $\lambda = 3$

$$f(x) = 3e^{-3x} \quad F(x) = P(X \leq x) = 1 - e^{-3x}$$

$$P(X > 5) = 1 - P(X \leq 5) = 1 - F(5) = 1 - (1 - e^{-3 \cdot 5}) = e^{-15}$$

$$P(X < 10) = F(10) = 1 - e^{-3 \cdot 10} = 1 - e^{-30}$$

$$P(X < 10 \mid X > 5) = \frac{P(5 < X < 10)}{P(X > 5)} = \frac{F(10) - F(5)}{1 - F(5)} = \frac{1 - e^{-30} - (1 - e^{-15})}{e^{-15}} = \frac{e^{-15} - e^{-30}}{e^{-15}} = 1 - e^{-15}$$

Adviértase que $P(X < 10 \mid X > 5) = P(X \leq 5)$ lo que significa que la variable aleatoria exponencial no tiene memoria.

81. Una web tiene un número promedio de 5 visitantes por hora. Se piden las probabilidades:

- Sea visitada por un mínimo de 6 personas.
- Qué pase más de una hora sin que sea visitada

Solución:

a) X = "Número promedio de visitantes", distribución de Poisson de parámetro $\lambda = 5$

$$P(X \geq 6) = 1 - P(X \leq 5) = 1 - \sum_{k=0}^5 P(X = k) = 1 - \sum_{k=0}^5 \frac{5^k}{k!} e^{-5}$$

b) $X =$ " Horas sin ser visitada la web", distribución exponencial de parámetro $\lambda = 5$

$$f(x) = 5e^{-5x} \quad F(x) = P(X \leq x) = 1 - e^{-5x}$$

82. Sea X la variable aleatoria que describe el número de clientes que llega a un supermercado durante un día (24 horas). Sabiendo que la probabilidad de que llegue un cliente en un día es equivalente a 100 veces la que no llegue ningún cliente en un día, se pide:

- Probabilidad de que lleguen al menos 3 clientes al día.
- Si acaba de llegar un cliente, calcular la probabilidad que pase más de 25 minutos hasta que llegue el siguiente cliente (o hasta que llegue el primer cliente).
- En dos semanas, ¿cuál es la probabilidad aproximada de que lleguen como mucho 1300 clientes al supermercado?

Solución:

a) Se trata de una distribución de Poisson: $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$

$$P(X = 1) = 100 \cdot P(X = 0) \rightarrow \frac{\lambda}{1!} e^{-\lambda} = 100 \cdot \frac{\lambda^0}{0!} e^{-\lambda} \rightarrow \lambda \cdot e^{-\lambda} = 100 \cdot e^{-\lambda} \rightarrow \lambda = 100$$

$$P(X \geq 3) = P\left[z \geq \frac{3 - 100}{10}\right] = P(z \geq -9,7) = P(z \leq 9,7) \approx 1$$

b) Es una función exponencial, es decir, el tiempo de espera hasta que ocurre un suceso (que llegue el siguiente cliente), donde λ es el número de sucesos de Poisson por unidad de tiempo.
 $X \sim \text{Exp}(\lambda)$

Para calcular el parámetro λ se establece una proporción:

$$\frac{100}{24 \cdot 60} = \frac{\lambda}{25} \rightarrow \lambda = \frac{100 \cdot 25}{24 \cdot 60} = \frac{250}{144} = 1,736$$

$$P(X \leq x) = F(x) = 1 - e^{-\lambda \cdot x} \rightarrow P(X \leq 1) = F(1) = 1 - e^{-1,736 \cdot 1} = 0,8238$$

$$P(X \leq 1300) = P\left[z \leq \frac{1300 - 1400}{\sqrt{1400}}\right] = P(z \leq -2,67) = P(z \geq 2,67) = 0,00379$$

RELACIÓN ENTRE LA DISTRIBUCIÓN DE POISSON Y LA DISTRIBUCIÓN EXPONENCIAL



- Si el tiempo de llegadas consecutivas sigue una distribución exponencial $\text{Exp}(\lambda)$, entonces el número de llegadas en un intervalo de tiempo t sigue una distribución de Poisson $P(\lambda t)$
- Si el número de llegadas en una hora sigue una distribución de Poisson $P(\lambda)$, entonces el número de horas entre dos llegadas consecutivas sigue una distribución exponencial $\text{Exp}(\lambda)$

- Una propiedad que caracteriza de manera única a la distribución exponencial dentro del conjunto de distribuciones continuas es que satisface la *pérdida de memoria*, esto es, si $X \sim \text{exp}(\lambda)$ para cualquiera tiempos $s, t \geq 0$ se cumple la igualdad $P(X > t + s | X > s) = P(X > t)$
Por tanto, $P(X > t + s) = P(X > t) P(X > s)$

En consecuencia, la distribución exponencial $\text{exp}(\lambda)$ carece de memoria y es la única distribución continua con tal propiedad, ya que $P(X > t + s) = e^{-\lambda(t+s)} = e^{-\lambda t} e^{-\lambda s} = P(X > t) P(X > s)$

- La distribución exponencial tiene la *propiedad reproductiva*, que hace referencia a que la suma de distribuciones exponenciales independientes e idénticamente distribuidas sigue una distribución gamma $\Gamma(n, \lambda)$

En efecto, si $(X_1, X_2, \dots, X_n)$ son n variables independientes distribuidas exponencialmente,

$X_i \sim \text{Exp}(\lambda) \forall i$, entonces $X_1 + X_2 + \dots + X_n = \sum_{i=1}^n X_i \sim \Gamma(n, \lambda)$, cuya función de densidad es

$$f(t) = \frac{(\lambda t)^{n-1}}{(n-1)!} \lambda e^{-\lambda t} = \frac{\lambda^n t^{n-1}}{(n-1)!} e^{-\lambda t} \quad \forall t, n > 0$$

- Si $(X_1, X_2, \dots, X_n)$ son n variables aleatorias independientes con distribución exponencial $X_i \sim \text{exp}(\lambda) \forall i$, entonces $X = \min(X_1, X_2, \dots, X_n) \sim \text{exp}(\sum_i \lambda_i)$, es decir, el mínimo de n variables aleatorias exponenciales independientes se distribuye exponencialmente.

$$P(X > x) = P[\min(X_1, X_2, \dots, X_n) > x] = P(X_i > x, \forall i) = \prod_{i=1}^n P(X_i > x) = \prod_{i=1}^n e^{-\lambda_i x} = e^{-\left(\sum_{i=1}^n \lambda_i\right) x}$$

- Sean X_1 y X_2 dos variables aleatorias independientes con distribución exponencial de parámetros λ_1 y λ_2 respectivamente, por el teorema de probabilidad total, se tiene:

$$P(X_1 < X_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

83. La duración de vida de una pieza de un motor sigue una distribución exponencial, sabiendo que la probabilidad de que sobrepase las 100 horas de uso es 0,9. Se pide:

- a) Probabilidad de que sobrepase las 200 horas de uso.
b) ¿Cuántas horas se mantiene funcionando con probabilidad 0,95?

Solución:

a) Sea v.a. X = "tiempo de vida de la pieza del motor", donde $X \sim \text{Exp}(\lambda)$

Respectivamente, la función de densidad y la función de distribución:

$$f(x) = \lambda \cdot e^{-\lambda x} \quad F(x) = 1 - e^{-\lambda x} \quad \forall x > 0$$

$$\text{Siendo } P[X > 100] = 1 - P[X \leq 100] = 1 - F(100) = 1 - [1 - e^{-100\lambda}] = e^{-100\lambda} = 0,9$$

$$e^{-100\lambda} = 0,9 \rightarrow -100\lambda = \ln 0,9 \Rightarrow -100\lambda = -0,105 \Rightarrow \lambda = 0,00105$$

$$\text{Por tanto, } X \sim \text{Exp}(0,00105) \quad f(x) = 0,00105 \cdot e^{-0,00105 \cdot x} \quad F(x) = 1 - e^{-0,00105 \cdot x} \quad \forall x > 0$$

$$P[X > 200] = 1 - P[X \leq 200] = 1 - F(200) = 1 - [1 - e^{-0,00105 \cdot 200}] = 0,81$$

$$\text{O bien, } P[X > 200] = \int_{200}^{\infty} f(x) dx = \int_{200}^{\infty} 0,00105 \cdot e^{-0,00105 \cdot x} dx = -[e^{-0,00105 \cdot x}]_{200}^{\infty} = 0,81$$

$$\text{b) } P[X > t] = 0,95 \rightarrow P[X > t] = 1 - P[X \leq t] = 1 - F(t) = e^{-0,00105 \cdot t} = 0,95$$

$$e^{-0,00105 \cdot t} = 0,95 \rightarrow -0,00105 t = \ln 0,95 \rightarrow -0,00105 t = -0,05129 \rightarrow t = 48,85$$

84. En una tela las fallas siguen un proceso de Poisson a razón de 1 falla cada 10 metros. ¿Cuál es la probabilidad de que la distancia entre la cuarta y quinta falla sea mayor a un metro?

Solución:

En un proceso de Poisson el intervalo entre dos eventos consecutivos es una variable exponencial $\text{Exp}(\lambda)$ con $\lambda = 1/10 = 0,1$

$$P(X > 1) = \int_1^{\infty} \lambda e^{-\lambda x} dx = \int_1^{\infty} 0,1 e^{-0,1x} dx = -e^{-0,1x} \Big|_1^{\infty} = 0,9048$$

$$P(X > 1) = 1 - P(X \leq 1) = 1 - F(1) = 1 - (1 - e^{-0,1 \times 1}) = e^{-0,1} = 0,9048$$

85. La ciudad de Segovia se comunica con UAM mediante 3 autobuses por hora, distribuidos según un proceso de Poisson. ¿Cuál es la probabilidad de tener que esperar un autobús más de 20 minutos?

Solución:

Sea la variable aleatoria $X \equiv$ Tiempo de espera a un autobús donde $X \sim \text{Exp}(\lambda)$

$$\lambda = \frac{3 \text{ autobuses}}{\text{hora}} = \frac{3 \text{ autobuses}}{60 \text{ minutos}} = 0,05 \text{ por tanto } X \sim \text{Exp}(0,05)$$

$$P(X > 20) = \int_{20}^{\infty} \lambda e^{-\lambda x} dx = \int_{20}^{\infty} 0,05 e^{-0,05x} dx = -e^{-0,05x} \Big|_{20}^{\infty} = 0,3678$$

$$P(X > 1) = 1 - P(X \leq 20) = 1 - F(20) = 1 - (1 - e^{-0,05 \times 20}) = e^{-1} = 0,3678$$

86. El promedio de clientes que llegan a solicitar servicio a un banco es de cinco por minuto. ¿Cuál es la probabilidad de que dos clientes tarden entre 30 a 45 segundos en solicitar servicio?

Solución:

El número de clientes que llegan por minuto a solicitar servicio bancario es una variable aleatoria discreta de Poisson de parámetro λ

El tiempo que tardan n clientes en llegar a solicitar servicio al banco es una variable aleatoria continua Gamma $\Gamma(n, 1/\lambda)$

$$\lambda = \frac{5 \text{ clientes}}{\text{minuto}} = \frac{5 \text{ clientes}}{60 \text{ segundos}} = \frac{1}{12}$$

$$X \sim \Gamma(2, 12) \text{ con función de densidad } f(x) = \frac{\lambda^n x^{n-1}}{(n-1)!} e^{-\lambda x} \quad n=2 \quad \lambda = 1/12$$

$$\text{con lo cual, } P(30 \leq X \leq 45) = \int_{30}^{45} \frac{(1/12)^2}{(2-1)!} x^{2-1} e^{-x/12} dx = \frac{1}{12^2} \int_{30}^{45} x e^{-x/12} dx = 0,17558$$

$$\frac{1}{12^2} \int_{30}^{45} x e^{-x/12} dx \quad \begin{array}{l} \text{integración partes} \\ u=x \quad dv=e^{-x/12} dx \end{array} = \left(\frac{x}{12} + 1 \right) e^{-x/12} \Big|_{30}^{45} = -0,11171 + 0,28729 = 0,17558$$

87. El número promedio de recepción de solicitudes en una ventanilla de atención al cliente es de tres al día.

- ¿Cuál es la probabilidad de que el tiempo antes de recibir una solicitud exceda cinco días?
- ¿Cuál es la probabilidad de que el tiempo antes de recibir una solicitud sea menor de diez días?
- ¿Cuál es la probabilidad de que el tiempo antes de recibir una solicitud sea menor de diez días, si ya han pasado cinco días sin recibir solicitudes?

Solución:

a) $X =$ "Días antes de recibir una solicitud", $X \sim \text{Exp}(\lambda = 3)$

$$f(x) = 3e^{-3x} \quad F(x) = P(X \leq x) = 1 - e^{-3x}$$

$$P(X > 5) = 1 - P(X \leq 5) = 1 - F(5) = 1 - (1 - e^{-3 \cdot 5}) = e^{-15}$$

$$b) P(X < 10) = F(10) = 1 - e^{-3 \cdot 10} = 1 - e^{-30}$$

$$c) P(X < 10 \mid X > 5) = \frac{P(5 < X < 10)}{P(X > 5)} = \frac{F(10) - F(5)}{1 - F(5)} = \frac{1 - e^{-30} - (1 - e^{-15})}{e^{-15}} = \frac{e^{-15} - e^{-30}}{e^{-15}} = 1 - e^{-15}$$

Adviértase que $P(X < 10 \mid X > 5) = P(X \leq 5)$ lo que significa que la variable aleatoria exponencial no tiene memoria.

88. Sea X la variable aleatoria que describe el número de clientes que llega a un supermercado durante un día (24 horas). Sabiendo que la probabilidad de que llegue un cliente en un día es equivalente a 100 veces la que no llegue ningún cliente en un día, se pide:

- Probabilidad de que lleguen al menos 3 clientes al día.
- Si acaba de llegar un cliente, calcular la probabilidad que pase más de 25 minutos hasta que llegue el siguiente cliente (o hasta que llegue el primer cliente).
- En dos semanas, ¿cuál es la probabilidad aproximada de que lleguen como mucho 1300 clientes al supermercado?

Solución:

a) Se trata de una distribución de Poisson: $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$

$$P(X = 1) = 100 \cdot P(X = 0) \rightarrow \frac{\lambda}{1!} e^{-\lambda} = 100 \cdot \frac{\lambda^0}{0!} e^{-\lambda} \rightarrow \lambda \cdot e^{-\lambda} = 100 \cdot e^{-\lambda} \rightarrow \lambda = 100$$

$$P(X \geq 3) = P\left[z \geq \frac{3 - 100}{10}\right] = P(z \geq -9,7) = P(z \leq 9,7) \approx 1$$

b) Es una función exponencial, es decir, el tiempo de espera hasta que ocurre un suceso (que llegue el siguiente cliente), donde λ es el número de sucesos de Poisson por unidad de tiempo.
 $X \sim \text{Exp}(\lambda)$

Para calcular el parámetro λ se establece una proporción:

$$\frac{100}{24 \cdot 60} = \frac{\lambda}{25} \rightarrow \lambda = \frac{100 \cdot 25}{24 \cdot 60} = \frac{250}{144} = 1,736$$

$$P(X \leq x) = F(x) = 1 - e^{-\lambda \cdot x} \rightarrow P(X \leq 1) = F(1) = 1 - e^{-1,736 \cdot 1} = 0,8238$$

$$c) P(X \leq 1300) = P\left[z \leq \frac{1300 - 1400}{\sqrt{1400}}\right] = P(z \leq -2,67) = P(z \geq 2,67) = 0,00379$$

89. El tiempo de revisión del motor de un avión sigue aproximadamente una distribución exponencial, con media 22 minutos.

- Hallar la probabilidad de que el tiempo de la revisión sea menor de 10 minutos.
- El costo de la revisión es de 200 euros por cada media hora o fracción. ¿Cuál es la probabilidad de que una revisión cueste 400 euros?
- Para efectuar una programación sobre las revisiones del motor, ¿cuánto tiempo se debe asignar a cada revisión para que la probabilidad de que cualquier tiempo de revisión mayor que el tiempo asignado sea solo de 0,1?

Solución:

b) Sea $X =$ "Tiempo de revisión del motor de un avión en minutos"

$$\mu_x = E(X) = \frac{1}{\lambda} = 22 \text{ minutos} \rightarrow \lambda = \frac{1}{22} \quad X \sim \text{Exp}\left[\frac{1}{22}\right]$$

Función de densidad:

$$f(x) = \begin{cases} \frac{1}{22} e^{-x/22} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Función distribución:

$$F(x) = P(X \leq x) = \begin{cases} 1 - e^{-x/22} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$P[X < 10] = F(10) = 1 - e^{-10/22} = 1 - e^{-5/11} = 0,365$$

$$\text{O bien, } P[X < 10] = \int_0^{10} f(x) dx = \int_0^{10} \left[\frac{1}{22} e^{-x/22} \right] dx = -\left[e^{-x/22} \right]_0^{10} = -e^{-10/22} + 1 = 1 - e^{-5/11} = 0,365$$

b) Como el costo de la revisión del motor es de 200 euros por cada media hora o fracción, para que la revisión cueste 400 euros la duración de la revisión debe de ser inferior o igual a 60 minutos. Es decir, se tendrá que calcular $P[30 < X \leq 60]$

$$P[30 < X \leq 60] = F(60) - F(30) = [1 - e^{-60/22}] - [1 - e^{-30/22}] = e^{-30/22} - e^{-60/22} = e^{-15/11} - e^{-30/11} = 0,19$$

O bien,

$$P[30 < X \leq 60] = \int_{30}^{60} f(x) dx = \int_{30}^{60} \left[\frac{1}{22} e^{-x/22} \right] dx = - \left[e^{-x/22} \right]_{30}^{60} = -e^{-60/22} + e^{-30/22} = e^{-15/11} - e^{-30/11} = 0,19$$

c) Sea t = "Tiempo que se debe asignar a la revisión", verificando $P[X > t] = 0,1$

$$P[X > t] = \int_t^{\infty} f(x) dx = \int_t^{\infty} \left[\frac{1}{22} e^{-x/22} \right] dx = - \left[e^{-x/22} \right]_t^{\infty} = 0 + e^{-t/22} = 0,1$$

$$e^{-t/22} = 0,1 \rightarrow -t/22 = \ln(0,1) \rightarrow -t/22 = -2,30 \rightarrow t = 50,6 \approx 51 \text{ minutos}$$

90. Un sistema esta formado con dos procesadores, los procesadores 1 y 2 funcionan correctamente durante un tiempo exponencial de media de 5 y 6 años, respectivamente. Se pide:

a) ¿Cuál es la probabilidad de que ambos procesadores funcionen correctamente más de 4 años?

b) ¿Cuál es la probabilidad de que el procesador 1 deje de funcionar correctamente antes que el procesador 2?

Solución:

a) Sea $X_i \equiv$ tiempo de funcionamiento correcto del procesor i -ésimo, $X_i \sim \text{Exp}(\lambda_i)$ con $\lambda_1 = \frac{1}{5}$

$$\text{y } \lambda_2 = \frac{1}{6}$$

$$P(X > 4) = P[\min(X_1, X_2) > 4] = \prod_{i=1}^2 e^{-\lambda_i \cdot 4} = e^{-\left(\sum_{i=1}^2 \lambda_i\right) \cdot 4} = e^{-\left(\frac{1}{5} + \frac{1}{6}\right) \cdot 4} = 0,2307$$

$$\text{b) } P(X_1 < X_2) = \frac{\lambda_1}{\lambda_1 + \lambda_2} \rightarrow P(X_1 < X_2) = \frac{1/5}{1/5 + 1/6} = 0,545$$

91. La variable aleatoria poblacional "renta de las familias" del municipio de Madrid se distribuye siguiendo un modelo $N(\mu, \sigma)$. Se extraen muestras aleatorias simples de tamaño 4.

Como estimadores del parámetro μ , se proponen los siguientes:

$$\hat{\mu}_1 = \frac{x_1 + 2x_2 + 3x_3}{6} \quad \hat{\mu}_2 = \frac{x_3 - 4x_2}{-3} \quad \hat{\mu}_3 = \bar{x}$$

Se pide:

- Comprobar si los estimadores son insesgados
- ¿Cuál es el estimador más eficiente?
- Si tuviera que escoger entre ellos, ¿cuál escogería?. Razone su respuesta a partir del Error Cuadrático Medio.

Solución:

a) Un estimador $\hat{\theta}$ es insesgado (o centrado) cuando se verifica $E(\hat{\theta}) = \theta$

$$E(\hat{\mu}_1) = E\left[\frac{x_1 + 2x_2 + 3x_3}{6}\right] = \frac{1}{6} E[x_1 + 2x_2 + 3x_3] = \frac{1}{6} [E(x_1) + 2E(x_2) + 3E(x_3)] = \frac{1}{6} [6\mu] = \mu$$

$$E(\hat{\mu}_2) = E\left[\frac{x_1 - 4x_2}{-3}\right] = -\frac{1}{3} E[x_1 - 4x_2] = -\frac{1}{3} [E(x_1) - 4E(x_2)] = -\frac{1}{3} [-3\mu] = \mu$$

$$E(\hat{\mu}_3) = E\left[\frac{x_1 + x_2 + x_3 + x_4}{4}\right] = \frac{1}{4} E[x_1 + x_2 + x_3 + x_4] = \frac{1}{4} [E(x_1) + E(x_2) + E(x_3) + E(x_4)] = \\ = \frac{1}{4} [4\mu] = \mu$$

Los tres estimadores son insesgados o centrados.

b) El estimador *más eficiente* es el que tenga menor varianza.

$$V[\hat{\mu}_1] = V\left[\frac{x_1 + 2x_2 + 3x_3}{6}\right] = \frac{1}{36} V[x_1 + 2x_2 + 3x_3] = \frac{1}{36} [V(x_1) + 4V(x_2) + 9V(x_3)] = \\ = \frac{1}{36} [14\sigma^2] = \frac{14}{36} \sigma^2 = 0,39\sigma^2$$

$$V[\hat{\mu}_2] = V\left[\frac{x_1 - 4x_2}{-3}\right] = \frac{1}{9} V[x_1 - 4x_2] = \frac{1}{9} [V(x_1) + 16V(x_2)] = \frac{1}{9} [17\sigma^2] = \frac{17}{9} \sigma^2 = 1,89\sigma^2$$

$$V[\hat{\mu}_3] = V\left[\frac{x_1 + x_2 + x_3 + x_4}{4}\right] = \frac{1}{16} V[x_1 + x_2 + x_3 + x_4] = \frac{1}{16} [V(x_1) + V(x_2) + V(x_3) + V(x_4)] = \\ = \frac{1}{16} [4\sigma^2] = \frac{4}{16} \sigma^2 = 0,25 \sigma^2$$

El estimador $\hat{\mu}_3$ es el más eficiente.

c) Se elige el estimador que presente menor Error Cuadrático Medio (ECM)

$$ECM(\hat{\theta}) = E(\hat{\theta} - \theta)^2 = V(\hat{\theta}) + \underbrace{\left[E(\hat{\theta}) - \theta\right]^2}_{\text{sesgo}} \quad \text{sesgo} \equiv b(\hat{\theta}) = [E(\hat{\theta}) - \theta]$$

Adviértase que si el estimador $\hat{\theta}$ es insesgado: $\underbrace{E(\hat{\theta})}_{\text{insesgado}} = \theta \Rightarrow ECM(\hat{\theta}) = V(\hat{\theta})$

Al ser los tres estimadores insesgados (centrados), se elige al que menor varianza presenta, que coincidirá con el que tiene menor ECM, se opta por el estimador $\hat{\mu}_3$

92. El peso en kilos de los jamones vendidos por una empresa sigue una distribución normal con varianza 4 y peso medio desconocido. Se conoce que el peso medio de los jamones vendidos es superior a 5 kg, y se toman m.a.s. de tamaño 4 para estimar θ . ¿Cuál de los dos estimadores sería el mejor respondiendo a la insesgader y eficiencia?

$$\hat{\theta}_1 = \frac{x_1 + x_2 + x_3}{4} \quad \hat{\theta}_2 = \frac{x_1 + x_2}{2}$$

Solución:

La v.a X_i = 'peso en kg de los jamones' sigue una distribución normal de varianza $\sigma^2 = 4$

Para estudiar la insesgader de los estimadores se calculan sus esperanzas:

$$E(\hat{\theta}_1) = E\left[\frac{x_1 + x_2 + x_3}{4}\right] = \frac{1}{4} [E(x_1) + E(x_2) + E(x_3)] = \frac{3}{4} \theta$$

$$\text{El sesgo del estimador } \hat{\theta}_1 \text{ será: } b(\hat{\theta}_1) = E(\hat{\theta}_1) - \theta = \frac{3}{4} \theta - \theta = -\frac{1}{4} \theta$$

$$E(\hat{\theta}_2) = E\left[\frac{x_1 + x_2}{2}\right] = \frac{1}{2} [E(x_1) + E(x_2)] = \frac{2}{2} \theta = \theta$$

El estimador $\hat{\theta}_2$ es insesgado, $b(\hat{\theta}_2) = 0$

Atendiendo al sesgo se elige $\hat{\theta}_2$

Para analizar la eficiencia relativa de los dos estimadores se calculan las respectivas varianzas

$$\begin{aligned}
 V(\hat{\theta}_1) &= V\left[\frac{x_1 + x_2 + x_3}{4}\right] = \frac{1}{16} \underbrace{[V(x_1 + x_2 + x_3)]}_{\text{Observaciones independientes}} = \\
 &= \frac{1}{16} [V(x_1) + V(x_2) + V(x_3)] \underset{V(x_i)=4}{=} \frac{1}{16} 12 = \frac{12}{16} = \frac{3}{4}
 \end{aligned}$$

$$V(\hat{\theta}_2) = V\left[\frac{x_1 + x_2}{2}\right] = \frac{1}{4} [V(x_1 + x_2)] = \frac{1}{4} [V(x_1) + V(x_2)] = \frac{1}{4} 8 = 2$$

Respecto a la varianza se elige el estimador $\hat{\theta}_1$ por ser el de menor varianza.

Aparecen decisiones contrapuestas, de modo que el estimador se elige en base al error cuadrático medio: $ECM = Varianza + (\text{sesgo})^2$

$$ECM(\hat{\theta}_1) = \frac{3}{4} + \left(-\frac{\theta}{4}\right)^2 = \frac{\theta^2 + 12}{16} \quad ECM(\hat{\theta}_2) = 2 + 0 = 2$$

Se analiza cuando es mayor el ECM en el primer estimador $\hat{\theta}_1$:

$$ECM(\hat{\theta}_1) > ECM(\hat{\theta}_2) \rightarrow \frac{\theta^2 + 12}{16} > 2 \rightarrow \theta^2 > 20 \rightarrow |\theta| > \sqrt{20} \approx 4,47$$

Si θ es en valor absoluto mayor que 4,47, el error cuadrático medio de $\hat{\theta}_1$ es mayor, con lo que se elige el estimador $\hat{\theta}_2$.

Se conoce que el peso medio de los jamones es superior a 5 kg, no queda duda que el estimador a elegir (con menor error cuadrático medio) es $\hat{\theta}_2$

93. La distribución del peso de las manzanas de una determinada cosecha sigue una distribución normal, cuyo peso medio es desconocido y cuya desviación típica es 7 gramos. Se pide:

a) Analizar cuál de los estimadores $\hat{\mu}_1$ y $\hat{\mu}_2$ del peso medio es mejor respecto del sesgo y de la eficiencia, para una muestra aleatoria simple de tamaño cinco.

b) Si $\hat{\mu}_1 = \frac{\sum_{i=1}^5 X_i}{5}$ y $\hat{\mu}_2 = X_1 + 2X_2 + 3X_3 - 4X_4 - X_5$, obtener los pesos medios estimados a partir de la siguiente muestra (125, 135, 130, 137, 142).

Solución:

a) El peso de las manzanas sigue una distribución $N(\mu, 7)$

Se calculan las esperanzas de los estimadores para analizar el sesgo de los estimadores

$$E(\hat{\mu}_1) = E\left(\frac{1}{5} \sum_{i=1}^5 X_i\right) = \frac{1}{5} E\left[\sum_{i=1}^5 X_i\right] = \frac{1}{5} \sum_{i=1}^5 E[X_i] = \frac{1}{5} (5\mu) = \mu$$

$$E(\hat{\mu}_2) = E(X_1 + 2X_2 + 3X_3 - 4X_4 - X_5) = E(X_1) + 2E(X_2) + 3E(X_3) - 4E(X_4) - E(X_5) = \\ = \mu + 2\mu + 3\mu - 4\mu - \mu = \mu$$

Los estimadores $\hat{\mu}_1$, $\hat{\mu}_2$ son insesgados (centrados).

b) Para analizar la eficiencia de los estimadores se calculan sus varianzas:

$$V(\hat{\mu}_1) = V\left[\frac{1}{5} \sum_{i=1}^5 X_i\right] = \frac{1}{25} V\left[\sum_{i=1}^5 X_i\right] = \frac{1}{25} \sum_{i=1}^5 V[X_i] = \frac{1}{25} (5 \cdot 49) = \frac{49}{5}$$

$$V(\hat{\mu}_2) = V(X_1 + 2X_2 + 3X_3 - 4X_4 - X_5) = V(X_1) + 2^2 V(X_2) + 3^2 V(X_3) + (-4)^2 V(X_4) + (-1)^2 V(X_5) = \\ = 49 + 4 \cdot 49 + 9 \cdot 49 + 16 \cdot 49 + 49 = 31 \cdot 49 = 1519$$

Como los dos estimadores son insesgados y $V(\hat{\mu}_1) < V(\hat{\mu}_2)$ se elige como mejor el estimador $\hat{\mu}_1$ que es el peso medio de la muestra de las cinco manzanas.

94. Sea x_1, x_2, x_3, x_4, x_5 una muestra aleatoria simple de una variable aleatoria X con distribución normal con media $(\mu - 5)$ y varianza σ^2 .

Se proponen los siguientes estimadores: $\hat{\mu}_1 = \sum_{i=1}^5 x_i$ $\hat{\mu}_2 = 8x_2 - 3x_5$

Determinar cuál es el mejor estimador para μ . Justificar la respuesta.

Solución:

a) Estudio de la insesgadez

$$E(\hat{\mu}_1) = E\left(\sum_{i=1}^5 x_i\right) = E(x_1 + x_2 + x_3 + x_4 + x_5) = 5(\mu - 5)$$

$$E(\hat{\mu}_2) = E[8x_2 - 3x_5] = 8E(x_2) - 3E(x_5) = 8(\mu - 5) - 3(\mu - 5) = 5(\mu - 5)$$

ambos estimadores son sesgados, con idéntico sesgo:

$$b(\hat{\mu}_1) = b(\hat{\mu}_2) = 5(\mu - 5) - \mu = 4\mu - 25$$

b) Estudio de la varianza

$$V(\mu_1) = V\left(\sum_{i=1}^5 x_i\right) = 5V(x) = 5\sigma^2$$

$$V(\mu_2) = V(8x_2 - 3x_5) = 8^2 V(x_2) + 3^2 V(x_5) = 64\sigma^2 + 9\sigma^2 = 73\sigma^2$$

Dado que los dos estimadores tienen el mismo sesgo y el primer estimador μ_1 tiene menor varianza, será el estimador óptimo.

95. Sea una población con media μ de la que se extraen m.a.s. de tamaño n . Considere los

siguientes estimadores de la media: $\hat{\mu}_1 = \bar{x}$ $\hat{\mu}_2 = \frac{1}{n+1} \sum_{i=1}^n x_i$

- a) Estudie la insesgader, la eficiencia relativa y la consistencia de ambos estimadores
b) Elija uno de los dos en término del error cuadrático medio

Solución:

⇒ Insesgader

$$E(\hat{\mu}_1) = E(\bar{x}) = E\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n} E\left(\sum_{i=1}^n x_i\right) = \frac{1}{n} \sum_{i=1}^n E(x_i) = \frac{1}{n} (n\mu) = \mu$$

$$b(\hat{\mu}_1) = E(\hat{\mu}_1) - \mu = \mu - \mu = 0 \rightarrow \text{Estimador } \hat{\mu}_1 \text{ es centrado}$$

$$E(\hat{\mu}_2) = E\left(\frac{1}{n+1} \sum_{i=1}^n x_i\right) = \frac{1}{n+1} E\left(\sum_{i=1}^n x_i\right) = \frac{1}{n+1} \sum_{i=1}^n E(x_i) = \frac{1}{n+1} (n\mu) = \frac{n\mu}{n+1}$$

Un estimador $\hat{\theta}$ es *asintóticamente insesgado* si el sesgo tiende a cero al aumentar el tamaño muestral, es decir: $\lim_{n \rightarrow \infty} b(\hat{\theta}) = 0$

$$b(\hat{\mu}_2) = E(\hat{\mu}_2) - \mu = \frac{n\mu}{n+1} - \mu = \frac{n\mu - n\mu - \mu}{n+1} = -\frac{\mu}{n+1}$$

$$\lim_{n \rightarrow \infty} b(\hat{\mu}_2) = \lim_{n \rightarrow \infty} \left(-\frac{\mu}{n+1}\right) = 0 \rightarrow \text{Estimador } \hat{\mu}_2 \text{ asintóticamente insesgado}$$

⇒ La eficiencia relativa se mide por el ratio del cociente de varianzas: $\frac{V(\hat{\mu}_1)}{V(\hat{\mu}_2)}$

$$V(\hat{\mu}_1) = V(\bar{x}) = V\left(\frac{1}{n} \sum_{i=1}^n x_i\right) = \frac{1}{n^2} \sum_{i=1}^n V(x_i) = \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n}$$

$$V(\hat{\mu}_2) = V\left(\frac{1}{n+1} \sum_{i=1}^n x_i\right) = \frac{1}{(n+1)^2} \sum_{i=1}^n V(x_i) = \frac{1}{(n+1)^2} (n\sigma^2) = \frac{n}{(n+1)^2} \sigma^2$$

$$\text{Eficiencia relativa} \equiv \frac{V(\hat{\mu}_1)}{V(\hat{\mu}_2)} = \frac{\sigma^2/n}{n\sigma^2/(n+1)^2} = \frac{(n+1)^2}{n^2} > 1 \rightarrow V(\hat{\mu}_1) > V(\hat{\mu}_2)$$

El estimador $\hat{\mu}_2$ tiene menor varianza, por lo que es *más eficiente* que $\hat{\mu}_1$

⇒ **Consistencia:** Un estimador $\hat{\theta}$ *consistente* es un estimador asintóticamente insesgado cuya varianza tiende a cero al aumentar el tamaño muestral.

El estimador $\hat{\theta}$ es *consistente* cuando $\lim_{n \rightarrow \infty} E(\hat{\theta}) = \theta$ y $\lim_{n \rightarrow \infty} V(\hat{\theta}) = 0$

$$\hat{\mu}_1 \equiv \lim_{n \rightarrow \infty} E(\hat{\mu}_1) = \lim_{n \rightarrow \infty} \mu = \mu \quad \lim_{n \rightarrow \infty} V(\hat{\mu}_1) = \lim_{n \rightarrow \infty} \frac{\sigma^2}{n} = 0 \quad \text{es consistente}$$

$$\hat{\mu}_2 \equiv \lim_{n \rightarrow \infty} E(\hat{\mu}_2) = \lim_{n \rightarrow \infty} \left(\mu - \frac{1}{n+1} \mu \right) = \mu \quad \lim_{n \rightarrow \infty} V(\hat{\mu}_2) = \lim_{n \rightarrow \infty} \left[\frac{n}{(n+1)^2} \sigma^2 \right] = 0 \quad \text{es consistente}$$

$$\text{b) } ECM(\hat{\mu}_1) = V(\hat{\mu}_1) + [b(\hat{\mu}_1)]^2 = \frac{\sigma^2}{n} + 0 = \frac{\sigma^2}{n}$$

$$ECM(\hat{\mu}_2) = V(\hat{\mu}_2) + [b(\hat{\mu}_2)]^2 = \frac{n}{(n+1)^2} \sigma^2 + \left(-\frac{1}{n+1} \mu \right)^2 = \frac{n\sigma^2 + \mu^2}{(n+1)^2}$$

Para analizar que estimador presenta menor ECM:

$$\frac{\sigma^2}{n} \leq \frac{n\sigma^2 + \mu^2}{(n+1)^2} = \frac{n\sigma^2}{(n+1)^2} + \frac{\mu^2}{(n+1)^2} \rightarrow \frac{\sigma^2}{n} - \frac{n\sigma^2}{(n+1)^2} \leq \frac{\mu^2}{(n+1)^2}$$

$$\frac{(n+1)^2 \sigma^2 - n^2 \sigma^2}{n(n+1)^2} \leq \frac{\mu^2}{(n+1)^2} \rightarrow \frac{(n+1)^2 - n^2}{n} \sigma^2 \leq \mu^2$$

$$\frac{2n+1}{n} \sigma^2 \leq \mu^2 \rightarrow \frac{2n+1}{n} \leq \frac{\mu^2}{\sigma^2}$$

$$\text{Se elige } \hat{\mu}_1 \text{ cuando } \frac{2n+1}{n} \leq \frac{\mu^2}{\sigma^2} \quad \text{Se elige } \hat{\mu}_2 \text{ cuando } \frac{2n+1}{n} \geq \frac{\mu^2}{\sigma^2}$$

96. Supongamos que la distribución de ingresos de una cierta población es una variable aleatoria con media μ desconocida y varianza σ^2 también desconocida.

Si se desea estimar el ingreso medio de la población mediante una m.a.s. (muestra aleatoria simple) de tamaño n , respecto de la insesgadez y de la eficiencia.

¿Cuál de los dos estimadores elegiríamos?. ¿Son consistentes?

$$\hat{\mu}_1 = \frac{\sum_{i=1}^n x_i}{n-1} \quad \hat{\mu}_2 = \frac{\sum_{i=1}^n x_i}{n}$$

Solución:

La v.a X_i = 'Ingresos de cierta población' sigue una distribución normal $N(\mu, \sigma)$

▪ Sesgo de los estimadores:

$$E(\hat{\mu}_1) = E\left[\sum_{i=1}^n x_i / (n-1)\right] = \frac{1}{n-1} E\left[\sum_{i=1}^n x_i\right] = \frac{1}{n-1} \sum_{i=1}^n E(x_i) = \frac{1}{n-1} (n\mu) = \frac{n}{n-1} \mu$$

El sesgo del estimador $\hat{\mu}_1$ será: $b(\hat{\mu}_1) = E(\hat{\mu}_1) - \mu = \frac{n}{n-1} \mu - \mu = \frac{1}{n-1} \mu$

$$E(\hat{\mu}_2) = E\left[\sum_{i=1}^n x_i / n\right] = \frac{1}{n} E\left[\sum_{i=1}^n x_i\right] = \frac{1}{n} \sum_{i=1}^n E(x_i) = \frac{1}{n} (n\mu) = \mu$$

El estimador $\hat{\mu}_2$, que es la media muestral, es insesgado (centrado).

▪ La eficiencia de los estimadores se analiza a través de su varianza:

$$V(\hat{\mu}_1) = V\left[\sum_{i=1}^n x_i / (n-1)\right] = \frac{1}{(n-1)^2} V\left[\sum_{i=1}^n x_i\right] = \frac{1}{(n-1)^2} \sum_{i=1}^n V(x_i) = \frac{1}{(n-1)^2} (n\sigma^2) = \frac{n\sigma^2}{(n-1)^2}$$

$$V(\hat{\mu}_2) = V\left[\sum_{i=1}^n x_i / n\right] = \frac{1}{n^2} V\left[\sum_{i=1}^n x_i\right] = \frac{1}{n^2} \sum_{i=1}^n V(x_i) = \frac{1}{n^2} (n\sigma^2) = \frac{\sigma^2}{n}$$

El estimador más eficiente será el de menor varianza. Comparando las varianzas de los estimadores:

$$V(\hat{\mu}_2) = \frac{\sigma^2}{n} < \frac{n\sigma^2}{(n-1)^2} = V(\hat{\mu}_1)$$

El estimador $\hat{\mu}_2$, que es la media muestral, es el mejor tanto al sesgo como a la eficiencia.

- Los dos estimadores son consistentes:

$$\hat{\mu}_1 \text{ es consistente: } \lim_{n \rightarrow \infty} E(\hat{\mu}_1) = \mu \quad \lim_{n \rightarrow \infty} V(\hat{\mu}_1) = \lim_{n \rightarrow \infty} \frac{n\sigma^2}{(n-1)^2} = 0$$

$$\hat{\mu}_2 \text{ es consistente: } \lim_{n \rightarrow \infty} E(\hat{\mu}_2) = \mu \quad \lim_{n \rightarrow \infty} V(\hat{\mu}_2) = \lim_{n \rightarrow \infty} \frac{\sigma^2}{n} = 0$$

97. La variable aleatoria X representa los gastos mensuales de una empresa, cuya función de densidad es $f(\theta, x) = \theta x^{\theta-1}$ con $\theta > 0$ y $0 < x < 1$. Se realiza una m.a.s. de tamaño 3, y se proponen tres estimadores:

$$\hat{\theta}_1 = \bar{x} \quad \hat{\theta}_2 = \frac{x_1^2 + 2x_2^2 + 3x_3^2}{6} \quad \hat{\theta}_3 = \frac{x_3 - 2x_1 + 4x_2}{6}$$

- a) Calcule los sesgos.
b) Si la muestra que se obtiene es $(0,7, 0,1, 0,3)$, calcule las estimaciones puntuales.
c) ¿Cuáles son las funciones estimadas para las estimaciones anteriores?

Solución:

a) Un estimador $\hat{\theta}$ es insesgado (centrado) cuando $E(\hat{\theta}) = \theta$.

Un estimador $\hat{\theta}$ es sesgado cuando $E(\hat{\theta}) = \theta + \underbrace{b(\hat{\theta})}_{\text{sesgo}} \rightarrow b(\hat{\theta}) = E(\hat{\theta}) - \theta$

X = "gastos mensuales de la empresa" $f(\theta, x) = \theta x^{\theta-1}$ con $\theta > 0$ y $0 < x < 1$ m.a.s. con $n = 3$

- Sesgo del estimador $\hat{\theta}_1 = \bar{x}$

$$\hat{\theta}_1 = \bar{x} \rightarrow E(\hat{\theta}_1) = E\left[\frac{x_1 + x_2 + x_3}{3}\right] = \frac{1}{3} E[x_1 + x_2 + x_3] = \frac{1}{3} (3\mu) = \mu \text{ media poblacional}$$

$$\text{donde } \mu = \int_{-\infty}^{\infty} x f(x, \theta) dx = \int_0^1 x f(x, \theta) dx = \int_0^1 x \theta x^{\theta-1} dx = \int_0^1 \theta x^{\theta} dx = \left[\frac{\theta x^{\theta+1}}{\theta+1} \right]_0^1 = \frac{\theta}{\theta+1}$$

$$E(\hat{\theta}_1) = \mu = \frac{\theta}{\theta+1} \rightarrow \text{Sesgo } \hat{\theta}_1 = b(\hat{\theta}_1) = E(\hat{\theta}_1) - \theta = \frac{\theta}{\theta+1} - \theta = -\frac{\theta^2}{\theta+1}$$

- Sesgo del estimador $\hat{\theta}_2 = \frac{x_1^2 + 2x_2^2 + 3x_3^2}{6}$

$$E(\hat{\theta}_2) = E\left[\frac{x_1^2 + 2x_2^2 + 3x_3^2}{6}\right] = \frac{1}{6} \left[\underbrace{E(x_1^2)}_{\alpha_2} + 2\underbrace{E(x_2^2)}_{\alpha_2} + 3\underbrace{E(x_3^2)}_{\alpha_2} \right] = \frac{1}{6}(6\alpha_2) = \alpha_2 (*)$$

donde α_2 es el momento de orden 2 respecto al origen.

$$\alpha_2 = E(x^2) = \int_{-\infty}^{\infty} x^2 f(x, \theta) dx = \int_0^1 x^2 f(x, \theta) dx = \int_0^1 x^2 \theta x^{\theta-1} dx = \int_0^1 \theta x^{\theta+1} dx = \left[\frac{\theta x^{\theta+2}}{\theta+2} \right]_0^1 = \frac{\theta}{\theta+2}$$

$$\text{entonces, } E(\hat{\theta}_2) = E\left[\frac{x_1^2 + 2x_2^2 + 3x_3^2}{6}\right] = \frac{1}{6} \left[\underbrace{E(x_1^2)}_{\alpha_2} + 2\underbrace{E(x_2^2)}_{\alpha_2} + 3\underbrace{E(x_3^2)}_{\alpha_2} \right] = \alpha_2 = \frac{\theta}{\theta+2}$$

$$\text{Sesgo } \hat{\theta}_2 = b(\hat{\theta}_2) = E(\hat{\theta}_2) - \theta = \frac{\theta}{\theta+2} - \theta = -\frac{\theta^2 + \theta}{\theta+2}$$

- Sesgo del estimador $\hat{\theta}_3 = \frac{x_3 - 2x_1 + 4x_2}{6}$

$$E(\hat{\theta}_3) = E\left[\frac{x_3 - 2x_1 + 4x_2}{6}\right] = \frac{1}{6} E[x_3 - 2x_1 + 4x_2] = \frac{1}{6} (3\mu) = \frac{1}{2} \mu$$

$$\mu = \int_{-\infty}^{\infty} x f(x, \theta) dx = \int_0^1 x f(x, \theta) dx = \int_0^1 x \theta x^{\theta-1} dx = \int_0^1 \theta x^{\theta} dx = \left[\frac{\theta x^{\theta+1}}{\theta+1} \right]_0^1 = \frac{\theta}{\theta+1}$$

$$\text{Sesgo } \hat{\theta}_3 = b(\hat{\theta}_3) = E(\hat{\theta}_3) - \theta = \frac{1}{2} \left(\frac{\theta}{\theta+1} \right) - \theta = -\frac{2\theta^2 + \theta}{2(\theta+1)}$$

b) Si la muestra que se obtiene es (0,7 , 0,1 , 0,3) las estimaciones puntuales:

$$\hat{\theta}_1 = \frac{0,7 + 0,1 + 0,3}{3} = 0,367$$

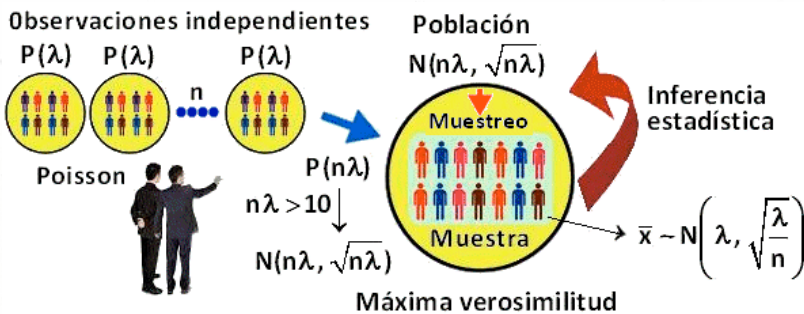
$$\hat{\theta}_2 = \frac{0,7^2 + 2 \cdot 0,1^2 + 3 \cdot 0,3^2}{6} = 0,13$$

$$\hat{\theta}_3 = \frac{0,3 - 2 \cdot 0,7 + 4 \cdot 0,1}{6} = -0,117 \mapsto \text{no puede ser, puesto que } \hat{\theta} > 0$$

c) Las funciones estimadas para las estimaciones anteriores son:

$$\hat{\theta}_1 \Rightarrow f(0,367, x) = 0,367 x^{0,367-1} = 0,367 x^{-0,633}$$

$$\hat{\theta}_2 \Rightarrow f(0,13, x) = 0,13 x^{0,13-1} = 0,13 x^{-0,87}$$



El estimador de máxima verosimilitud (probabilidad conjunta) de θ es el formado por los valores $(\hat{\theta}_1, \dots, \hat{\theta}_n)$ que maximizan la *función de verosimilitud* de la muestra $(x_1, \dots, x_n)$ obtenida:

$$L(\theta) = L(X; \theta) = L(x_1, \dots, x_n; \theta) = P(x_1, \theta) \cdots P(x_n, \theta) \quad \text{caso discreto}$$

$$L(\theta) = L(x_1, \dots, x_n; \theta) = L(X; \theta) = f_{\theta}(x_1) f_{\theta}(x_2) \cdots f_{\theta}(x_n) \quad \text{caso continuo}$$

En muchas ocasiones, es más práctico encontrar el estimador de máxima verosimilitud al considerar la función *Soporte* o *Log-verosimilitud* $\ln L(X, \theta)$, en lugar de la función de verosimilitud $L(\theta)$, ya que es más fácil de manejar y presenta los mismos máximos y mínimos.

Se despeja $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_n)$ de la ecuación: $\frac{\partial \ln L(\theta)}{\partial \theta} \bigg|_{\theta=\hat{\theta}} = 0$ y se obtiene el estimador de

máxima verosimilitud E.M.V($\hat{\theta}$)

Para que un estimador sea *eficiente* tiene que ser *centrado* y de *varianza mínima*. La varianza mínima se analiza en virtud de la acotación de Cramer-Rao:

$$V(\hat{\theta}) \geq \text{CCR} = \frac{[1 + b(\hat{\theta})]^2}{E \left[\frac{\partial \ln P(x_1, x_2, \dots, x_n; \theta)}{\partial \theta} \right]^2}$$

La Cota de Cramer-Rao no tiene por qué tomar siempre un valor muy pequeño (cercano a cero).

Un estimador es eficiente cuando $V(\hat{\theta}) = \text{CCR}$

Un estimador es asintóticamente eficiente si $\lim_{n \rightarrow \infty} V(\hat{\theta}) = \text{CCR}$

El denominador de la Cota de Cramer-Rao es la cantidad de información de Fisher $I(\theta)$ en una muestra aleatoria simple (m.a.s.) de tamaño n :

$$I(\theta) = E \left[\frac{\partial \ln L(X, \theta)}{\partial \theta} \right]^2 \quad \text{o bien} \quad i(\theta) = n E \left[\frac{\partial \ln L(x, \theta)}{\partial \theta} \right]^2 \quad \text{donde} \quad I(\theta) = n \cdot i(\theta)$$

$$V(\hat{\theta}) \geq \text{CCR} = \frac{[1 + b(\hat{\theta})]^2}{n \cdot E \left[\frac{\partial \ln P(x; \theta)}{\partial \theta} \right]^2} \xrightarrow{b(\hat{\theta})=0} V(\hat{\theta}) \geq \text{CCR} = \frac{1}{n \cdot E \left[\frac{\partial \ln P(x; \theta)}{\partial \theta} \right]^2} = \frac{1}{I(\theta)}$$

La cantidad de información de Fisher según sea el caso discreto o continuo, toma la expresión:

Discreto: $E \left[\frac{\partial \ln P(x_1, x_2, \dots, x_n; \theta)}{\partial \theta} \right]^2 = n \cdot E \left[\frac{\partial \ln P(x; \theta)}{\partial \theta} \right]^2$

Continuo: $E \left[\frac{\partial \ln f(x_1, x_2, \dots, x_n; \theta)}{\partial \theta} \right]^2 = n \cdot E \left[\frac{\partial \ln f(x; \theta)}{\partial \theta} \right]^2$

Un estimador $\hat{\theta}$ es suficiente cuando no da lugar a una pérdida de información.

En otras palabras,

Cuando la información basada en $\hat{\theta}$ es tan buena como la que hiciera uso de toda la muestra.



Para identificar estadísticos suficientes se utiliza el Teorema de Factorización, que establece que dada una muestra aleatoria $(x_1, \dots, x_n)$ de una población X con función de densidad f_{θ} (caso continuo) un estadístico $\hat{\theta}$ es suficiente para θ si y sólo si:

$$f_{\theta}(x_1, \dots, x_n) = g[\hat{\theta}(x_1, \dots, x_n), \theta] \cdot h(x_1, \dots, x_n)$$

En el caso discreto: $P_{\theta}(x_1, \dots, x_n) = g[\hat{\theta}(x_1, \dots, x_n), \theta] \cdot h(x_1, \dots, x_n)$

Para encontrar un estadístico suficiente $\hat{\theta}$ hay que factorizar la función de verosimilitud de la forma: $L(\theta) = g(\hat{\theta}, \theta) \cdot h(x_1, \dots, x_n)$

98. Un atleta olímpico de salto de altura se enfrenta a un listón de 2,3 metros. Su entrenador desea estudiar el comportamiento del saltador. Sabe que el número de saltos fallidos por hora es una variable aleatoria distribuida como una Poisson de parámetro λ . Se pide:

- Calcular el estimador máximo verosímil del parámetro λ
- Analizar las propiedades de Insensibilidad, Consistencia y Eficiencia
- A partir de la eficiencia del estimador elaborar un intervalo de confianza para el parámetro poblacional λ

Solución:

- a) Sea la v.a. X = 'Número de saltos fallidos por hora'

En la distribución de Poisson: $P(X=x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad \begin{cases} E(x) = \lambda \\ V(x) = \lambda \end{cases}$

La función de verosimilitud $L(X, \lambda)$ en una muestra aleatoria simple de tamaño n :

$$L(\lambda) = L(X, \lambda) = \prod_{i=1}^n P(x_i, \lambda) = \frac{\lambda^{x_1}}{x_1!} e^{-\lambda} \dots \frac{\lambda^{x_n}}{x_n!} e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-n\lambda}$$

$$L(\lambda) = L(X, \lambda) = \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} \cdot \underbrace{e^{-n\lambda}}_{g(\hat{\lambda}, \lambda)} \text{ es un estadístico suficiente.}$$

$$\ln L(X, \lambda) = \ln \left[\frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-n\lambda} \right] = \ln \left(\lambda^{\sum_{i=1}^n x_i} \right) - \ln \left(\prod_{i=1}^n x_i! \right) + \ln(e^{-n\lambda}) = \sum_{i=1}^n x_i \ln \lambda - \sum_{i=1}^n \ln(x_i!) - n\lambda$$

$$\ln L(X, \lambda) = \sum_{i=1}^n x_i \ln \lambda - \sum_{i=1}^n \ln(x_i!) - n\lambda$$

Para obtener el estimador de máxima verosimilitud $EMV(\hat{\lambda})$, se deriva la expresión anterior respecto del parámetro, sustituyendo λ por $\hat{\lambda}$

$$\left. \frac{\partial \ln L(X, \lambda)}{\partial \lambda} \right|_{\lambda=\hat{\lambda}} = 0 \equiv \sum_{i=1}^n x_i \frac{1}{\lambda} - n \Big|_{\lambda=\hat{\lambda}} = 0 \rightarrow \hat{\lambda} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$

El estimador de máxima verosimilitud (estimador que ofrece mayor credibilidad) viene dado por la media muestral $EMV(\hat{\lambda}) = \bar{x}$

b) Propiedades de Insesgadez, Consistencia y Eficiencia

▪ Insesgadez

El estimador $\hat{\lambda}$ es insesgado (centrado) si $E(\hat{\lambda}) = \lambda$

$$E(\hat{\lambda}) = E(\bar{x}) = E\left(\frac{\sum_{i=1}^n x_i}{n}\right) = \frac{1}{n} E\left(\sum_{i=1}^n x_i\right) = \frac{1}{n} \sum_{i=1}^n E(x_i) = \frac{1}{n} (n\lambda) = \lambda$$

$$V(\hat{\lambda}) = V(\bar{x}) = V\left(\sum_{i=1}^n x_i / n\right) = \frac{1}{n^2} \sum_{i=1}^n V(x_i) = \frac{1}{n^2} (n\lambda) = \frac{\lambda}{n}$$

▪ Consistencia

Cuando no es posible emplear estimadores de máxima verosimilitud, el requisito mínimo deseable para un estimador es que sea consistente.

El estimador $\hat{\lambda}$ es *consistente* cuando $\lim_{n \rightarrow \infty} E(\hat{\lambda}) = \lambda$ y $\lim_{n \rightarrow \infty} V(\hat{\lambda}) = 0$

$$\lim_{n \rightarrow \infty} E(\hat{\lambda}) = \lim_{n \rightarrow \infty} \lambda = \lambda \quad \text{y} \quad \lim_{n \rightarrow \infty} V(\hat{\lambda}) = \lim_{n \rightarrow \infty} \frac{\lambda}{n} = 0$$

El estimador $\hat{\lambda}$ es consistente

▪ Eficiencia

$$P(X=x) = \frac{\lambda^x}{x!} e^{-\lambda} \rightarrow \ln P(x, \lambda) = \ln \left[\frac{\lambda^x}{x!} e^{-\lambda} \right] = x \ln \lambda - \ln(x!) - \lambda$$

$$\frac{\partial \ln P(x, \lambda)}{\partial \lambda} = \frac{x}{\lambda} - 1 = \frac{x - \lambda}{\lambda}$$

$$E \left[\frac{\partial \ln P(x, \lambda)}{\partial \lambda} \right]^2 = E \left[\frac{x - \lambda}{\lambda} \right]^2 = \frac{1}{\lambda^2} E(x - \lambda)^2 = \frac{1}{\lambda^2} E(x - \bar{x})^2 = \frac{1}{\lambda^2} V(x) = \frac{\lambda}{\lambda^2} = \frac{1}{\lambda}$$

$$\text{En consecuencia, } V(\hat{\lambda}) \geq \frac{1}{n \frac{1}{\lambda}} = \frac{\lambda}{n}$$

El menor valor de la varianza del estimador será $\frac{\lambda}{n}$

Se sabe que $V(\hat{\lambda}) = V(\bar{x}) = \frac{\lambda}{n}$, lo que muestra $V(\hat{\lambda}) = \text{CCR}$, el estimador empleado es *eficiente*.

$$\text{c) } I = [(\text{media muestral}) \pm (\text{error estimación})] = \left[\hat{\lambda} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{\lambda}}{n}} \right] \equiv \left[\bar{x} \pm z_{\alpha/2} \cdot \sqrt{\frac{\bar{x}}{n}} \right]$$

99. Una muestra aleatoria $(x_1, \dots, x_n)$ de la población tiene como función de densidad

$$f_{\theta}(x) = \begin{cases} \theta x^{\theta-1} & \text{si } x \in (0,1) \\ 0 & \text{en el resto} \end{cases} \quad \theta > 0$$

- a) Hallar un estadístico suficiente
b) Estimador de máxima verosimilitud de θ

Solución:

a) Función de verosimilitud:

$$L(\theta) = f_{\theta}(x_1) f_{\theta}(x_2) \cdots f_{\theta}(x_n) = (\theta x_1^{\theta-1}) (\theta x_2^{\theta-1}) \cdots (\theta x_n^{\theta-1}) = \underbrace{\theta^n}_{g(\hat{\theta}, \theta)} \cdot \underbrace{(x_1 \cdots x_n)^{\theta-1}}_{h(x_1, x_2, \dots, x_n)}$$

Por tanto, $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_n)$ es un estadístico suficiente.

$$b) L(\theta) = \theta^n (x_1, \dots, x_n)^{\theta-1} = \theta^n \prod_{i=1}^n x_i^{\theta-1}$$

Función soporte o Log-verosimilitud:

$$\begin{aligned} \ln L(\theta) &= \ln \left[\theta^n \prod_{i=1}^n x_i^{\theta-1} \right] = \ln \theta^n + \ln \prod_{i=1}^n x_i^{\theta-1} = \ln \theta^n + \sum_{i=1}^n \ln x_i^{\theta-1} = \\ &= n \cdot \ln \theta + (\theta-1) \cdot \ln \sum_{i=1}^n \ln x_i \end{aligned}$$

$$\frac{\partial \ln L(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} = \frac{n}{\hat{\theta}} + \sum_{i=1}^n \ln x_i = 0 \rightarrow \hat{\theta} = - \frac{n}{\sum_{i=1}^n \ln x_i}$$

100. En una distribución $N(\mu, \sigma)$ se estima la media poblacional μ en una muestra aleatoria simple de tamaño n $(x_1, x_2, \dots, x_n)$ por medio de la función muestral:

$$\hat{\mu} = \frac{\sum_{i=1}^n i x_i}{n}$$

Estúdiese la eficiencia del estimador.

Solución:

Un estimador es eficiente cuando $V(\hat{\theta}) = \text{CCR}$

Para hallar la cota de Cramer-Rao se necesita saber en primer lugar si el estimador es insesgado o no, se calcula para ello la esperanza matemática:

$$\begin{aligned} E(\hat{\mu}) &= E\left[\frac{\sum_{i=1}^n i x_i}{n}\right] = \frac{1}{n} \sum_{i=1}^n E(i x_i) = \frac{1}{n} \sum_{i=1}^n i E(x_i) = \frac{1}{n} [E(x_1) + 2E(x_2) + 3E(x_3) + \dots + nE(x_n)] = \\ &= \frac{1}{n} [\mu + 2\mu + 3\mu + \dots + n\mu] = \frac{\mu}{n} [1 + 2 + 3 + \dots + n] = \frac{\mu}{n} \frac{(n+1)n}{2} = \frac{(n+1)}{2} \mu \end{aligned}$$

$$E(\hat{\mu}) = \mu + b(\hat{\mu}) \rightarrow \text{Sesgo: } b(\hat{\mu}) = E(\hat{\mu}) - \mu = \frac{n+1}{2} \mu - \mu = \frac{n-1}{2} \mu$$

Varianza del estimador:

$$\begin{aligned} V(\hat{\mu}) &= V\left[\frac{\sum_{i=1}^n i x_i}{n}\right] = \frac{1}{n^2} \sum_{i=1}^n V(i x_i) = \frac{1}{n^2} \sum_{i=1}^n i^2 V(x_i) = \frac{1}{n^2} [1^2 V(x_1) + 2^2 V(x_2) + 3^2 V(x_3) + \dots + n^2 V(x_n)] = \\ &= \frac{1}{n^2} [1^2 \sigma^2 + 2^2 \sigma^2 + 3^2 \sigma^2 + \dots + n^2 \sigma^2] = \frac{\sigma^2}{n^2} [1^2 + 2^2 + 3^2 + \dots + n^2] = \frac{\sigma^2}{n^2} \frac{n(n+1)(2n+1)}{6} \end{aligned}$$

$$V(\hat{\mu}) = \frac{(n+1)(2n+1)\sigma^2}{6n}$$

$$\text{Cota de Cramer-Rao: } \text{CCR} = \frac{[1 + b(\hat{\mu})]^2}{n E \left[\frac{\partial \ln L(x, \mu)}{\partial \mu} \right]^2}$$

$$[1 + b(\hat{\mu})]^2 = \left[1 + \frac{n-1}{2} \mu \right]^2 = \frac{(n+1)^2}{4}$$

La Información de Fisher $I(\mu) = ni(\mu)$ de la muestra: $I(\mu) = E \left[\frac{\partial \ln L(X, \mu)}{\partial \mu} \right]^2 = n E \left[\frac{\partial \ln L(x, \mu)}{\partial \mu} \right]^2$

se obtiene a partir de la función de densidad poblacional $f(x, \mu) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Tomando logaritmos neperianos, se tiene:

$$\ln f(x, \mu) = \ln \left(\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right) = -\ln \sigma - \frac{1}{2} \ln 2\pi - \frac{(x-\mu)^2}{2\sigma^2}$$

Derivando respecto a μ :

$$\frac{\partial \ln f(x, \mu)}{\partial \mu} = -\frac{1}{2\sigma^2} [-2(x-\mu)] = \frac{x-\mu}{\sigma^2} \rightarrow \left(\frac{\partial \ln f(x, \mu)}{\partial \mu} \right)^2 = \frac{(x-\mu)^2}{\sigma^4}$$

La esperanza matemática: $E \left[\frac{\partial \ln f(x, \mu)}{\partial \mu} \right]^2 = E \left[\frac{(x-\mu)^2}{\sigma^4} \right] = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}$

Atendiendo a la aditividad de esta medida: $I(\mu) = ni(\mu) = \frac{n}{\sigma^2}$

$$CCR = \frac{[1 + b(\hat{\mu})]^2}{n E \left[\frac{\partial \ln L(x, \mu)}{\partial \mu} \right]^2} = \frac{\frac{(n+1)^2}{4}}{\frac{n}{\sigma^2}} = \frac{(n+1)^2 \sigma^2}{4n} \rightarrow V(\hat{\mu}) = \frac{(n+1)(2n+1) \sigma^2}{6n}$$

Al ser la varianza del estimador mayor que la cota de Cramer-Rao:

$$V(\hat{\mu}) = \frac{(n+1)(2n+1) \sigma^2}{6n} > \frac{(n+1)^2 \sigma^2}{4n} = CCR$$

El estimador *no es eficiente*, siempre que $n > 1$

Para $n = 1$: $V(\hat{\mu}) = CCR = \frac{\sigma^2}{n}$ el estimador es eficiente.

101. Una muestra aleatoria $(x_1, \dots, x_n)$ de la población tiene como función de densidad

$$f_{\theta}(x) = \frac{x}{\theta^2} e^{-x/\theta} \quad \text{siendo } x > 0$$

- a) Hallar un estadístico suficiente
b) Estimador de máxima verosimilitud de θ

Solución:

a) Función de verosimilitud:

$$L(\theta) = f_{\theta}(x_1) f_{\theta}(x_2) \cdots f_{\theta}(x_n) = \left(\frac{x_1}{\theta^2} e^{-\frac{x_1}{\theta}} \right) \left(\frac{x_2}{\theta^2} e^{-\frac{x_2}{\theta}} \right) \cdots \left(\frac{x_n}{\theta^2} e^{-\frac{x_n}{\theta}} \right) = \left(\frac{1}{\theta^2} \right)^n e^{-\frac{1}{\theta} \sum_{i=1}^n x_i}$$

Por tanto, $\hat{\theta} = (\hat{\theta}_1, \dots, \hat{\theta}_n)$ es un estadístico suficiente.

$$b) \quad L(\theta) = \left(\frac{1}{\theta^2} \right)^n e^{-\frac{1}{\theta} \sum_{i=1}^n x_i}$$

Función soporte o Log-verosimilitud:

$$\begin{aligned} \ln L(\theta) &= \ln \left[\left(\frac{1}{\theta^2} \right)^n e^{-\frac{1}{\theta} \sum_{i=1}^n x_i} \right] = \ln \left(\frac{1}{\theta^2} \right)^n + \ln e^{-\frac{1}{\theta} \sum_{i=1}^n x_i} = -n \cdot \ln \left(\frac{1}{\theta^2} \right) - \frac{1}{\theta} \sum_{i=1}^n x_i = \\ &= -2n \cdot \ln \theta - \frac{1}{\theta} \sum_{i=1}^n x_i \end{aligned}$$

$$\frac{\partial \ln L(\theta)}{\partial \theta} \Big|_{\theta=\hat{\theta}} = \frac{-2n}{\hat{\theta}} + \frac{1}{\hat{\theta}^2} \sum_{i=1}^n x_i = 0 \rightarrow \hat{\theta} = \frac{\sum_{i=1}^n x_i}{2n} = \frac{1}{2} \bar{x}$$

102. En un estacionamiento el número de veces que se abre la barrera en un intervalo de 10 minutos, para que pasen vehículos en un sector de seguridad, se considera una variable aleatoria con distribución de Poisson de parámetro λ desconocido.

a) En una muestra aleatoria de 8 intervalos de 10 minutos cada uno, elegidos de forma independiente, se registra para cada intervalo el valor que toma la variable en estudio.

3 5 8 7 4 5 6 2

Encontrar la estimación máximo verosímil de λ

b) Sea $(x_1, \dots, x_n)$ una muestra aleatoria de tamaño n que sigue una distribución de Poisson.

Sí $\hat{\lambda}_1 = \sum_{i=1}^n \frac{x_i}{n}$, $\hat{\lambda}_2 = \frac{x_1 + 3x_n}{4}$ son estimadores. Determinar el mejor estimador del parámetro λ

Solución:

a) $X =$ "Número de veces que se abre la barrera en un intervalo de 10 minutos", $X \sim P(\lambda)$

$$P(X, \lambda) = \begin{cases} \frac{\lambda^x}{x!} e^{-\lambda} & x = 0, 1, 2, \dots \\ 0 & \text{otro caso} \end{cases}$$

$$\text{Función de verosimilitud: } L(\lambda) = L(X, \lambda) = \prod_{i=1}^n P(x_i, \lambda) = \frac{\lambda^{x_1}}{x_1!} e^{-\lambda} \dots \frac{\lambda^{x_n}}{x_n!} e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-n\lambda}$$

$$\ln L(X, \lambda) = \ln \left(\frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!} e^{-n\lambda} \right) = \ln(\lambda^{\sum_{i=1}^n x_i}) - \ln \left(\prod_{i=1}^n x_i! \right) + \ln(e^{-n\lambda}) = \sum_{i=1}^n x_i \ln \lambda - \sum_{i=1}^n \ln(x_i!) - n\lambda$$

Para obtener el estimador de máxima verosimilitud $EMV(\hat{\lambda})$, se deriva la expresión anterior respecto del parámetro para obtener, sustituyendo λ por $\hat{\lambda}$

$$\left. \frac{\partial \ln L(X, \lambda)}{\partial \lambda} \right|_{\lambda=\hat{\lambda}} = 0 = \sum_{i=1}^n x_i \frac{1}{\lambda} - n \left|_{\lambda=\hat{\lambda}} = 0 \Rightarrow \hat{\lambda} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$

El estimador de máxima verosimilitud viene dado por la media muestral $EMV(\hat{\lambda}) = \bar{x}$

Utilizando la muestra aleatoria de ocho intervalos de 10 minutos, se obtiene el estimador máximo verosímil:

$$\hat{\lambda} = \frac{3+5+8+7+4+5+6+2}{8} = 5$$

En consecuencia, en una muestra aleatoria de ocho intervalos de 10 minutos cada uno, elegidos de forma independiente, la estimación máxima verosímil corresponde a que la barrera se abre 5 veces.

b) Se analizan si los estimadores son o no insesgados, esto es, si la esperanza del estimador coincide con el parámetro a estimar.

$$E(\hat{\lambda}_1) = E\left(\sum_{i=1}^n \frac{x_i}{n}\right) = \frac{1}{n} \sum_{i=1}^n E(x_i) = \frac{n\lambda}{n} = \lambda$$

$$E(\hat{\lambda}_2) = E\left(\frac{x_1 + 3x_n}{4}\right) = \frac{1}{4} [E(x_1) + 3E(x_n)] = \frac{\lambda + 3\lambda}{4} = \lambda$$

Ambos estimadores son insesgados.

Varianza de cada estimador:

$$V(\hat{\lambda}_1) = V\left(\sum_{i=1}^n \frac{x_i}{n}\right) = \frac{1}{n^2} \sum_{i=1}^n V(x_i) = \frac{n\lambda}{n^2} = \frac{\lambda}{n}$$

$$V(\hat{\lambda}_2) = V\left(\frac{x_1 + 3x_n}{4}\right) = \frac{1}{16} [V(x_1) + 9V(x_n)] = \frac{\lambda + 9\lambda}{16} = \frac{10\lambda}{16} = \frac{5\lambda}{8}$$

La efectividad de los estimadores depende del tamaño de la muestra:

- Si la muestra es igual a 1 ($n = 1$) el estimador más eficiente es $\hat{\lambda}_2$
- Si la muestra es mayor que 1 ($n > 1$) el estimador más eficiente es $\hat{\lambda}_1$

103. Sea $(x_1, x_2, \dots, x_n)$ una muestra aleatoria de tamaño n , distribuida según $f(x, \theta)$ con θ desconocido, donde X representa el tiempo máximo necesario para determinar un proceso en

segundos:
$$f(X, \theta) = \begin{cases} (\theta + 1)x^\theta & 0 \leq x \leq 1, \theta > -1 \\ 0 & \text{otro caso} \end{cases}$$

- a) Determinar el estimador máximo verosímil de θ
b) Determinar la estimación máximo verosímil de θ en una muestra aleatoria simple constituida por los datos: 0,7 0,9 0,6 0,8 0,9 0,7 0,9.

Estimar la probabilidad del tiempo máximo necesario para terminar un proceso, que no exceda de 0,25 segundos ni supere los 0,75 segundos.

- c) Determinar el estimador máximo verosímil de $\theta + 1$ y $\frac{2\theta + 1}{\theta - 1}$

Solución:

a) La función de verosimilitud $L(\theta) = L(X, \theta) = \prod_{i=1}^n f(x_i, \theta)$

$$L(x_1, x_2, \dots, x_n; \theta) = [(\theta + 1)x_1^\theta] [(\theta + 1)x_2^\theta] \dots [(\theta + 1)x_n^\theta] = (\theta + 1)^n \prod_{i=1}^n x_i^\theta$$

Calculando el logaritmo natural de $L(x_1, x_2, \dots, x_n; \theta)$:

$$\ln[L(x_1, x_2, \dots, x_n; \theta)] = \ln \left((\theta + 1)^n \prod_{i=1}^n x_i^\theta \right) = \ln (\theta + 1)^n + \ln \left(\prod_{i=1}^n x_i^\theta \right) = n \ln (\theta + 1) + \theta \sum_{i=1}^n \ln x_i$$

Para obtener el estimador de máxima verosimilitud $EMV(\hat{\theta})$, se deriva la expresión anterior respecto del parámetro para obtener, sustituyendo θ por $\hat{\theta}$

$$\frac{\partial \ln[L(x_1, x_2, \dots, x_n; \theta)]}{\partial \theta} \bigg|_{\theta = \hat{\theta}} = \frac{n}{\hat{\theta} + 1} + \sum_{i=1}^n \ln x_i = 0 \rightarrow \frac{n}{\hat{\theta} + 1} = - \sum_{i=1}^n \ln x_i$$

$$\hat{\theta} = EMV(\theta) = - \frac{n}{\sum_{i=1}^n \ln x_i} - 1$$

- b) El estimador máximo verosímil de θ con los datos de la muestra:

$$\hat{\theta} = - \frac{8}{\ln 0,7 + \ln 0,9 + \ln 0,6 + \ln 0,8 + \ln 0,9 + \ln 0,7 + \ln 0,9 + \ln 0,8} - 1 = 4,027 - 1 = 3,027$$

El estimador máximo verosímil $EMV(\theta) = \hat{\theta} = 3,027$

De otra parte,

$$P(0,25 \leq X \leq 0,75) = \int_{0,25}^{0,75} f(X, \hat{\theta}) dx = \int_{0,25}^{0,75} (\hat{\theta} + 1) x^{\hat{\theta}} dx = (\hat{\theta} + 1) \frac{x^{\hat{\theta}+1}}{(\hat{\theta} + 1)} \Big|_{0,25}^{0,75} = x^{\hat{\theta}+1} \Big|_{0,25}^{0,75} = x^{3,027+1} \Big|_{0,25}^{0,75} =$$

$$= 0,75^{4,027} - 0,25^{4,047} = 0,31$$

La probabilidad del tiempo máximo necesario para terminar el proceso (entre 0,25 y 0,75 segundos) es 0,31

c) El estimador máximo verosímil de $\theta + 1$ y $\frac{2\theta + 1}{\theta - 1}$

$$EMV[\theta + 1] = EMV[\theta] + 1 = \hat{\theta} + 1 = 3,027 + 1 = 4,027$$

$$EMV\left(\frac{2\theta + 1}{\theta - 1}\right) = \frac{2 EMV(\theta) + 1}{EMV(\theta) - 1} = \frac{2\hat{\theta} + 1}{\hat{\theta} - 1} = \frac{2 \times 3,027 + 1}{3,027 - 1} = 2,493$$

104. En una distribución $N(\mu, \sigma)$, se estima la media poblacional μ mediante la media de una muestra aleatoria simple $(x_1, x_2, \dots, x_n)$. El estimador es insesgado y su varianza $V(\bar{x}) = \frac{\sigma^2}{n}$. Demostrar que la media muestral es un estimador eficiente.

Solución:

La varianza de un estimador verifica siempre la Cota de Cramer-Rao: $V(\hat{\theta}) \geq CCR$.

Un estimador es eficiente cuando $V(\hat{\theta}) = CCR$

Para obtener la cota de Cramer-Rao se parte de la función de densidad poblacional:

$$CCR = \frac{1}{n E \left[\frac{\partial \ln f(x, \mu)}{\partial \mu} \right]^2}$$

La función de densidad poblacional: $f(x, \mu) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

tomando logaritmos neperianos, se tiene:

$$\ln f(x, \mu) = \ln \left(\frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right) = -\ln \sigma - \frac{1}{2} \ln 2\pi - \frac{(x-\mu)^2}{2\sigma^2}$$

Derivando respecto a μ :

$$\frac{\partial \ln f(x, \mu)}{\partial \mu} = -\frac{1}{2\sigma^2} [-2(x - \mu)] = \frac{x - \mu}{\sigma^2} \rightarrow \left(\frac{\partial \ln f(x, \mu)}{\partial \mu} \right)^2 = \frac{(x - \mu)^2}{\sigma^4}$$

La esperanza matemática: $E \left[\frac{\partial \ln f(x, \mu)}{\partial \mu} \right]^2 = E \left[\frac{(x - \mu)^2}{\sigma^4} \right] = \frac{\sigma^2}{\sigma^4} = \frac{1}{\sigma^2}$

Sustituyendo el valor de la esperanza matemática en la expresión de la cota para estimadores insesgados:

$$CCR = \frac{1}{n E \left[\frac{\partial \ln f(x, \mu)}{\partial \mu} \right]^2} = \frac{1}{n \frac{1}{\sigma^2}} = \frac{\sigma^2}{n}$$

La varianza del estimador coincide con la cota de Cramer-Rao, concluyendo que la media muestral es un estimador eficiente de la media poblacional en la distribución normal.

105. Sea la distribución $N(\mu, \sigma)$, con media μ conocida y varianza desconocida.

Calcular la estimación máximo-verosímil de la varianza en muestras aleatorias simples de tamaño n .

Solución:

La función de verosimilitud es:

$$L(X; \mu, \sigma^2) = \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_1 - \mu)^2}{2\sigma^2}} \right] \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_2 - \mu)^2}{2\sigma^2}} \right] \dots \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_n - \mu)^2}{2\sigma^2}} \right] = \frac{1}{(2\pi)^{\frac{n}{2}} (\sigma^2)^{\frac{n}{2}}} e^{-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}}$$

tomando logaritmos neperianos, se tiene:

$$\ln [L(X; \mu, \sigma^2)] = \ln \left[\frac{1}{(2\pi)^{\frac{n}{2}} (\sigma^2)^{\frac{n}{2}}} e^{-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}} \right] = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}$$

y derivando respecto a σ^2 e igualando a cero:

$$\left. \frac{d \ln [L(X; \mu, \sigma^2)]}{d \sigma^2} \right|_{\sigma = \hat{\sigma}} = -\frac{n}{2\hat{\sigma}^2} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\hat{\sigma}^4} = 0 \rightarrow \hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \mu)^2}{n}$$

Conviene observar que el estimador no es la varianza muestral, dado que las desviaciones de los valores muestrales lo son con respecto a la media poblacional μ y no respecto a la media muestral $\bar{x}$.

106. Calcular el estimador máximo verosímil del parámetro 'a' de las siguientes funciones:

a) $f(x; a) = a^2 e^{-ax}$ siendo $x \geq 0$ en muestras aleatorias simples de tamaño n.

b) $f(x; a) = ae^{-ax}$ para $x \geq 0$, $a > 0$ en muestras aleatorias simples de tamaño 2.

Solución::

a) $f(x; a) = a^2 e^{-ax}$ donde $x \geq 0$ en m.a.s. de tamaño n

Función de verosimilitud $L = L(x_1, x_2, \dots, x_n; a) = (a^2 e^{-ax_1}) \cdot (a^2 e^{-ax_2}) \dots (a^2 e^{-ax_n}) = a^{2n} e^{-a \sum_{i=1}^n x_i}$

Tomando logaritmos neperianos: $\ln L = \ln \left(a^{2n} e^{-a \sum_{i=1}^n x_i} \right) = 2n \ln a - a \sum_{i=1}^n x_i$

Derivando respecto del parámetro 'a' e igualando a cero:

$$\left. \frac{d(\ln L)}{da} \right|_{a=\hat{a}} = \frac{2n}{\hat{a}} - \sum_{i=1}^n x_i = 0 \rightarrow \hat{a} = \frac{2n}{\sum_{i=1}^n x_i} = \frac{2}{\bar{x}} \rightarrow \hat{a} = \frac{2}{\bar{x}}$$

b) Sea $f(x; a) = ae^{-ax}$ para $x \geq 0$, $a > 0$ en m.a.s. de tamaño 2

Función de verosimilitud $L = L(x_1, x_2; a) = (ae^{-ax_1}) \cdot (ae^{-ax_2}) = a^2 e^{-a(x_1+x_2)}$

Tomando logaritmos neperianos: $\ln L = \ln (a^2 e^{-a(x_1+x_2)}) = 2 \ln a - a(x_1+x_2)$

Derivando respecto de 'a' e igualando a cero:

$$\left. \frac{d(\ln L)}{da} \right|_{a=\hat{a}} = \frac{2}{\hat{a}} - (x_1+x_2) = 0 \rightarrow \hat{a} = \frac{2}{x_1+x_2} = \frac{1}{\bar{x}}$$

107. En la distribución $B(m, p)$ se considera como estimador del parámetro p el estadístico

$\hat{p} = \frac{\bar{x}}{n}$, siendo $\bar{x}$ la media muestral en muestras aleatorias simples de tamaño n . Hallar la eficiencia del estimador.

Solución:

Un estimador es eficiente cuando su varianza coincide con la Cota de Cramer-Rao $V(\hat{p}) = CCR$

$$CCR = \frac{1}{n \cdot E \left[\frac{\partial \ln P(x; p)}{\partial \theta} \right]^2} \quad I(\theta) = n \cdot E \left[\frac{\partial \ln P(x; p)}{\partial \theta} \right]^2 \quad \text{Información de Fisher}$$

$$\text{Inssegadez: } E(\hat{p}) = E\left(\frac{\bar{x}}{n}\right) = \frac{np}{n} = p$$

El estimador es inssegado o centrado

Para determinar la CCR se considera la función de cuantía de la distribución binomial:

$$P(x, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\text{Función soporte: } \ln P(x, p) = \ln \binom{n}{x} + x \ln p + (n-x) \ln(1-p)$$

$$\frac{\partial \ln P(x, p)}{\partial p} = \frac{x}{p} - \frac{(n-x)}{1-p} = \frac{(1-p)x - p(n-x)}{p(1-p)} = \frac{x - np}{p(1-p)}$$

$$E \left[\frac{\partial \ln P(x, p)}{\partial p} \right]^2 = E \left[\frac{x - np}{p(1-p)} \right]^2 = \frac{1}{p^2(1-p)^2} E(x - np)^2 = \frac{1}{p^2(1-p)^2} np(1-p) = \frac{n}{p(1-p)}$$

En una distribución binomial: $E(X) = np$, $V(X) = E(X - np)^2 = np(1-p)$

$$CCR = \frac{1}{n \cdot E \left[\frac{\partial \ln P(x; p)}{\partial \theta} \right]^2} = \frac{1}{n \cdot \frac{n}{p(1-p)}} = \frac{p(1-p)}{n^2}$$

$$\text{La varianza del estimador: } V(\hat{p}) = V\left(\frac{\bar{x}}{n}\right) = \frac{1}{n^2} V(\bar{x}) = \frac{1}{n^2} \frac{np(1-p)}{n} = \frac{p(1-p)}{n^2}$$

Como la varianza del estimador coincide con la cota de Cramer-Rao el estadístico $\hat{p} = \frac{\bar{x}}{n}$ es eficiente.

108. Una urna contiene bolas blancas y negras. Sea p la probabilidad de extraer una bola blanca cuando se realiza una extracción al azar. Asociado a este experimento aleatorio se encuentra la variable aleatoria X que puede tomar los valores:

$X = 1$ si la bola extraída es blanca $X = 0$ si la bola extraída es negra

Se selecciona una muestra aleatoria con reemplazamiento de tamaño 3 (x_1, x_2, x_3) , siendo X_i la variable aleatoria a la extracción i -ésima, y se supone que ha resultado la siguiente relación (B, N, B) . Como el parámetro p es desconocido se pretende saber entre los valores $p = 0,65$ y $p = 0,73$ qué valor hace más probable la aparición de dicha extracción.

Solución:

La distribución de probabilidad será una $B(1, p)$: $P(X = x) = p^x (1-p)^{1-x}$

Si la muestra (B, N, B) es independiente, siendo $P(B) = p$ $P(N) = 1 - p$

$$P(B, N, B) = P(B \cap N \cap B) = P(B) \cdot P(N) \cdot P(B) = p \cdot (1-p) \cdot p = p^2 \cdot (1-p)$$

$$\text{entonces } \begin{cases} p = 0,65: P(B, N, B) = 0,65^2 \cdot 0,35 = 0,1479 \\ p = 0,73: P(B, N, B) = 0,73^2 \cdot 0,27 = 0,1439 \end{cases}$$

Resulta más probable ($p = 0,65$), siendo más verosímil.

Se considera la m.a.s. (x_1, x_2, x_3) , siendo las variables aleatorias X_i independientes, tomando los valores 0, 1, con distribución $B(1, p)$, la distribución de probabilidad asociada será:

$$\left. \begin{aligned} P(x_1, p) &= P(X = x_1) = p^{x_1} (1-p)^{1-x_1} \\ P(x_2, p) &= P(X = x_2) = p^{x_2} (1-p)^{1-x_2} \\ P(x_3, p) &= P(X = x_3) = p^{x_3} (1-p)^{1-x_3} \end{aligned} \right\} x_i = 1, 0 \text{ sea bola blanca o negra}$$

Función de verosimilitud

$$L(p) = \prod_{i=1}^3 P(x_i, p) = p^{x_1} (1-p)^{1-x_1} \cdot p^{x_2} (1-p)^{1-x_2} \cdot p^{x_3} (1-p)^{1-x_3} = p^{x_1+x_2+x_3} (1-p)^{3-(x_1+x_2+x_3)}$$

En la muestra (B, N, B) el valor que toma la función de verosimilitud será:

$$L(p) = p^{1+0+1} (1-p)^{3-(1+0+1)} = p^2 \cdot (1-p)$$

109. A partir de una muestra aleatoria simple X de tamaño n , determinar la matriz de información de Fisher de una población con distribución $N(\mu, \sigma)$, respecto a los dos parámetros μ y σ .

Como consecuencia, determinar la distribución asintótica del Valor en Riesgo (Value at Risk) de una empresa que cotiza en Bolsa y considera como pérdidas los rendimientos inferiores a 3 euros por acción.

Solución:

Función de densidad poblacional: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Función de verosimilitud:

$$L(\mu, \sigma) = L(X; \mu, \sigma) = \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right] \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right] \cdots \left[\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right] =$$

$$= \frac{1}{(2\pi)^{\frac{n}{2}} \sigma^{\frac{n}{2}}} e^{-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}}$$

Función soporte o Log-verosimilitud:

$$\ln L(\mu, \sigma) = \ln \left[\frac{1}{(2\pi)^{\frac{n}{2}} \sigma^{\frac{n}{2}}} e^{-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}} \right] = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

Para determinar la información de Fisher se deriva la expresión anterior respecto de los parámetros μ y σ

$$\left. \frac{\partial \ln L(\mu, \sigma)}{\partial \mu} \right|_{\mu=\hat{\mu}} = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \hat{\mu}) = 0 \rightarrow E(\hat{\mu}) = \bar{x}$$

$$\left. \frac{\partial \ln L(\mu, \sigma)}{\partial \sigma} \right|_{\sigma=\hat{\sigma}} = -\frac{n}{\hat{\sigma}} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{\hat{\sigma}^3} = 0 \rightarrow E(\hat{\sigma}) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\text{Cálculo de las segundas derivadas} \left[\mu = \bar{x}, \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right]$$

$$\left. \frac{\partial^2 \ln L(X; \mu, \sigma)}{\partial \mu \partial \mu} \right|_{(\bar{x}, \hat{\sigma}^2)} = \frac{\partial}{\partial \mu} \left(\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) \right) \bigg|_{(\bar{x}, \hat{\sigma}^2)} = -\frac{1}{\sigma^2} \bigg|_{(\bar{x}, \hat{\sigma}^2)} = -\frac{n}{\hat{\sigma}^2}$$

$$\left. \frac{\partial^2 \ln L(X; \mu, \sigma)}{\partial \mu \partial \sigma} \right|_{(\bar{x}, \hat{\sigma}^2)} = \frac{\partial^2 \ln L(X; \mu, \sigma)}{\partial \sigma \partial \mu} \bigg|_{(\bar{x}, \hat{\sigma}^2)} = \frac{\partial}{\partial \sigma} \left(\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \bar{x}) \right) \bigg|_{(\bar{x}, \hat{\sigma}^2)} = -\frac{2 \sum_{i=1}^n (x_i - \bar{x})}{\hat{\sigma}^3} = 0$$

$$\begin{aligned} \left. \frac{\partial^2 \ln L(X; \mu, \sigma)}{\partial \sigma \partial \sigma} \right|_{(\bar{x}, \hat{\sigma}^2)} &= \frac{\partial}{\partial \sigma} \left(-\frac{n}{\sigma} + \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sigma^3} \right) \bigg|_{(\bar{x}, \hat{\sigma}^2)} = \frac{n}{\sigma^2} + \sum_{i=1}^n (x_i - \bar{x})^2 \left(\frac{-3\sigma^2}{\sigma^6} \right) \bigg|_{(\bar{x}, \hat{\sigma}^2)} = \\ &= \frac{n}{\hat{\sigma}^2} - \frac{3 \sum_{i=1}^n (x_i - \bar{x})^2}{\hat{\sigma}^4} = \frac{n}{\hat{\sigma}^2} - \frac{3n\hat{\sigma}^2}{\hat{\sigma}^4} = -\frac{2n}{\hat{\sigma}^2} \end{aligned}$$

Matriz de Información de Fisher $I(\mu, \sigma)$:

$$I(\mu, \sigma) = -E \begin{bmatrix} \frac{\partial^2}{\partial \mu \partial \mu} \ln L(\mu, \sigma) & \frac{\partial^2}{\partial \mu \partial \sigma} \ln L(\mu, \sigma) \\ \frac{\partial^2}{\partial \sigma \partial \mu} \ln L(\mu, \sigma) & \frac{\partial^2}{\partial \sigma \partial \sigma} \ln L(\mu, \sigma) \end{bmatrix} = - \begin{bmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{2n}{\hat{\sigma}^2} \end{bmatrix} = \begin{bmatrix} \frac{n}{\hat{\sigma}^2} & 0 \\ 0 & \frac{2n}{\hat{\sigma}^2} \end{bmatrix}$$

Se puede obtener la distribución de $(\hat{\mu}, \hat{\sigma})$: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Función soporte:

$$\begin{aligned} \ln f(x) &= \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right) = \ln \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) + \ln \left(e^{-\frac{(x-\mu)^2}{2\sigma^2}} \right) = -\ln \sqrt{2\pi\sigma^2} - \left[\frac{(x-\mu)^2}{2\sigma^2} \right] = \\ &= -\ln(\sigma\sqrt{2\pi}) - \frac{1}{2} \left[\frac{(x-\mu)^2}{\sigma^2} \right] = -\ln \sqrt{2\pi} - \ln \sigma - \frac{1}{2} \left[\frac{(x-\mu)^2}{\sigma^2} \right] \end{aligned}$$

Para determinar la información de Fisher se deriva la expresión anterior respecto de los parámetros μ y σ

$$\ln f(x) = -\ln \sqrt{2\pi} - \ln \sigma - \frac{1}{2} \left[\frac{(x-\mu)^2}{\sigma^2} \right]$$

$$\frac{\partial}{\partial \mu} \ln f(x) = \frac{(x-\mu)}{\sigma^2}$$

$$\frac{\partial}{\partial \sigma} \ln f(x) = \frac{-1}{\sigma} - \frac{(x-\mu)^2}{2} \left(\frac{-2\sigma}{\sigma^4} \right) = \frac{-1}{\sigma} + \frac{(x-\mu)^2}{\sigma^3}$$

Cálculo de las segundas derivadas:

$$\frac{\partial^2}{\partial \mu \partial \mu} f(x) = -\frac{1}{\sigma^2}$$

$$\frac{\partial^2}{\partial \mu \partial \sigma} f(x) = \frac{\partial^2}{\partial \sigma \partial \mu} f(x) = (x-\mu) \left(\frac{-2\sigma}{\sigma^4} \right) = -\frac{2(x-\mu)}{\sigma^3}$$

$$\frac{\partial^2}{\partial \sigma \partial \sigma} f(x) = \frac{1}{\sigma^2} + (x-\mu)^2 \left(\frac{-3\sigma^2}{\sigma^6} \right) = \frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4}$$

Matriz de Información de Fisher $I(\mu, \sigma)$:

$$\begin{aligned} i(\mu, \sigma) &= -E \begin{bmatrix} \frac{\partial^2}{\partial \mu \partial \mu} f(x) & \frac{\partial^2}{\partial \mu \partial \sigma} f(x) \\ \frac{\partial^2}{\partial \sigma \partial \mu} f(x) & \frac{\partial^2}{\partial \sigma \partial \sigma} f(x) \end{bmatrix} = -E \begin{bmatrix} -\frac{1}{\sigma^2} & -\frac{2(x-\mu)}{\sigma^3} \\ -\frac{2(x-\mu)}{\sigma^3} & \frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4} \end{bmatrix} = \\ &= E \begin{bmatrix} \frac{1}{\sigma^2} & \frac{2(x-\mu)}{\sigma^3} \\ \frac{2(x-\mu)}{\sigma^3} & -\frac{1}{\sigma^2} + \frac{3(x-\mu)^2}{\sigma^4} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & -\frac{1}{\sigma^2} + \frac{3\sigma^2}{\sigma^4} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix} \end{aligned}$$

Atendiendo a la aditividad de esta medida:

$$I(\mu, \sigma) = n i(\mu, \sigma) = n \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{2}{\sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{2n}{\sigma^2} \end{bmatrix}$$

Por el Teorema Central del Límite la distribución asintótica de $(\hat{\mu}, \hat{\sigma})$ es:

$$\sqrt{n} \left(\begin{bmatrix} \hat{\mu} \\ \hat{\sigma} \end{bmatrix} - \begin{bmatrix} \mu \\ \sigma \end{bmatrix} \right) \sim N(0, i(\mu, \sigma)^{-1}) = N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 \\ 0 & \frac{\sigma^2}{2} \end{bmatrix} \right)$$

Consecuencia: En una cartera de inversiones, asumiendo mercados normales y que no se produce negociación en la cartera, el VaR (Valor en Riesgo) se define como un valor límite tal que la probabilidad de que una pérdida a precios de mercados en la cartera sobre un horizonte temporal dado exceda ese valor sea el nivel de probabilidad dado.

Parámetros comunes para el VaR son probabilidades de 1% y 5% y horizontes temporales de un día y de dos semanas, aunque también se utilizan otras combinaciones.

Las pérdidas máximas esperadas para un nivel de significación α son $\text{VaR} = \mu - 3 - z_\alpha \sigma$.

La distribución del estimador: $\widehat{\text{VaR}} = \hat{\mu} - 3 - z_\alpha \hat{\sigma} = (1 \quad -z_\alpha) \begin{pmatrix} \hat{\mu} \\ \hat{\sigma} \end{pmatrix} - 3$

La distribución asintótica de $(\hat{\mu}, \hat{\sigma})^t$: $\sqrt{n} (\widehat{\text{VaR}} - \text{VaR}) \sim N\left(0, \sigma^2 + z_\alpha^2 \frac{\sigma^2}{2}\right)$

El Método de los Momentos es un procedimiento consistente en igualar momentos poblacionales respecto al origen α_r a los correspondientes momentos muestrales respecto al origen a_r , formando así tantas ecuaciones como parámetros poblacionales se pretenden estimar:



$$\begin{aligned} \alpha_1 = E(X) = \mu &\Rightarrow \hat{\alpha}_1 = a_1 = \frac{\sum_{i=1}^n x_i}{n} = \bar{x} \\ \alpha_2 = E(X^2) &\Rightarrow \hat{\alpha}_2 = a_2 = \frac{\sum_{i=1}^n x_i^2}{n} \\ \dots\dots\dots \\ \alpha_k = E(X^k) &\Rightarrow \hat{\alpha}_k = a_k = \frac{\sum_{i=1}^n x_i^k}{n} \end{aligned}$$

110. Sea una población definida por: $P(X=-1) = \frac{1-\theta}{2}$, $P(X=0) = \frac{\theta+\lambda}{2}$, $P(X=1) = \frac{1-\lambda}{2}$, donde $0 < \theta < 1$, $0 < \lambda < 1$. Estimar los parámetros θ y λ por el método de los momentos, estudiando si son insesgados.

Solución:

Puesto que hay que estimar dos parámetros θ y λ hay que calcular los dos primeros momentos.

Momentos poblacionales

$$\begin{aligned} \alpha_1 = \mu = E(X) &= \sum_i x_i P(X=x_i) = \\ &= (-1) \left(\frac{1-\theta}{2} \right) + (0) \left(\frac{\theta+\lambda}{2} \right) + (1) \left(\frac{1-\lambda}{2} \right) = \frac{\theta-\lambda}{2} \end{aligned}$$

Momentos poblacionales

$$\begin{aligned}\alpha_2 &= E(X^2) = \sum_i x_i^2 P(X=x_i) = \\ &= (-1)^2 \left(\frac{1-\theta}{2}\right) + (0)^2 \left(\frac{\theta+\lambda}{2}\right) + (1)^2 \left(\frac{1-\lambda}{2}\right) = \frac{2-\theta-\lambda}{2}\end{aligned}$$

Momentos muestrales

$$a_1 = \bar{x} = \frac{\sum_i x_i}{n} \quad a_2 = \frac{\sum_i x_i^2}{n}$$

$$\begin{cases} \alpha_1 = a_1 \rightarrow \frac{\theta-\lambda}{2} = \bar{x} \rightarrow \theta-\lambda = 2\bar{x} \\ \alpha_2 = a_2 \rightarrow \frac{2-\theta-\lambda}{2} = a_2 \rightarrow -\theta-\lambda = 2a_2-2 \end{cases} \rightarrow \begin{cases} \theta-\lambda = 2\bar{x} \\ -\theta-\lambda = 2a_2-2 \end{cases} \rightarrow \begin{cases} \hat{\lambda} = 1-a_2-\bar{x} \\ \hat{\theta} = 1-a_2+\bar{x} \end{cases}$$

▪ **Insesgadez**

$$E(\hat{\theta}) = E(1-a_2+\bar{x}) = 1-E(a_2)+E(\bar{x}) = 1-\alpha_2+\mu = 1-\left(\frac{2-\theta-\lambda}{2}\right)+\left(\frac{\theta-\lambda}{2}\right) = \theta$$

$$E(\hat{\lambda}) = E(1-a_2-\bar{x}) = 1-E(a_2)-E(\bar{x}) = 1-\alpha_2-\mu = 1-\left(\frac{2-\theta-\lambda}{2}\right)-\left(\frac{\theta-\lambda}{2}\right) = \lambda$$

Los estimadores $\hat{\theta}$ y $\hat{\lambda}$ son insesgados.

111. Una muestra aleatoria $(x_1, \dots, x_n)$ de la población tiene como función de densidad

$$f_{\theta}(x) = \begin{cases} e^{-x+\theta} & \text{si } x > 0 \\ 0 & \text{en el resto} \end{cases}$$

Se pide:

- Hallar un estimador por el método de los momentos de θ .
- Estudiar si el estimador encontrado en el apartado anterior es insesgado para estimar el parámetro θ .

Solución:

$$\text{a) Se plantea la ecuación: } \bar{x} = E[x] = \int_0^{\infty} x f_{\theta}(x) dx = \overbrace{\int_0^{\infty} x e^{-x+\theta} dx}^{\text{integración por partes}} = \theta + 1 \rightarrow \hat{\theta} = \bar{x} - 1$$

Integración por partes:

$$\int_{\theta}^{\infty} x e^{-x+\theta} dx = \underbrace{x}_{u} \underbrace{(-e^{-x+\theta})}_{v} - \int \underbrace{(-e^{-x+\theta})}_{v} \underbrace{1}_{du} dx = -x e^{-x+\theta} - e^{-x+\theta} = -(1+x) e^{-x+\theta} = -e^{-\theta} \left(\frac{1+x}{e^x} \right)$$

$$\int_{\theta}^{\infty} x e^{-x+\theta} dx = -e^{-\theta} \left(\frac{1+x}{e^x} \right)_{\theta}^{\infty} = \theta + 1$$

b) Un estimador es insesgado o centrado cuando su valor probable coincide con el valor del parámetro a estimar. Es decir, $E(\hat{\theta}) = \theta$

$$E(\hat{\theta}) = E(\bar{x} - 1) = E(\bar{x}) - 1 = (\theta + 1) - 1 = \theta$$

112. El coseno X del ángulo con el que se emiten los electrones en un proceso radioactivo es

una variable aleatoria con función de densidad $f_{\theta}(x) = \begin{cases} (1+\theta x)/2 & -1 \leq x \leq 1 \\ 0 & \text{en el resto} \end{cases} \quad -1 \leq \theta \leq 1$

Considerando una muestra aleatoria $(x_1, \dots, x_n)$ de esta variable aleatoria, se pide:

- Obtener el estimador θ por el método de los momentos.
- Calcular la varianza de este estimador y demostrar que es consistente.

Solución:

a) Se plantea la ecuación: $\bar{x} = E[x] = \int_{-1}^1 x \frac{1+\theta x}{2} dx = \left[\frac{x^2}{2} + \frac{\theta x^3}{6} \right]_{-1}^1 = \frac{\theta}{3} \Rightarrow \hat{\theta} = 3\bar{x}$

b) $V(\hat{\theta}) = V(3\bar{x}) = 9V(\bar{x}) = 9 \frac{V(x)}{n} = \frac{9}{n} V(x)$

$$V(x) = E(x^2) - [E(x)]^2 = \int_{-1}^1 x^2 \frac{1+\theta x}{2} dx - \left[\frac{\theta}{3} \right]^2 = \left[\frac{x^3}{6} + \frac{\theta x^4}{8} \right]_{-1}^1 - \left[\frac{\theta}{3} \right]^2 = \frac{3-\theta^2}{9}$$

de donde, $V(\hat{\theta}) = \frac{9}{n} V(x) = \frac{9}{n} \left[\frac{3-\theta^2}{9} \right] = \frac{3-\theta^2}{n}$

Para probar que $\hat{\theta}$ es consistente para estimar θ es suficiente probar $\lim_{n \rightarrow \infty} E(\hat{\theta}) = \theta$
 $\lim_{n \rightarrow \infty} V(\hat{\theta}) = 0$

$$\lim_{n \rightarrow \infty} E(\hat{\theta}) = \lim_{n \rightarrow \infty} E(3\bar{x}) = \lim_{n \rightarrow \infty} 3E(\bar{x}) = 3E\left(\frac{\theta}{3}\right) = 3 \frac{\theta}{3} = \theta$$

$$\lim_{n \rightarrow \infty} V(\hat{\theta}) = \lim_{n \rightarrow \infty} V(3\bar{x}) = \lim_{n \rightarrow \infty} \frac{3-\theta^2}{n} = 0$$

Por tanto, queda probado que $\hat{\theta}$ es consistente para estimar θ

113. Sea una variable aleatoria X procedente de una población con densidad de probabilidad $N(\mu, 5)$. Efectuadas dos hipótesis sobre el valor de μ : $H_0: \mu = 12$ $H_1: \mu = 15$

Mediante un muestreo aleatorio simple de tamaño 25, se contrasta la hipótesis H_0 respecto de la hipótesis H_1 , estableciéndose que si la media muestral es menor que 14 se aceptaría la hipótesis nula. Determinar:

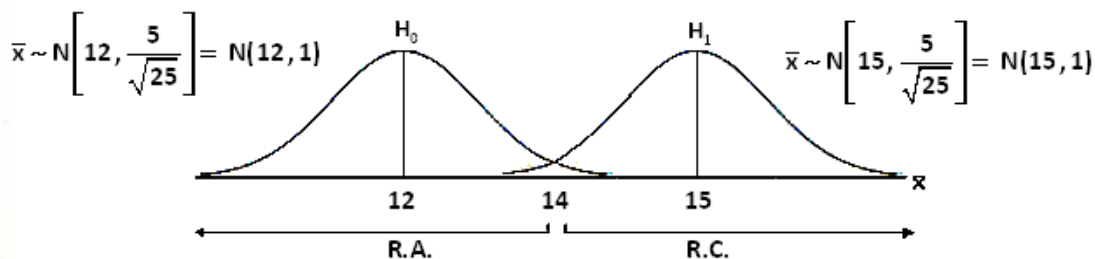
- La probabilidad de cometer el error tipo I
- La probabilidad de cometer el error tipo II
- La potencia del contraste

Solución:

a) Se trata de un contraste unilateral de hipótesis sobre la media poblacional

Regla de decisión $\begin{cases} \bar{x} \leq 14 & \text{se acepta } H_0 \equiv \text{región aceptación R.A.} \\ \bar{x} > 14 & \text{se rechaza } H_0 \equiv \text{región crítica R.C.} \end{cases}$

$H_0: \mu = 12$ $H_1: \mu = 15$



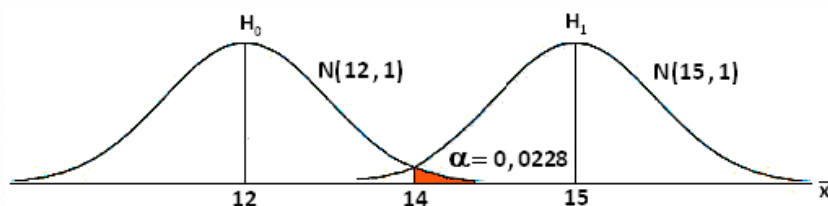
La distribución de la media muestral $\bar{x}$ de tamaño 25 bajo la hipótesis nula y la hipótesis alternativa viene dada:

$$H_0: \bar{x} \sim N\left[\mu_0, \frac{\sigma}{\sqrt{n}}\right] = N\left[12, \frac{5}{\sqrt{25}}\right] = N(12, 1) \quad H_1: \bar{x} \sim N\left[\mu_1, \frac{\sigma}{\sqrt{n}}\right] = N\left[15, \frac{5}{\sqrt{25}}\right] = N(15, 1)$$

Error Tipo I: $\alpha = P(\text{ET I}) = P(\text{Rechazar } H_0 / H_0 \text{ Cierta})$

$$\alpha = P(\text{ET I}) = P[\bar{x} > 14 \mid H_0: \mu = 12] = P\left[z \geq \frac{14 - 12}{1}\right] = P(z \geq 2) = 0,0228$$

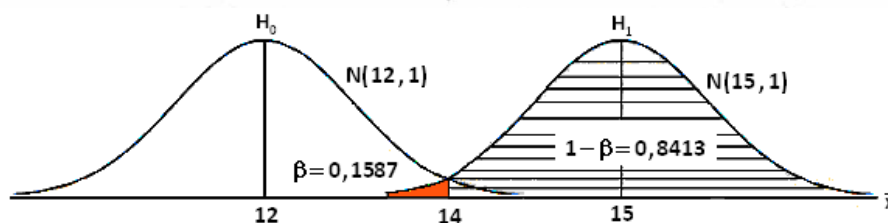
H_0 se rechaza cuando es cierta el 2,28% de los casos



b) Error Tipo II: $\beta = P(\text{ET II}) = P(\text{Aceptar } H_0 / H_0 \text{ Falsa})$

$$\beta = P(\text{ET II}) = P[\bar{x} \leq 14 \mid H_1: \mu = 15] = P\left[z \leq \frac{14 - 15}{1}\right] = P(z \leq -1) = P(z \geq 1) = 0,1587$$

H_0 se acepta cuando es falsa el 15,87% de los casos



c) Potencia del Contraste: $\text{Potencia} = P[\text{Rechazar } H_0 \mid H_0 \text{ Falsa}] = 1 - \beta$

$$1 - \beta = P[\text{Rechazar } H_0 \mid H_0 \text{ Falsa}] = 1 - 0,1587 = 0,8413 \quad (\text{Bondad del tipo de ajuste})$$

O bien,

$$\begin{aligned} \text{Potencia} &= P[\bar{x} \geq 14 \mid N(15, 1)] = P\left[\frac{\bar{x} - 15}{1} \geq \frac{14 - 15}{1}\right] = P[z \geq -1] = 1 - P[z \leq -1] = \\ &= 1 - P[z \leq -1] = 1 - P[z \geq 1] = 1 - 0,1587 = 0,8413 \end{aligned}$$

H_0 se rechaza cuando es falsa el 84,13% de los casos

Resaltar que es más grave cometer un Error Tipo I (α) que un Error Tipo II (β)

114. Las latas de mejillones de una determinada marca indican que el peso escurrido de dicho producto es de 250 gramos. No obstante, un consumidor está convencido de que el peso escurrido medio de dicho producto es menor que el que indican las latas. Si el peso escurrido sigue una ley normal con desviación típica 9 gramos. Se pide:

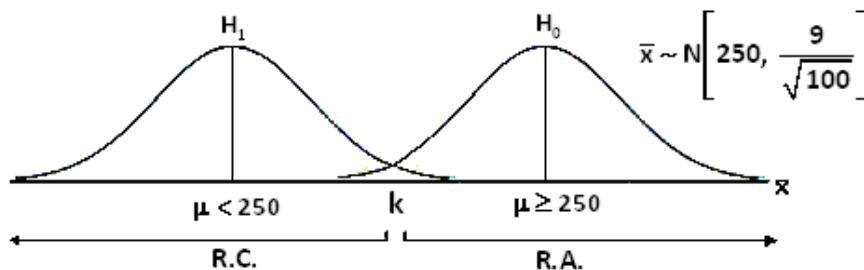
- Determinar, si existe, la mejor región crítica para contrastar, con un nivel de significación del 5% y muestras aleatorias simples de tamaño 100.
- Tomar una decisión acerca del rechazo o no de la hipótesis nula a partir de una muestra aleatoria simple de tamaño 100 en la cual se ha observado un peso escurrido promedio de 245 gramos.
- Determinar la función de potencia del contraste.

Solución:

a) Sea la variable aleatoria $X =$ "Peso escurrido de las latas de mejillones"

Contraste unilateral: $H_0: \mu \geq 250$ $H_1: \mu < 250$

Regla de decisión del muestreo: $\begin{cases} \bar{x} \geq k & \text{se acepta } H_0 \text{ (R.A.)} \\ \bar{x} < k & \text{se rechaza } H_0 \text{ (R.C.)} \end{cases}$



Bajo la hipótesis nula $\bar{x} \sim N\left[250, \frac{9}{\sqrt{100}}\right] \equiv N(250, 0,9)$

El valor crítico k , bajo la hipótesis nula, se determina con el nivel de significación α :

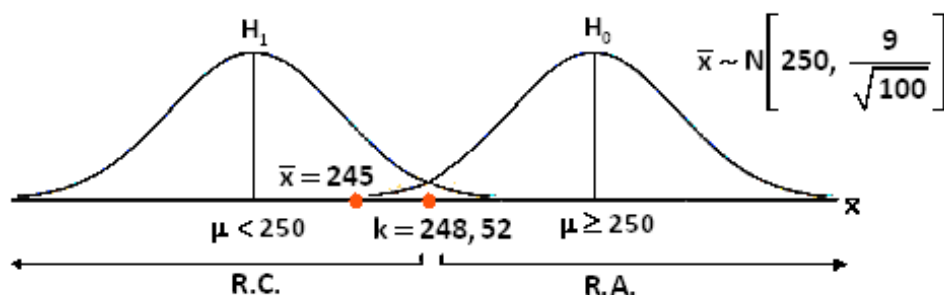
$$\begin{aligned} \alpha &= P[ET I] = P[\text{Rechazar } H_0 / H_0 \text{ Cierta}] = P[\bar{x} \leq k / H_0 \text{ Cierta}] = \\ &= P[\bar{x} \leq k / N(250; 0,9)] = P\left[\frac{\bar{x} - 250}{0,9} \leq \frac{k - 250}{0,9}\right] = P\left[z \leq \frac{k - 250}{0,9}\right] = P\left[z \geq \frac{250 - k}{0,9}\right] = 0,05 \end{aligned}$$

Observando las tablas de la $N(0,1)$ se tiene: $\frac{250 - k}{0,9} = 1,645 \rightarrow k = 248,52$

La región crítica más potente para muestras de tamaño 100 es $\bar{x} \geq 248,52$

b) Dado que $\bar{x} = 245 < 248,52$, el peso escurrido promedio se encuentra en la región de rechazo de la hipótesis nula.

$$\begin{cases} \bar{x} \geq 248,52 \text{ se acepta } H_0 \text{ (R.A.)} \\ \bar{x} < 248,52 \text{ se rechaza } H_0 \text{ (R.C.)} \end{cases}$$



c) La función potencia del contraste se establece como:

$$P(\mu) = P(\bar{x} \leq 248,52) = P\left[\bar{x} \leq k / N\left(\mu; \frac{9}{\sqrt{100}}\right)\right] = P\left[\frac{\bar{x} - \mu}{0,9} \leq \frac{248,52 - \mu}{0,9}\right] = P\left[z \leq \frac{248,52 - \mu}{0,9}\right]_{\mu \leq 250}$$

(bondad del tipo de ajuste)

115. Sea una variable aleatoria X procedente de una población con densidad de probabilidad $N(\mu, 4)$. Se quiere contrastar la hipótesis nula $H_0: \mu = 10$ frente a la hipótesis alternativa $H_1: \mu = 12$, con un nivel de significación $\alpha = 0,05$, con un muestreo simple de tamaño 25.

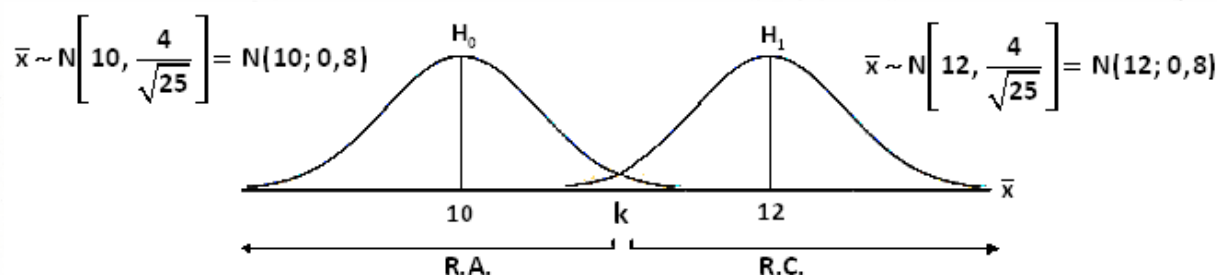
Determinar:

- La probabilidad de cometer el error tipo II
- La potencia del contraste

Solución:

$$\text{a) Regla de decisión } \begin{cases} \bar{x} \leq k & \text{se acepta } H_0 \rightarrow \text{región aceptación (R.A.)} \\ \bar{x} > k & \text{se rechaza } H_0 \rightarrow \text{región crítica (R.C.)} \end{cases}$$

Las hipótesis sobre la media poblacional (contraste unilateral):



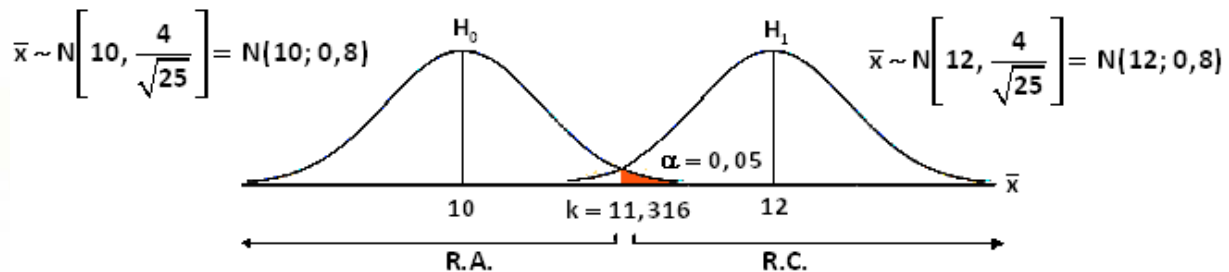
La distribución de la media muestral $\bar{x}$ de tamaño 25 bajo la hipótesis nula y alternativa, respectivamente:

$$\bar{x} \sim N\left[\mu_0, \frac{\sigma}{\sqrt{n}}\right] = N\left[10, \frac{4}{\sqrt{25}}\right] = N(10; 0,8) \quad \bar{x} \sim N\left[\mu_1, \frac{\sigma}{\sqrt{n}}\right] = N\left[12, \frac{4}{\sqrt{25}}\right] = N(12; 0,8)$$

Para hallar el valor crítico 'k' se recurre al Error Tipo I:

$$\alpha = P[\text{ET I}] = P[\text{Rechazar } H_0 / H_0 \text{ cierta}] = 0,05 \rightarrow$$

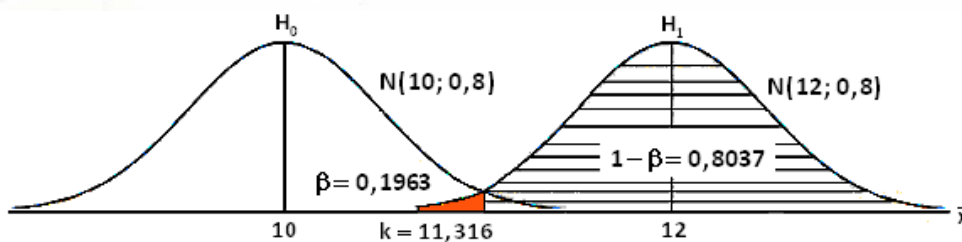
$$\alpha = P[\text{ET I}] = P[\bar{x} > k \mid H_0 : \mu_0 = 10] = P\left[z > \frac{k-10}{0,8}\right] = 0,05 \rightarrow \frac{k-10}{0,8} = 1,645 \Rightarrow k = 11,316$$



$$\beta = P[\text{ET II}] = P[\text{Aceptar } H_0 / H_0 \text{ Falsa}] =$$

$$\text{Error Tipo II: } = P[\text{ET II}] = P[\bar{x} \leq 11,316 \mid H_1 : \mu_1 = 12] = P\left[z \leq \frac{11,316 - 12}{0,8}\right] = P(z \leq -0,855) =$$

$$= P(z \geq 0,855) = 0,1963$$



b) Potencia del Contraste: Potencia = $P[\text{Rechazar } H_0 \mid H_0 \text{ Falsa}] = 1 - \beta = 1 - 0,1963 = 0,8037$

116. Un agricultor sabe que el peso en kg. De las patatas sigue una distribución $N(\mu, 1)$. Una muestra de patatas dio un peso medio de 330 gramos. Con la muestra se realizó un contraste, con un nivel de significación del 5% y una potencia de 0,6406, en el que la hipótesis nula era $\mu = 0,4$ Kg y la alternativa $\mu = 0,3$ Kg. Se pide:

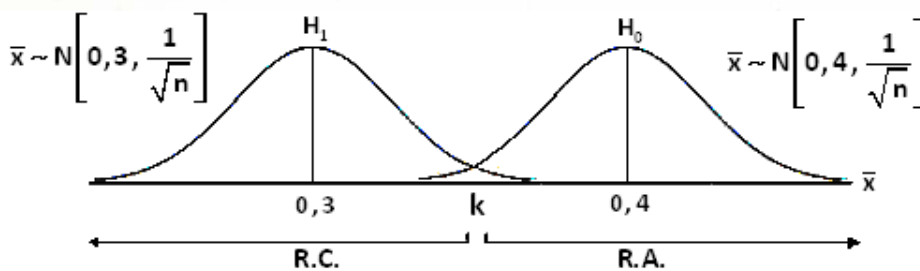
- a) ¿Cuál es el tamaño de la muestra utilizada por el agricultor?
b) Qué hipótesis fue aceptada

Solución:

a) La variable aleatoria $X =$ 'Peso en kg. de las patatas', $X \sim N(\mu, 1)$

Hipótesis sobre μ : $H_0: \mu_0 = 0,4$ $H_1: \mu_1 = 0,3$

Regla de decisión $\begin{cases} \bar{x} \geq k & \rightarrow \text{Se acepta } H_0 \text{ (R.A.)} \\ \bar{x} < k & \rightarrow \text{Se rechaza } H_0 \text{ (R.C.)} \end{cases}$



La distribución de la media muestral, bajo la hipótesis nula, sigue una ley $N\left[0,4, \frac{1}{\sqrt{n}}\right]$

El valor crítico 'k' se obtiene a partir del nivel de significación:

$$\begin{aligned} \alpha &= P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P[\bar{x} < k \mid H_0 \text{ Cierta}] = P[\bar{x} < k \mid \mu = 0,4] = \\ &= P\left[\frac{\bar{x} - 0,4}{1/\sqrt{n}} < \frac{k - 0,4}{1/\sqrt{n}}\right] = P\left[\frac{\bar{x} - 0,4}{1/\sqrt{n}} \geq \frac{0,4 - k}{1/\sqrt{n}}\right] = 0,05 \xrightarrow{N(0,1)} \frac{0,4 - k}{1/\sqrt{n}} = 1,64 \end{aligned}$$

Por otro lado, como la potencia del contraste es 0,6406:

$$\text{Potencia} = P[\text{Rechazar } H_0 \mid H_0 \text{ Falsa}] \equiv P[\text{Rechazar } H_0 \mid H_1 \text{ Cierta}]$$

La media muestral bajo la hipótesis alternativa sigue una ley $N\left[0,3, \frac{1}{\sqrt{n}}\right]$, con lo cual,

$$\text{Potencia} = P\left[\frac{\bar{x} - 0,3}{1/\sqrt{n}} < \frac{k - 0,3}{1/\sqrt{n}}\right] = 0,6406 \rightarrow P\left[z \geq \frac{k - 0,3}{1/\sqrt{n}}\right] = 0,3594 \xrightarrow{N(0,1)} \frac{k - 0,3}{1/\sqrt{n}} = 0,36$$

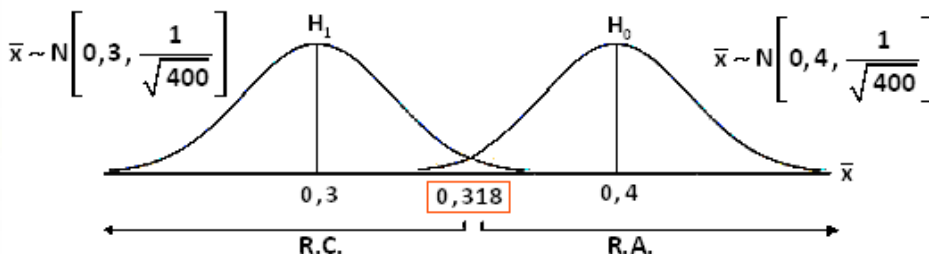
Resolviendo el sistema:

$$\begin{cases} 0,4 - k = 1,64 / \sqrt{n} \\ k - 0,3 = 0,36 / \sqrt{n} \end{cases} \rightarrow 0,1 = 2 / \sqrt{n} \rightarrow \sqrt{n} = 2 / 0,1 \rightarrow n = 400$$

El tamaño de la muestra de patatas utilizada por el agricultor es de 400 patatas.

Se determina el valor crítico k: $\frac{0,4 - k}{1 / \sqrt{400}} = 1,64 \rightarrow k = 0,318$

b) Regla de decisión $\begin{cases} \bar{x} \geq 0,318 \rightarrow \text{Se acepta } H_0 \text{ (R.A.)} \\ \bar{x} < 0,318 \rightarrow \text{Se rechaza } H_0 \text{ (R.C.)} \end{cases}$



Siendo $\bar{x} = 0,330 > 0,318$ se acepta la hipótesis nula de que el peso medio de las patatas es de 400 gramos.

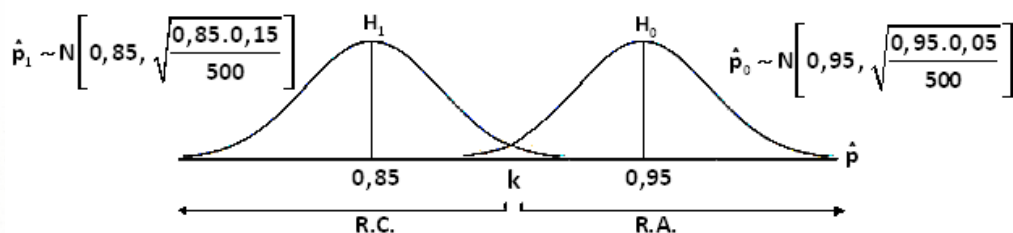
117. Un laboratorio farmacéutico quiere lanzar un nuevo medicamento para la hipertensión, llamado Hipotensil. El director de dicho laboratorio cree que la eficacia del medicamento sería de un 95%, medida ésta como la proporción de pacientes a los que se les suministra y que experimentan una mejoría. Sin embargo, el inspector de sanidad del Ministerio no es tan optimista y opina que la eficacia es sólo del 85%. Para analizar la eficacia del medicamento antes de su comercialización, se selecciona una muestra aleatoria de 500 pacientes, a los que se les administra Hipotensil, de los cuales mejoran 467. ¿Tiene razón el director del laboratorio? Suponga un nivel de significación del 5%.

Solución:

Sea la variable aleatoria $X = \text{'eficacia Hipotensil'} \sim B(1, p)$

Las hipótesis sobre la proporción (contraste unilateral izquierda): $H_0: p_0 = 0,95 \quad H_1: p_1 = 0,85$

Regla de decisión para el valor crítico k $\begin{cases} \hat{p} \geq k \text{ se acepta } H_0 \text{ (R.A.)} \\ \hat{p} < k \text{ se rechaza } H_0 \text{ (R.C.)} \end{cases}$



La distribución en el muestreo del estadístico $\hat{p} = \frac{\sum_{i=1}^n x_i}{n} \underset{\text{Teorema Central Límite}}{\sim} N\left[p, \sqrt{\frac{p \cdot q}{n}}\right]$

Bajo la hipótesis H_0 : $\hat{p}_0 \sim N\left[0,95, \sqrt{\frac{0,95 \cdot 0,05}{500}}\right] = N(0,95, 0,00974)$

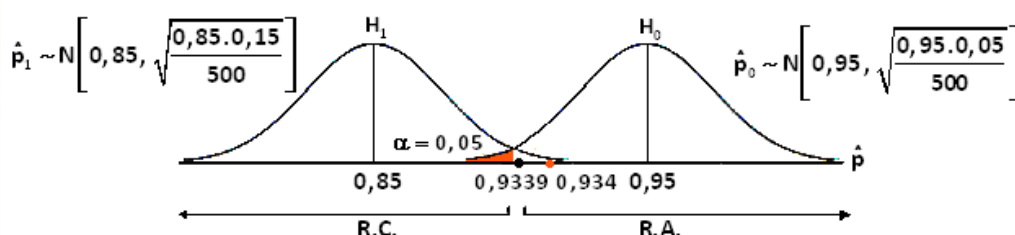
Bajo la hipótesis H_1 : $\hat{p}_1 \sim N\left[0,85, \sqrt{\frac{0,85 \cdot 0,15}{500}}\right] = N(0,85, 0,01597)$

Se determina el valor crítico k a partir del nivel de significación α :

$$\alpha = P[\text{ET I}] = P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P[\hat{p} < k] = P\left[z < \frac{k - 0,95}{0,00974}\right] = P\left[z \geq \frac{0,95 - k}{0,00974}\right] = 0,05$$

Observando en las tablas de la normal $N(0, 1)$, resulta:

$$\frac{0,95 - k}{0,00974} = 1,645 \rightarrow k = 0,95 - (0,00974 \cdot 1,645) = 0,9339$$



Regla de decisión $\begin{cases} \hat{p} \geq 0,9339 \text{ se acepta } H_0 \text{ (R.A.)} \\ \hat{p} < 0,9339 \text{ se rechaza } H_0 \text{ (R.C.)} \end{cases}$

El valor del estadístico muestral $\hat{p}$ (evidencia empírica) es $\hat{p} = \frac{467}{500} = 0,934$, siendo

$\hat{p} = 0,934 > 0,9339$, concluyendo que existe evidencia empírica suficiente para aceptar la hipótesis nula H_0 . Es decir, el Hipotensil es eficaz en un 95% de los casos.

118. Un portal e-business sabe que el 60% de todos sus visitantes a la web están interesados en adquirir sus productos pero no reacios al comercio electrónico y no realizan finalmente la compra vía internet. Sin embargo, en la dirección del portal se piensa que en el último año, el porcentaje de gente que está dispuesta a comprar por internet ha aumentado y esto se debe reflejar en sus resultados empresariales.

- a) Contrastar con un nivel de significación del 2%, si en el último año se ha reducido el porcentaje de gente que no está dispuesta a comprar por internet, si para ello se tomó una muestra de 500 visitantes para conocer su opinión y se observó que el 55% no estaba dispuesto a realizar compras vía on-line.
- b) Realizar el contraste con el p_valor

Solución:

Sea el parámetro p = "Proporción del número de visitantes al portal". Al realizar el contraste sobre la proporción, se parte de una muestra aleatoria $(x_1, x_2, \dots, x_{500})$, donde $X \sim B(1, p)$

La distribución del parámetro muestral $\hat{p} = \sum_{i=1}^{500} x_i / n$ al ser de tamaño suficientemente grande $n = 500$ y estar $\hat{p}$ definido como suma de variables independientes entre sí según una distribución de Bernoulli $B(1, p)$, por el teorema central del límite (TCL) se puede aproximar

$$\hat{p} = \frac{\sum_{i=1}^{500} x_i}{n} \sim N\left(p, \sqrt{\frac{pq}{n}}\right)$$

Se establecen las hipótesis: $H_0: p = 0,6$ $H_1: p < 0,6$

Se trata de un contraste unilateral por la derecha

Regla de decisión: $\begin{cases} \hat{p} \geq k & \text{se acepta } H_0 \\ \hat{p} < k & \text{se rechaza } H_0 \end{cases}$

Bajo la hipótesis nula $\hat{p} \sim N\left(0,6, \sqrt{\frac{0,6 \times 0,4}{500}}\right) = N(0,6, 0,022)$

A partir del nivel de significación $\alpha = 0,02$ se determina el valor crítico k :

$$\begin{aligned} \alpha &= P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P[\hat{p} < k \mid N(0,6, 0,022)] = P\left[\frac{\hat{p} - 0,6}{0,022} < \frac{k - 0,6}{0,022}\right] = \\ &= P\left[z < \frac{k - 0,6}{0,022}\right] = P\left[z \geq \frac{0,6 - k}{0,022}\right] = 0,02 \rightarrow \frac{0,6 - k}{0,022} = 2,055 \Rightarrow k = 0,555 \end{aligned}$$

El valor del estadístico muestral (evidencia empírica) es $\hat{p} = 0,55 < 0,555$, rechazando la hipótesis nula. En conclusión, existe evidencia empírica que la proporción de visitantes al portal

que están dispuestos a comprar on-line ha disminuido, es decir, el porcentaje de visitantes que son reacios a comprar por internet ha aumentado.

b) El p -valor es el nivel de significación más pequeño posible que puede escogerse para el que todavía se aceptaría la hipótesis alternativa con las observaciones actuales. Paquetes Estadísticos proporcionan el *significance level* (nivel de significación) cuando se refieren en realidad al p -valor (p -value).

$$\alpha_p = p\text{-valor} = P[\text{Rechazar el estadístico muestral observado} / H_0 \text{ es Cierta}]$$

$$\alpha_p = P\left[\hat{p} < 0,55 \mid N(0,6, 0,022)\right] = P\left[\frac{\hat{p} - 0,6}{0,022} < \frac{0,55 - 0,6}{0,022}\right] = P(z < -2,27) = P(z \geq 2,27) = 0,0116$$

Siendo $\alpha_p = 0,0116 < \alpha = 0,02$ se rechaza la hipótesis nula a un nivel de significación del 2%

119. Se trata de determinar si en una ciudad el 20% o el 30% de las familias dispone de lavavajillas; para dilucidarlo se toma al azar una muestra de 400 familias de la mencionada ciudad y se adopta el criterio de sí en la muestra hay menos de 100 familias con lavavajillas, se rechaza que el 20% de las familias poseen el mencionado electrodoméstico. Se pide:

- Nivel de significación del test.
- Potencia del test.

Solución:

a) Sea el parámetro p = " Proporción de familias con lavavajillas"

Al realizar un contraste sobre una proporción se parte de una muestra aleatoria

$(X_1, X_2, \dots, X_n)$ de tamaño $n = 400$, donde $X \sim B(1, p)$.

Para calcular la probabilidad interesa conocer la distribución del parámetro muestral

$$\hat{p} = \frac{\sum_{i=1}^n x_i}{n} \quad \text{donde} \quad X_i = \begin{cases} 1 & \text{si la familia tiene lavavajillas} \\ 0 & \text{si la familia no tiene lavavajillas} \end{cases}$$

Al ser el tamaño suficientemente grande $n = 400$ y estar definido $\hat{p}$ como suma de variables independientes según una distribución de Bernouilli $B(1, p)$, se tiene:

La distribución en el muestreo del estadístico $\hat{p} = \frac{\sum_{i=1}^n x_i}{n}$ $\xrightarrow{\text{Teorema Central Límite}} N\left[p, \sqrt{\frac{p \cdot q}{n}}\right]$

Bajo la hipótesis $H_0: \hat{p} \sim N\left[0,2, \sqrt{\frac{0,2 \cdot 0,8}{400}}\right] = N(0,2, 0,02)$

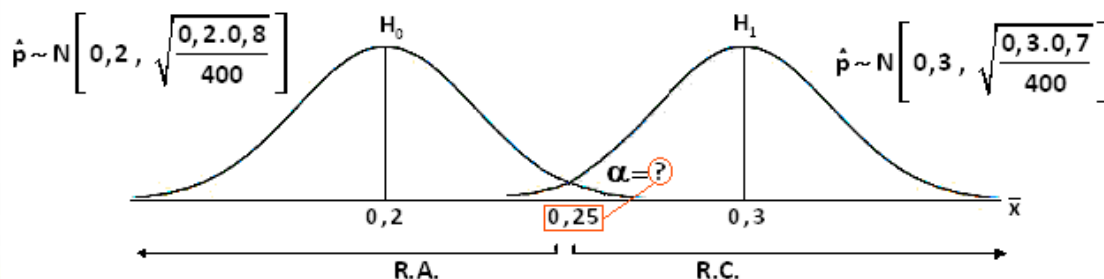
Bajo la hipótesis $H_1: \hat{p} \sim N\left[0,3, \sqrt{\frac{0,3 \cdot 0,7}{400}}\right] = N(0,3, 0,0229)$

Se establecen las hipótesis: $H_0: p = 0,2$ $H_1: p < 0,3$

Por el lema de Neyman-Pearson, la regla de decisión del muestreo $k = \frac{100}{400} = 0,25$ es:

$$\begin{cases} \hat{p} \leq 0,25 & \text{Se acepta } H_0 \text{ (R.A.)} \\ \hat{p} > 0,25 & \text{Se rechaza } H_0 \text{ (R.C.)} \end{cases}$$

Bajo la hipótesis nula, con el valor crítico $k = 0,25$ se determina el nivel de significación α :



$$\begin{aligned} \alpha &= P[\text{ET I}] = P[\text{Rechazar } H_0 / H_0 \text{ Cierta}] = P[\hat{p} > 0,25 / H_0 \text{ Cierta}] = \\ &= P[\hat{p} \geq 0,25 / N(0,2; 0,02)] = P\left[\frac{\hat{p} - 0,2}{0,02} \geq \frac{0,25 - 0,2}{0,02}\right] = P[z \geq 2,5] = 0,00621 \end{aligned}$$

$$\text{b) Potencia} = P[\text{Rechazar } H_0 \mid H_0 \text{ Falsa}] = 1 - \beta = 1 - P[\text{Aceptar } H_0 / H_0 \text{ Falsa}]$$

$$\begin{aligned} \beta &= P[\text{ET II}] = P[\hat{p} \leq 0,25 / N(0,3, 0,0229)] = P\left[\frac{\hat{p} - 0,3}{0,0229} \leq \frac{0,25 - 0,3}{0,0229}\right] = P[z \leq -2,1822] = \\ &= P[z \geq 2,1822] = 0,0144 \end{aligned}$$

En consecuencia, bondad del tipo de ajuste:

$$\text{Pot} = 1 - \beta = P[\text{Rechazar } H_0 \mid H_0 \text{ Falsa}] = 1 - 0,0144 = 0,9856$$

120. Las especificaciones de un tipo de báscula aseguran que los errores de los pesajes siguen una distribución $N(0, \sigma)$. Se quiere contrastar la afirmación sobre la dispersión que es igual a la unidad, frente a una hipótesis alternativa de que es el doble. Para ello se realizan 5 pesajes en las que el error cometido resultó ser: 1 0,9 -0,2 1,4 -0,7

Para un nivel de significación del 5% se pide enunciar una regla de decisión (obtener la región crítica) e indicar que hipótesis resulta aceptada.

Solución:

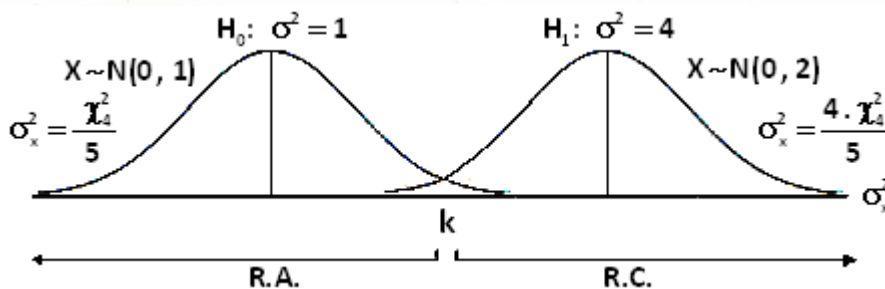
Sea la variable aleatoria $X = \text{'Errores en el peso'}$ $X \sim N(0, \sigma)$

$$\text{Distribución en el muestreo: } \chi_{n-1}^2 = \frac{\text{Varianza muestral}}{\text{Varianza teórica}} = \frac{\sigma_x^2}{(\sigma / \sqrt{n})^2} \rightarrow \sigma_x^2 = \frac{\sigma^2 \cdot \chi_{n-1}^2}{n}$$

Se establecen las hipótesis: $H_0: \sigma^2 = 1$ $H_1: \sigma^2 = 4$

$$\text{Bajo la hipótesis: } H_0: \sigma_x^2 = \frac{\chi_4^2}{5} \quad H_1: H_0: \sigma_x^2 = \frac{4 \cdot \chi_4^2}{5}$$

Regla de decisión: $\begin{cases} \sigma_x^2 \leq k & \text{se acepta } H_0 \\ \sigma_x^2 > k & \text{se rechaza } H_0 \end{cases}$



El valor crítico k se determina a partir del nivel de significación α (Error Tipo I):

$$\begin{aligned} \alpha &= P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P[\sigma_x^2 > k \mid H_0 \text{ Cierta}] = P[\sigma_x^2 > k \mid \sigma^2 = 1] = \\ &= P\left[\frac{\chi_4^2}{5} > k\right] = P[\chi_4^2 \geq 5k] = 0,05 \xrightarrow{\chi^2 \text{ Pearson}} 5k = 9,488 \rightarrow k = 1,898 \end{aligned}$$

Se compara el valor crítico $k = 1,898$ con la evidencia muestral

x_i	1	0,9	-0,2	1,4	-0,7	$\bar{x} = \sum_{i=1}^5 \frac{x_i}{5} = 0,48$
x_i^2	1	0,81	0,04	1,96	0,49	$\sum_{i=1}^5 \frac{x_i^2}{5} = 0,86$

$$\sigma_x^2 = \sum_{i=1}^5 \frac{x_i^2}{5} - \left(\sum_{i=1}^5 \frac{x_i}{5} \right)^2 = 0,6296$$

Como $\sigma_x^2 = 0,6296 < 1,898$ se sitúa en la región de aceptación (R.A), con un nivel de confianza del 95% no se puede rechazar la hipótesis de que la dispersión sea 1.

121. En una población $N(5, \sigma)$ se quiere contrastar la hipótesis nula $H_0: \sigma^2 = 2$ frente a la hipótesis alternativa $H_1: \sigma^2 = 3$, con un nivel de significación $\alpha = 0,025$, con una muestra aleatoria simple de tamaño 10:

5,1 6,2 4 2,8 2,9 5,6 3,7 3,4 2,5 5,2

Hallar la potencia del contraste.

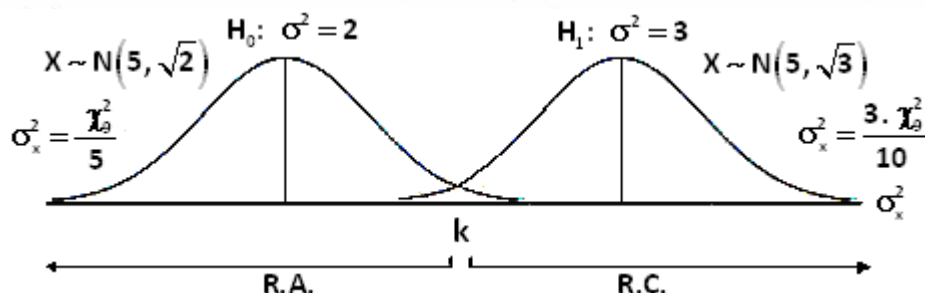
Solución:

Distribución en el muestreo: $\chi_{n-1}^2 = \frac{\text{Varianza muestral}}{\text{Varianza teórica}} = \frac{\sigma_x^2}{(\sigma / \sqrt{n})^2} \rightarrow \sigma_x^2 = \frac{\sigma^2 \cdot \chi_{n-1}^2}{n}$

Hipótesis: $H_0: \sigma^2 = 2$ $H_1: \sigma^2 = 3$

Bajo la hipótesis: $H_0: \sigma_x^2 = \frac{2 \cdot \chi_9^2}{10} = \frac{\chi_9^2}{5}$ $H_1: \sigma_x^2 = \frac{3 \cdot \chi_9^2}{10}$

Regla de decisión: $\begin{cases} \sigma_x^2 \leq k & \text{se acepta } H_0 \\ \sigma_x^2 > k & \text{se rechaza } H_0 \end{cases}$



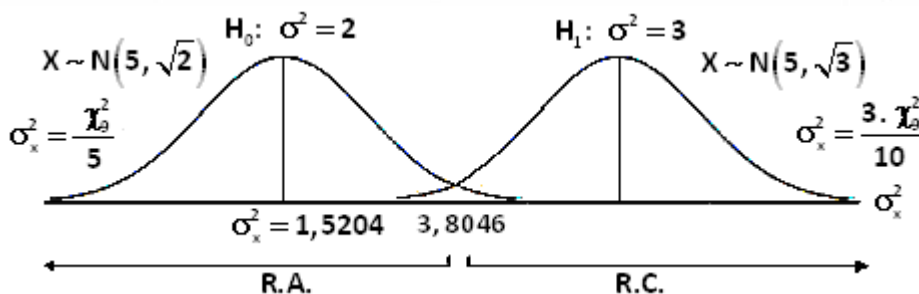
Determinación del valor crítico k a partir del nivel de significación α (Error Tipo I):

$$\begin{aligned}\alpha &= P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P[\sigma_x^2 > k \mid H_0 \text{ Cierta}] = P[\sigma_x^2 > k \mid \sigma^2 = 2] = \\ &= P\left[\frac{\chi_9^2}{5} > k\right] = P[\chi_9^2 > 5k] = 0,025 \xrightarrow{\chi^2 \text{ Pearson}} 5k = 19,023 \rightarrow k = 3,8046\end{aligned}$$

Se compara el valor crítico $k = 3,8046$ con la evidencia muestral:

x_i	5,1	6,2	4	2,8	2,9	5,6	3,7	3,4	2,5	5,2	$\bar{x} = \sum_{i=1}^{10} \frac{x_i}{10} = 4,14$
x_i^2	26,01	38,44	16	7,84	8,41	31,36	13,69	11,56	6,25	27,04	$\sum_{i=1}^{10} \frac{x_i^2}{10} = 18,66$

$$\sigma_x^2 = \sum_{i=1}^{10} \frac{x_i^2}{10} - \left(\sum_{i=1}^{10} \frac{x_i}{10}\right)^2 = 18,66 - 4,14^2 = 1,5204$$



Siendo $\sigma_x^2 = 1,5204 < 3,8046$ se sitúa en la región de aceptación (R.A), con un nivel de confianza del 97,5% se acepta que la la varianza es 2.

De otra parte, la potencia del contraste:

$$\begin{aligned}\text{Potencia} &= P[\text{Rechazar } H_0 \mid H_0 \text{ Falsa}] \equiv P[\text{Rechazar } H_0 \mid H_1 \text{ Cierta}] = P\left[\frac{3 \cdot \chi_9^2}{10} > 3,8046\right] = \\ &= P[\chi_9^2 \geq 12,682] = 0,252\end{aligned}$$

Abcisas	Áreas	
4,168 - 14,684 $\longleftrightarrow$ 0,90 - 0,10	-10,516 $\longleftrightarrow$ 0,80 $\rightarrow$ $x = 0,10 + \frac{2,002 \cdot 0,80}{10,516} = 0,252$	
12,682 - 14,684 $\longleftrightarrow$ x - 0,10	-2,002 $\longleftrightarrow$ x - 0,10	

122. En una población con distribución $N(\mu, \sigma)$ se desea contrastar la hipótesis nula $H_0: \sigma^2 = 25$ frente a la hipótesis alternativa $H_1: \sigma^2 = 36$, se toma una muestra aleatoria de tamaño 16, con varianza igual a 27.

- Tiene sentido utilizar el nivel de significación del 1%
- En caso negativo, utiliza el nivel de confianza más grande posible
- Hallar la potencia del contraste

Solución:

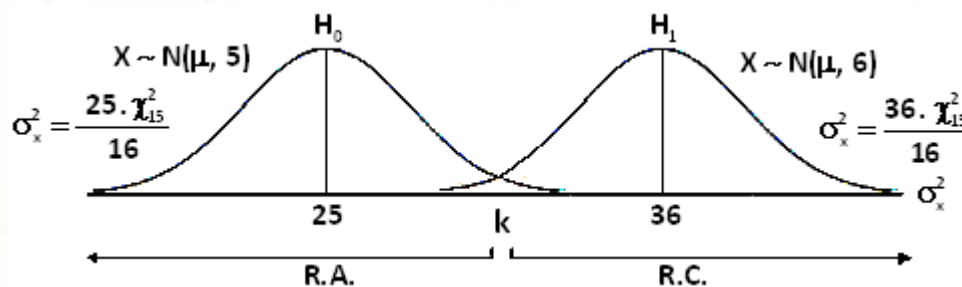
Por el Lema de Fisher-Cochran, la distribución en el muestreo:

$$\chi_{n-1}^2 = \frac{\text{Varianza muestral}}{\text{Varianza teórica}} = \frac{\sigma_x^2}{(\sigma / \sqrt{n})^2} \rightarrow \sigma_x^2 = \frac{\sigma^2 \cdot \chi_{n-1}^2}{n}$$

Hipótesis: $H_0: \sigma^2 = 25$ $H_1: \sigma^2 = 36$

Bajo la hipótesis: $H_0: \sigma_x^2 = \frac{25 \cdot \chi_{15}^2}{16}$ $H_1: \sigma_x^2 = \frac{36 \cdot \chi_{15}^2}{16}$

Regla de decisión: $\begin{cases} \sigma_x^2 \leq k & \text{se acepta } H_0 \\ \sigma_x^2 > k & \text{se rechaza } H_0 \end{cases}$



La determinación del valor crítico k a partir del nivel de significación $\alpha = 0,01$:

$$\begin{aligned} \alpha &= P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P[\sigma_x^2 > k \mid H_0 \text{ Cierta}] = P[\sigma_x^2 > k \mid \sigma^2 = 25] = \\ &= P\left[\frac{25 \cdot \chi_{15}^2}{16} > k\right] = P\left[\chi_{15}^2 > \frac{16 \cdot k}{25}\right] = 0,01 \xrightarrow{\chi^2 \text{ Pearson}} \frac{16 \cdot k}{25} = 30,578 \rightarrow k = 47,778 \end{aligned}$$

Adviértase que el valor crítico k se encuentra fuera del intervalo $[25, 36]$ con lo que no tiene sentido utilizar un nivel de confianza del 99%

- Para que el valor crítico se encuentre dentro del rango deseado se puede tomar un nivel de confianza del 90%, con lo cual $\alpha = 0,10$, resultando:

$$\alpha = P\left[\frac{25 \cdot \chi_{15}^2}{16} > k\right] = P\left[\chi_{15}^2 > \frac{16 \cdot k}{25}\right] = 0,10 \xrightarrow{\chi^2 \text{ Pearson}} \frac{16 \cdot k}{25} = 22,307 \rightarrow k = 34,855$$

Se compara el valor crítico $k = 34,855$ con la evidencia muestral $\sigma_x^2 = 27$

Como $\sigma_x^2 = 27 < 34,855$ se sitúa en la región de aceptación (R.A), con un nivel de confianza del 90% se acepta que la la varianza es 25.

c) Potencia del contraste con $\alpha = 0,10$

$$\begin{aligned} \text{Potencia} &= P[\text{Rechazar } H_0 \mid H_0 \text{ Falsa}] \equiv P[\text{Rechazar } H_0 \mid H_1 \text{ Cierta}] = \\ &= P\left[\frac{36 \cdot \chi_{15}^2}{16} > 34,855 \mid H_1 \text{ Cierta}\right] = P[\chi_{15}^2 \geq 15,491] = 0,496 \end{aligned}$$

Abscisas	Áreas	
8,547 – 22,307	↔ 0,90 – 0,10	–13,76 ↔ 0,80 → $x = 0,10 + \frac{6,816 \cdot 0,80}{13,76} = 0,496$
15,491 – 22,307	↔ x – 0,10	–6,816 ↔ x – 0,10

123. El propietario de un automóvil sospecha que su vehículo según una ley normal tiene un consumo medio de combustible en carretera superior a los 5,6 litros /100 km., que es lo que el fabricante indica en su publicidad. Para apoyar empíricamente su sospecha observa el consumo medio en 11 viajes seleccionados aleatoriamente entre todos los que realiza en el año, obteniendo los siguientes resultados: 6,1 6,5 5,1 6 5,9 5,2 5,8 5,3 6,2 5,9 6,3

Se pide:

- ¿Están fundadas las sospechas del propietario a un nivel de significación del 1%?
- Calcula el p_valor
- ¿En cuántas ocasiones debería observarse el consumo medio para que con un nivel de confianza del 99% se detectase un consumo medio de 5,9 litros/100 km?

Solución:

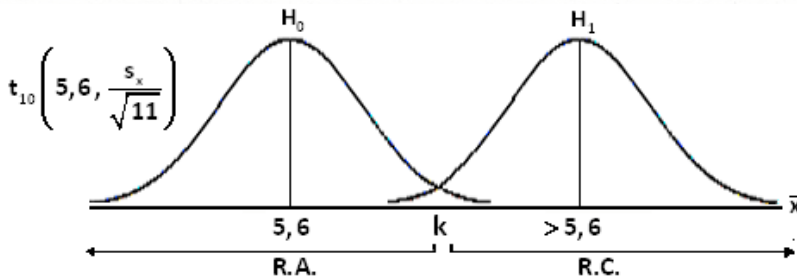
a) En el muestreo de una población normal $N(\mu, \sigma)$ con varianza poblacional desconocida, con

un tamaño muestral pequeño $n \leq 30$, la media muestral $\bar{x} \sim t_{n-1}\left(\mu, \frac{s_x}{\sqrt{n}}\right) \equiv t_{n-1}\left(\mu, \frac{\sigma_x}{\sqrt{n-1}}\right)$

El fabricante afirma que $H_0: \mu \leq 5,6$ y el propietario del vehículo cree que $H_1: \mu > 5,6$

Se trata, pues, de un contraste unilateral (cola a la derecha)

Regla de decisión: $\begin{cases} \bar{x} \leq k & \text{se acepta } H_0 \\ \bar{x} > k & \text{se rechaza } H_0 \end{cases}$



La media muestral bajo la hipótesis nula sigue una distribución t – Student con $(n-1)$ grados de

libertad: $t_{10}\left(5,6, \frac{0,4612}{\sqrt{11}}\right) \equiv t_{10}(5,6, 0,139)$

Con los datos muestrales:

x_i	6,1	6,5	5,1	6	5,9	5,2	5,8	5,3	6,2	5,9	6,3
x_i^2	37,21	42,25	26,01	36	34,81	27,04	33,64	28,09	38,44	34,81	39,69

$$\bar{x} = \sum_{i=1}^{11} \frac{x_i}{11} = \frac{64,3}{11} = 5,8454 \quad \sigma_x^2 = \sum_{i=1}^{11} \frac{x_i^2}{11} - \left(\sum_{i=1}^{11} \frac{x_i}{11} \right)^2 = \frac{377,99}{11} - \left(\frac{64,3}{11} \right)^2 = 0,19339$$

$$n \cdot \sigma_x^2 = (n-1) \cdot s_x^2 \rightarrow s_x^2 = \frac{n \cdot \sigma_x^2}{(n-1)} = \frac{11 \cdot 0,19339}{10} = 0,21273 \rightarrow s_x = \sqrt{0,21273} = 0,4612$$

El valor crítico k , bajo la hipótesis nula, se calcula a partir del nivel de significación $\alpha = 0,01$:

$$\begin{aligned} \alpha &= P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P[\bar{x} > k \mid H_0 \text{ Cierta}] = P[\bar{x} > k \mid \mu_0 = 5,6] = \\ &= P\left[\frac{\bar{x} - 5,6}{0,139} > \frac{k - 5,6}{0,139}\right] = P\left[t_{10} > \frac{k - 5,6}{0,139}\right] = 0,01 \rightarrow \frac{k - 5,6}{0,139} = 2,764 \rightarrow k = 5,9842 \end{aligned}$$

Siendo $\bar{x} = 5,8454 < 5,9842$ no se puede rechazar la hipótesis nula H_0 , con lo que se acepta las afirmaciones del fabricante sobre el consumo medio del automóvil.

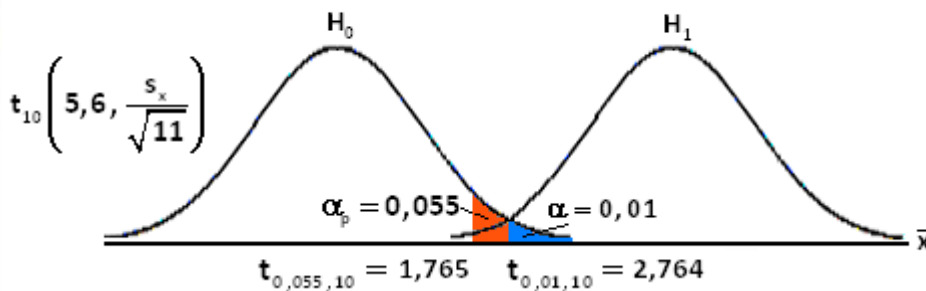
b) El p_valor (α_p) es el nivel de significación más pequeño posible que puede escogerse para el que todavía se aceptaría la hipótesis alternativa con las observaciones actuales.

Si $\alpha_p > \alpha$ se acepta la hipótesis nula H_0

$$\begin{aligned}\alpha_p = p_valor &= P[\text{Rechazar el estadístico muestral observado} / H_0 \text{ Cierta}] = \\ &= P[\bar{x} > 5,8454 \mid H_0 \text{ Cierta}] = P[\bar{x} > 5,8454 \mid t_{10}(5,6, 0,139)] = \\ &= P\left[\frac{\bar{x} - 5,6}{0,139} > \frac{5,8454 - 5,6}{0,139}\right] = P[t_{10} > 1,765] = 0,055\end{aligned}$$

En la tabla de la t -Student $\frac{1,372 - 1,812}{0,10 - 0,05} = \frac{1,765 - 1,812}{x - 0,05} \rightarrow x = 0,055$

Si $\alpha_p = 0,055 > \alpha = 0,01$ se acepta la hipótesis nula H_0



Para un nivel de significación $\alpha = 0,01$, el p -valor $\alpha_p = 0,055 > \alpha = 0,01$, aceptando la hipótesis nula. Es decir, con un nivel de confianza del 99% se acepta que el consumo medio de combustible en carretera superior es de 5,6 litros /100 km.

c) En esta ocasión se plantea la cuestión $P(\bar{x} > k / \mu = 5,9)$

Como el tamaño muestral es desconocido, el estadístico $\frac{\bar{x} - 5,9}{s_x / \sqrt{n}} \neq t_{10}$ no sigue una t -Student con 10 grados de libertad.

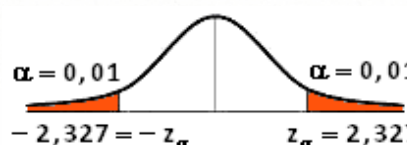
Por tanto, conociendo que $P\left[\frac{\bar{x} - 5,6}{s_x / \sqrt{n}} > 2,764\right] = 0,01$, se recurre a la siguiente estrategia:

$$\begin{aligned}P[\bar{x} > k / \mu = 5,9] &= P\left[\frac{\bar{x} - 5,6}{s_x / \sqrt{n}} > 2,764 \mid \mu = 5,9\right] = P\left[\frac{\bar{x} - 5,9 + 0,3}{s_x / \sqrt{n}} > 2,764\right] = \\ &= P\left[\frac{\bar{x} - 5,9}{s_x / \sqrt{n}} + \frac{0,3}{s_x / \sqrt{n}} > 2,764\right] = P\left[\frac{\bar{x} - 5,9}{s_x / \sqrt{n}} > 2,764 - \frac{0,3}{s_x / \sqrt{n}}\right] = 0,99\end{aligned}$$

$$\text{O bien, } P\left[\frac{\bar{x} - 5,9}{s_x/\sqrt{n}} \leq 2,764 - \frac{0,3}{s_x/\sqrt{n}}\right] = 0,01$$

En un punto donde no se pueden utilizar las tablas de la t – Student porque no se conoce el tamaño de la muestra. Para ello, se supone que el tamaño es suficientemente grande para que pueda ser aceptable la aproximación de la t mediante la distribución normal.

Con la aproximación normal y con la simetría de la $N(0, 1)$:



$$P\left[z \leq 2,764 - \frac{0,3}{s_x/\sqrt{n}}\right] = 0,01 \rightarrow 2,764 - \frac{0,3}{s_x/\sqrt{n}} = -2,327$$

$$\frac{0,3}{s_x/\sqrt{n}} = 5,091 \rightarrow n = \left[\frac{(5,091) \cdot s_x}{0,3}\right]^2$$

Otra vez en un callejón sin salida, aún es necesario conocer la cuasidesviación típica de una muestra sin saber su tamaño.

Se puede dar una salida, suponiendo que la cuasidesviación típica de esta nueva muestra es igual a la obtenida en la muestra anterior $s_x = 0,4612$.

$$\text{En este caso, } n = \left[\frac{(5,091) \cdot (0,4612)}{0,3}\right]^2 = 61,255$$

Para mayor seguridad en el logro del objetivo, se redondea con el entero inmediato superior, esto es, el tamaño de la muestra es 62.

124. El directorio de uno de los grandes operadores de Internet está considerando la posibilidad de ofrecer tarifa plana a sus clientes. Según sus conocimientos sobre el tema, sabe que está trabajando con una variable aleatoria que se distribuye como una normal. Mantiene la hipótesis de que los hogares que tienen Internet se conectan con una media de 5 horas mensuales. No obstante, existen otros estudios que sostienen que el tiempo de conexión es más alto. Para evaluar, a un 10% de significación, dicha hipótesis, el directorio decide encuestar a una muestra aleatoria de 30 hogares, obteniendo una media de 5,34 horas de conexión, con una dispersión de 7,24 horas.

- Formular el contraste a realizar
- Determinar la mejor región crítica del contraste
- ¿Se puede rechazar la hipótesis nula?
- Calcular y representar la función de potencia. ¿Qué representa la función de potencia?. Para facilitar los cálculos, suponer que el tamaño muestral es 300 hogares.

Solución:

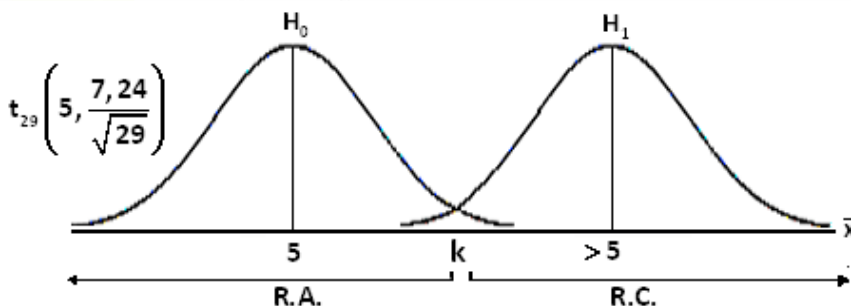
- a) Sea la variable aleatoria $X = \text{'Conexión a Internet por hogares'}$ $X \sim N(5, \sigma)$

Hipótesis de contraste: $H_0: \mu_0 \geq 5$ $H_1: \mu_1 > 5$

- b) Se trata de un contraste unilateral (cola derecha). Regla de decisión: $\begin{cases} \bar{x} \leq k & \text{se acepta } H_0 \\ \bar{x} > k & \text{se rechaza } H_0 \end{cases}$

En el muestreo de una población normal $N(\mu, \sigma)$ con varianza poblacional desconocida, con un tamaño muestral pequeño $n \leq 30$, la media muestral bajo la hipótesis nula sigue una

distribución $\bar{x} \sim t_{n-1} \left(\mu, \frac{\sigma_x}{\sqrt{n-1}} \right) \equiv t_{29} \left(5, \frac{7,24}{\sqrt{29}} \right)$



- c) Determinación del valor crítico k a partir del nivel de significación $\alpha = 0,10$:

$$\begin{aligned} \alpha &= P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P[\bar{x} > k \mid H_0 \text{ Cierta}] = P\left[\bar{x} > k \mid t_{29} \left(5, \frac{7,24}{\sqrt{29}} \right)\right] = \\ &= P\left[\frac{\bar{x} - 5}{7,24 / \sqrt{29}} > \frac{k - 5}{7,24 / \sqrt{29}}\right] = P\left[t_{29} > \frac{k - 5}{7,24 / \sqrt{29}}\right] = 0,10 \rightarrow \frac{\sqrt{29} \cdot (k - 5)}{7,24} = 1,311 \end{aligned}$$

$$k = 5 + \frac{1,311 \cdot 7,24}{\sqrt{29}} = 6,76$$

Regla de decisión: $\begin{cases} \bar{x} \leq 6,76 & \text{se acepta } H_0 \\ \bar{x} > 6,76 & \text{se rechaza } H_0 \end{cases}$

La media muestral $\bar{x}_0 = 5,34 < 6,76$ se encuentra en la región de aceptación, por lo que se acepta la hipótesis nula H_0 , concluyendo que tiene razón el Directivo, el tiempo medio de conexión es de 5 horas.

✓ Alternativamente, el problema se puede plantear desde la región de rechazo:

$$R = \left\{ \bar{x} - \mu_0 > t_{\alpha, n-1} \frac{s_x}{\sqrt{n}} \right\} \equiv \left\{ \bar{x} - \mu_0 > t_{\alpha, n-1} \frac{\sigma_x}{\sqrt{n-1}} \right\}$$

$$\text{Región de rechazo: } R = \left\{ 5,34 - 5 > \overbrace{1,311}^{t_{0,10,29}} \frac{7,24}{\sqrt{29}} \right\} = [0,34 > 1,763]$$

Como no se verifica la región de rechazo, no existe evidencia significativa para rechazar la hipótesis H_0 con un nivel de confianza del 90%.

✓ Una forma análoga de enfocar el problema consistiría en aceptar la hipótesis H_0 cuando el estadístico experimental fuera menor o igual que el estadístico teórico, es decir

Se acepta H_0 cuando se verifica: $\overbrace{t_{\alpha_p/2} = \frac{\bar{x} - \mu_0}{\frac{\sigma_x}{\sqrt{n-1}}}}^{\text{estadístico experimental}} \leq \overbrace{t_{\alpha, n-1}}^{\text{estadístico teórico}}$

En este caso, $t_{\alpha_p/2} = \frac{5,34 - 5}{7,24/\sqrt{29}} = 0,2537 < 1,311 = t_{0,10;29} \rightarrow$ Se acepta la hipótesis H_0

d) Potencia = $1 - \beta = P[\text{Rechazar } H_0 \mid H_0 \text{ Falsa}] = P[\text{Rechazar } H_0 \mid H_1 \text{ Cierta}]$

La hipótesis alternativa H_1 es compuesta (existen infinitos valores tal que $\mu > 5$), se construye una función, es decir:

$H_1: \mu_1 > 5$	β	$1 - \beta$
<i>Distintos valores</i>		

$$\beta = P[\text{ETII}] = P[\text{Aceptar } H_0 \mid H_0 \text{ Falsa}] = P[\text{Rechazar } H_1 \mid H_1 \text{ Cierta}] = P[\bar{x} \leq k \mid H_1 \text{ Cierta}]$$

En el muestreo de una población normal $N(\mu, \sigma)$ con varianza poblacional desconocida, cuando el tamaño muestral es grande $n = 300 > 30$, la distribución de la media muestral se aproxima a una distribución normal.

En esta línea, bajo la hipótesis alternativa $\bar{x} \sim N\left(\mu_1, \frac{\sigma_x}{\sqrt{n-1}}\right) \equiv N\left(\mu_1, \frac{7,24}{\sqrt{300-1}}\right)$

$$\begin{aligned} \beta &= P[\bar{x} \leq k \mid H_1 \text{ Cierta}] = P\left[\bar{x} \leq 6,76 \mid N\left(\mu_1, \frac{7,24}{\sqrt{300-1}}\right)\right] = P\left[\frac{\bar{x} - \mu_1}{7,24 / \sqrt{299}} \leq \frac{6,76 - \mu_1}{7,24 / \sqrt{299}}\right] = \\ &= P\left[z \leq \frac{6,76 - \mu_1}{7,24 / \sqrt{299}}\right] \end{aligned}$$

En consecuencia:

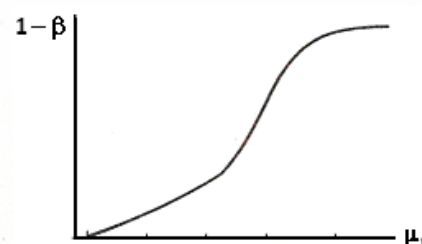
$$\Rightarrow H_1: \mu_1 = 5,5 \rightarrow \beta = P\left[z \leq \frac{6,76 - 5,5}{7,24 / \sqrt{299}}\right] = P(z \leq 3) = 0,9986$$

$$\Rightarrow H_1: \mu_1 = 6 \rightarrow \beta = P\left[z \leq \frac{6,76 - 6}{7,24 / \sqrt{299}}\right] = P(z \leq 1,82) = 0,9656$$

$$\Rightarrow H_1: \mu_1 = 7 \rightarrow \beta = P\left[z \leq \frac{6,76 - 7}{7,24 / \sqrt{299}}\right] = P(z \leq -0,57) = 0,2843$$

$$\Rightarrow H_1: \mu_1 = 8 \rightarrow \beta = P\left[z \leq \frac{6,76 - 8}{7,24 / \sqrt{299}}\right] = P(z \leq -2,96) = 0,00154$$

$H_1: \mu_1 > 5$	z	$1 - \beta$
5,5	0,9986	0,0014
6	0,9656	0,0034
7	0,2843	0,7157
8	0,00154	0,9985



A medida que se alejan las hipótesis H_0 y H_1 aumenta la potencia del contraste.

Es decir, cuanto más alejadas se encuentren las hipótesis para contrastar, mayor será la probabilidad de que se rechace H_0 cuando sea falsa, algo deseable (constante α y k).

125. El número de averías de un determinado tipo de avión se considera una variable aleatoria con distribución de Poisson de media 2 averías al mes. El equipo de mantenimiento intenta reducir esta media incorporando algunas mejoras. Para comprobar si con estas medidas se reduce el número medio de averías, se decide observar el número medio de averías en los 25 meses siguientes a la introducción de las mejoras. Si el número medio de averías en esos 25 meses fue de 1,5.

¿Qué decisión debe adoptar el servicio técnico a un nivel de significación del 1%?

¿Y si el servicio técnico relaja su nivel de exigencia al 85% de confianza?. ¿Cambiaría su decisión?.

Solución:

Sea la variable aleatoria $X =$ " Número de averías al mes " $X \sim P(\lambda = 2) \xrightarrow{\lambda > 10} N(2, \sqrt{2})$

Es un contraste unilateral, donde se plantean las hipótesis: $H_0: \lambda \geq 2$ $H_1: \lambda < 2$

En el muestreo, bajo la hipótesis nula H_0 $\bar{x} \sim N\left[\lambda, \sqrt{\frac{\lambda}{n}}\right] = N\left[2, \sqrt{\frac{2}{25}}\right] \equiv N(2, 0,2828)$

En la muestra de tamaño $n = 25$ meses $\hat{\lambda} = \bar{x} = 1,5$ averías.

Regla de decisión $\begin{cases} \text{Si } \bar{x} \geq k \text{ se acepta } H \\ \text{Si } \bar{x} < k \text{ se rechaza } H_0 \end{cases}$

Se determina el valor de k considerando el nivel de significación $\alpha = 0,01$:

$$\alpha = P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P[\bar{x} < k \mid \lambda = 2] = P\left[\frac{\bar{x} - 2}{0,2828} < \frac{k - 2}{0,2828}\right] = P\left[z < \frac{k - 2}{0,2828}\right] = 0,01$$

$$P\left[z \geq \frac{2 - k}{0,2828}\right] = 0,01 \rightarrow \frac{2 - k}{0,2828} = 2,32 \rightarrow k = 1,3439$$

Se advierte que $1,5 > 1,3439$, cae en la región de aceptación, con lo cual se acepta la hipótesis nula. Es decir, con nivel de significación del 1% se afirma que las averías mensuales se mantienen siendo 2, y en consecuencia, las mejoras no son operativas.

$$\text{Si } \alpha = 0,15 \text{ se replantean los cálculos: } \frac{2 - k}{0,2828} = 1,04 = z_{0,15} \rightarrow k = 1,71$$

Como $1,5 < 1,71$ la media muestral observada cae en la región de rechazo, con lo que no se acepta la hipótesis nula, y se concluye que las mejoras son operativas.

126. En los días previos a unas elecciones municipales, el candidato de un partido político está convencido de obtener el 60% de los votos electorales. No obstante, su partido encarga una encuesta entre 100 votantes potenciales, resultando que el 52% de ellos dijeron tener intención de votar a dicho candidato. Con un nivel de significación del 5%, se pide contrastar:

- $H_0: p = 0,60$ frente a $H_1: p = 0,50$
- $H_0: p = 0,60$ frente a $H_1: p \neq 0,60$
- Potencia del contraste efectuado en el apartado (a)

Solución:

a) Sea la variable $X =$ 'Porcentaje de votos al candidato'

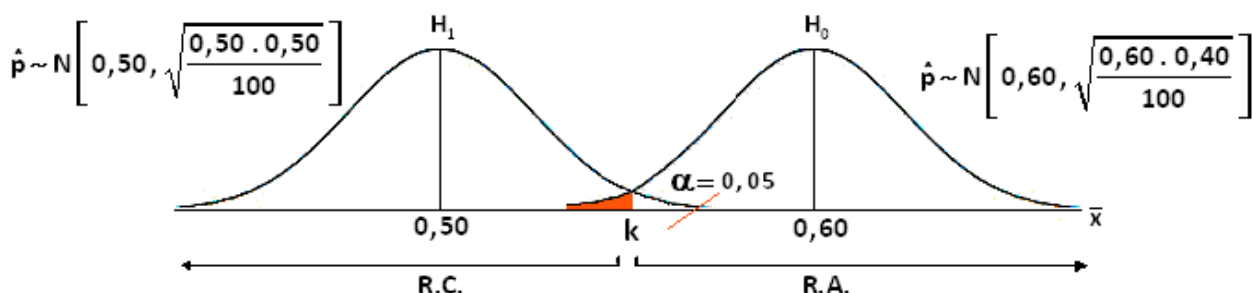
Al ser el tamaño suficientemente grande $n = 100$ y estar definido p como suma de variables aleatorias independientes, según una distribución de Bernoulli $B(1, p)$, se puede aproximar a la normal $N[np, \sqrt{npq}]$.

La distribución en el muestreo $\hat{p} = \frac{1}{100} \sum_{i=1}^{100} x_i$ siendo $x_i = \begin{cases} 1 & \text{vota Si} \\ 0 & \text{vota No} \end{cases} \quad \hat{p} \sim N\left[p, \sqrt{\frac{pq}{n}}\right]$

Contraste unilateral a la izquierda (cola a la izquierda): $H_0: p = 0,60 \quad H_1: p = 0,50$

Bajo la hipótesis nula: $\hat{p} \sim N\left[0,60, \sqrt{\frac{0,60 \cdot (1-0,60)}{100}}\right] \equiv N(0,60, 0,049)$

Regla de decisión: $\begin{cases} \hat{p} \geq k & \text{Se acepta } H_0 \text{ (R.A.)} \\ \hat{p} < k & \text{Se rechaza } H_0 \text{ (R.C.)} \end{cases}$



El valor crítico k se determina con el nivel de significación $\alpha = 0,05$:

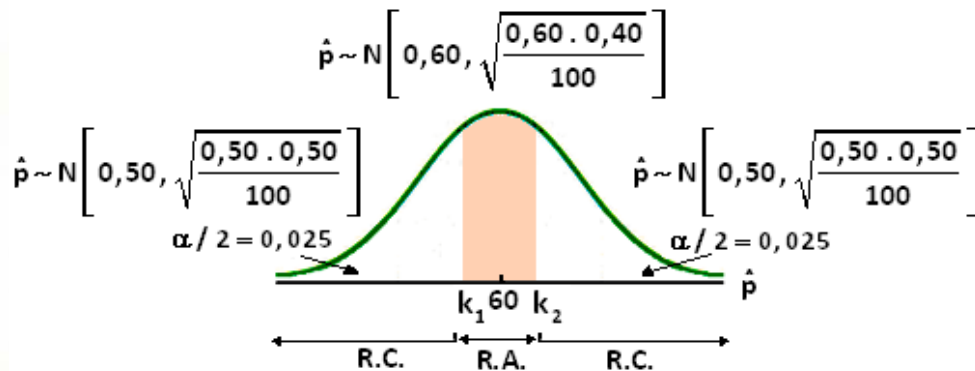
$$\alpha = P[ET I] = P[\text{Rechazar } H_0 / H_0 \text{ Cierta}] = P[\hat{p} < k \mid H_0: p = 0,6] = P\left[\frac{\hat{p} - 0,60}{0,049} < \frac{k - 0,60}{0,049}\right] =$$

$$= P\left[\frac{\hat{p} - 0,60}{0,049} \geq \frac{0,60 - k}{0,049}\right] = P\left[z \geq \frac{0,60 - k}{0,049}\right] = 0,05 \rightarrow \frac{0,60 - k}{0,049} = 1,645 \rightarrow k = 0,5193$$

La proporción muestral observada $\hat{p} = \bar{x} = 52/100 = 0,52 > 0,5193$ se encuentra dentro de la región de aceptación, es decir, se acepta la hipótesis nula con un nivel de confianza del 95%.

b) En este caso se trata de un contraste bilateral (dos colas): $H_0: p = 0,60$ $H_1: p \neq 0,60$

Regla de decisión: $\begin{cases} |\hat{p}| \leq k & \text{Se acepta } H_0 \text{ (R.A.)} \\ |\hat{p}| > k & \text{Se rechaza } H_0 \text{ (R.C.)} \end{cases} \rightarrow \begin{cases} k_1 \leq \hat{p} \leq k_2 & \text{Se acepta } H_0 \\ \hat{p} < k_1 \cup \hat{p} > k_2 & \text{Se rechaza } H_0 \end{cases}$



Los valores críticos k se determinan por el nivel de significación $\alpha/2 = 0,05/2 = 0,025$:

$$\begin{aligned} \alpha &= P[ET I] = P[\text{Rechazar } H_0 / H_0 \text{ Cierta}] = P[|\hat{p}| > k \mid H_0: p = 0,6] = P\left[\left|\frac{\hat{p} - 0,60}{0,049}\right| > \frac{k - 0,60}{0,049}\right] = \\ &= P\left[|z| > \frac{k - 0,60}{0,049}\right] = P\left[z < \frac{k_1 - 0,60}{0,049}\right] + P\left[z > \frac{k_2 - 0,60}{0,049}\right] = 0,025 + 0,025 = 0,05 \end{aligned}$$

$$P\left[z < \frac{k_1 - 0,60}{0,049}\right] = P\left[z > \frac{-k_1 + 0,60}{0,049}\right] = 0,025 \rightarrow \frac{-k_1 + 0,60}{0,049} = 1,96 \rightarrow k_1 = 0,504$$

$$P\left[z > \frac{k_2 - 0,60}{0,049}\right] = 0,025 \rightarrow \frac{k_2 - 0,60}{0,049} = 1,96 \rightarrow k_2 = 0,696$$

La proporción muestral observada $\hat{p} = \bar{x} = 52/100 = 0,52$ se encuentra dentro de la región de aceptación $0,504 < \hat{p} = 0,52 < 0,696$, concluyendo que se acepta la hipótesis nula.

✓ Otra forma de enfocar el problema consistiría en aceptar la hipótesis H_0 cuando el estadístico de contraste fuera menor o igual que el estadístico teórico, es decir:

Se acepta H_0 cuando se verifica:

$$\overbrace{z = \frac{|\hat{p} - p_0|}{\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}}}^{\text{estadístico contraste}} \leq \overbrace{z_{\alpha/2}}^{\text{estadístico teórico}}$$

Siendo $z = \frac{|0,52 - 0,60|}{\sqrt{\frac{0,52 \cdot 0,48}{100}}} = 1,6028 < 1,96 = z_{0,025}$ se acepta la hipótesis nula.

✓ Se puede elaborar un intervalo de confianza para el parámetro poblacional p de una distribución binomial y observar si el valor $p = 0,6$ de la hipótesis nula se encuentra dentro, con lo que se aceptaría la hipótesis.

$$I(p) = \left[\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}} \right] = \left[0,52 \pm 1,96 \cdot \sqrt{\frac{0,52 \cdot 0,48}{100}} \right] = [0,4221, 0,6179]$$

$0,6 \in [0,4221, 0,6179] \rightarrow$ Se acepta H_0

c) Potencia = $1 - \beta = P[\text{Rechazar } H_0 \mid H_0 \text{ Falsa}] = P[\text{Rechazar } H_0 \mid H_1 \text{ Cierta}]$

Regla de decisión: $\begin{cases} \hat{p} \geq 0,5193 & \text{Se acepta } H_0 \\ \hat{p} < 0,5193 & \text{Se rechaza } H_0 \end{cases}$

Con la hipótesis alternativa $H_1: p = 0,50 \quad \hat{p} \sim N\left[0,5, \sqrt{\frac{0,50 \cdot 0,50}{100}}\right] = N(0,5, 0,05)$

$$\begin{aligned} \text{Potencia} &= P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P[\hat{p} < 0,5193 \mid N(0,5, 0,05)] = P\left[\frac{\hat{p} - 0,5}{0,05} < \frac{0,5193 - 0,5}{0,05}\right] = \\ &= P(z < 0,386) = 1 - P(z \geq 0,386) = 1 - 0,34978 = 0,6502 \end{aligned}$$

Abscisas	Áreas				
0,38 - 0,39	↔ 0,3520 - 0,3483	-0,01	↔ 0,0037	→	$x = 0,3483 + \frac{0,004 \cdot 0,0037}{0,01} = 0,34978$
0,386 - 0,39	↔ x - 0,3483	-0,004	↔ x - 0,3483		

127. El dueño de los cines CINEFILÓN considera que, dado el aforo de la sala, una afluencia diaria a la misma del 85% sería óptima, en el sentido de que los clientes se sientan cómodos y para que a la vez no haya pérdidas económicas. Durante un período de tiempo, se analiza la afluencia a los cines, observándose que, en media, se ocupan 171 de las 200 butacas.

a) ¿Con qué confianza podrá afirmar el dueño de CINEFILÓN que la asistencia a sus cines es óptima?

b) ¿Qué pasaría si el dueño quisiera estar más seguro de su decisión, y ampliar el nivel de confianza al 99%?

Nota: La regla de la decisión adoptada es que si hay una desviación inferior al 5% de la cantidad óptima, se aceptaría la hipótesis de que la afluencia es, efectivamente, óptima.

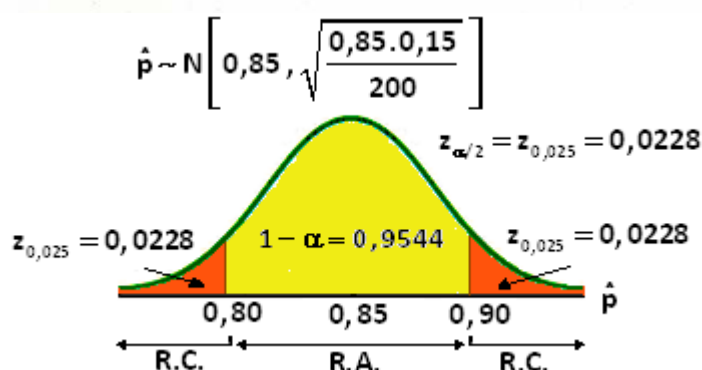
Solución:

La variable aleatoria X = 'Porcentaje asistencia diaria a los cines CINEFILÓN' $X \sim B(1, p)$.

En el contraste bilateral (dos colas) se establecen las hipótesis: $H_0: p_0 = 0,85$ $H_1: p_1 \neq 0,85$

- En el muestreo, la proporción que acude a los cines (200 butacas) en un día $\hat{p} = \frac{X}{n}$, bajo la hipótesis nula, por el TCL (Teorema Central Límite), sigue una distribución

$$N\left[p, \sqrt{\frac{p \cdot q}{n}}\right] \equiv N\left[0,85, \sqrt{\frac{0,85 \cdot 0,15}{200}}\right] = N(0,85, 0,025)$$



Considerando que una desviación inferior al 5% de la cantidad óptima significa que $0,80 < \hat{p} < 0,90$

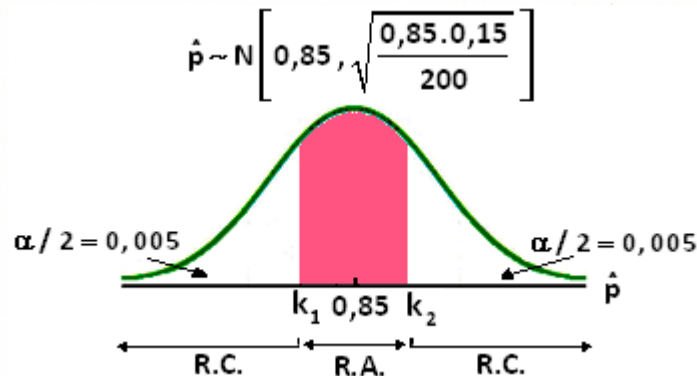
El nivel de confianza $(1 - \alpha)$ se determina a partir de los valores críticos:

$$\begin{aligned} \alpha &= P[ETI] = P[\text{Rechazar } H_0 / H_0 \text{ Cierta}] = P\left[\left(\frac{0,80 - 0,85}{0,025} > z\right) \cup \left(z > \frac{0,90 - 0,85}{0,025}\right)\right] \\ &= P[-2 > z] + P[z > 2] = 2 \cdot 0,0228 = 0,0456 \end{aligned}$$

Nivel de confianza: $1 - \alpha = 1 - 0,0456 = 0,9544$

b) Sí el nivel de confianza $(1 - \alpha) = 99\% \rightarrow \alpha = 1\%$ mayor exigencia

Regla de decisión: $\begin{cases} |\hat{p}| \leq k & \text{Se acepta } H_0 \text{ (R.A.)} \\ |\hat{p}| > k & \text{Se rechaza } H_0 \text{ (R.C.)} \end{cases} \rightarrow \begin{cases} k_1 \leq \hat{p} \leq k_2 & \text{Se acepta } H_0 \\ \hat{p} < k_1 \text{ o } \hat{p} > k_2 & \text{Se rechaza } H_0 \end{cases}$



Los valores críticos k se determina a partir del nivel de significación $\alpha = 0,01$

$$\begin{aligned} \alpha &= P[ET I] = P[\text{Rechazar } H_0 / H_0 \text{ Cierta}] = P[|\hat{p}| < k \mid N(0,85, 0,025)] = P\left[\left|\frac{\hat{p} - 0,85}{0,025}\right| < \frac{k - 0,85}{0,025}\right] = \\ &= P\left[\left|z\right| < \frac{k - 0,85}{0,025}\right] = P\left[\left|z\right| \geq \frac{0,85 - k}{0,025}\right] = P\left[\left(z \geq \frac{0,85 - k_1}{0,025}\right) \cup \left(z \geq -\frac{0,85 - k_2}{0,025}\right)\right] = 0,005 + 0,005 = 0,01 \end{aligned}$$

$$\frac{0,85 - k_1}{0,025} = 2,57 = z_{0,005} \rightarrow k_1 = 0,78575 \qquad -\frac{0,85 - k_2}{0,025} = 2,57 = z_{0,005} \rightarrow k_2 = 0,91425$$

$$\text{O bien } \left|\frac{0,85 - k}{0,025}\right| = 2,57 = z_{0,005} \rightarrow \begin{aligned} \frac{0,85 - k_1}{0,025} &= 2,57 \rightarrow k_1 = 0,78575 \\ \frac{0,85 - k_2}{0,025} &= -2,57 \rightarrow k_2 = 0,91425 \end{aligned}$$

Comparando el estadístico muestral (evidencia empírica) $\hat{p} = \frac{171}{200} = 0,855$ se observa que cae dentro de la región de aceptación $0,78575 < \hat{p} = 0,855 < 0,91425$ concluyendo que se acepta la hipótesis nula siendo óptima la afluencia a los cines CINEFILÓN.

128. En una población $N(10, \sigma)$ se desea contrastar la hipótesis nula $H_0: \sigma^2 = 3$ frente a la hipótesis alternativa $H_1: \sigma^2 \neq 3$ con un nivel de significación del 5%. En una muestra aleatoria simple de tamaño 4 se obtuvieron los resultados: 10 8 12 14

Solución:

Hipótesis compuesta σ^2 : $H_0: \sigma^2 = 3$ $H_1: \sigma^2 \neq 3$

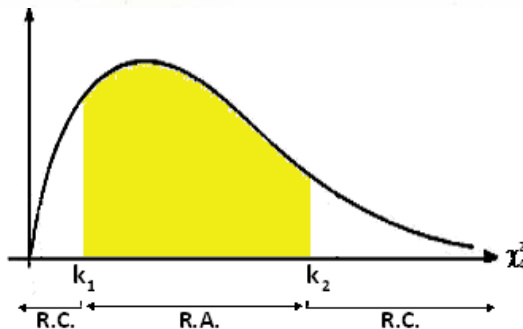
Como la hipótesis alternativa es $\sigma^2 \neq 3$ en la decisión deberán ser válidos los valores de σ^2 tanto mayores o menores que 3, por lo que el contraste debe ser bilateral o de dos colas.

Bajo la hipótesis nula, las variables aleatorias independientes $X_i \sim N(10, \sqrt{3})$ con lo que

$$\frac{X_i - 10}{\sqrt{3}} \sim N(0, 1) \rightarrow \sum_{i=1}^4 \left(\frac{X_i - 10}{\sqrt{3}} \right)^2 = \chi_4^2$$

Adviértase que la suma de 4 variables aleatorias $N(0, 1)$ independientes entre sí es una distribución Chi-cuadrado con 4 grados de libertad.

Regla decisión: $\left| \sum_{i=1}^4 \left(\frac{X_i - 10}{\sqrt{3}} \right)^2 \right| \leq k \rightarrow \text{Se acepta } H_0$ $\left| \sum_{i=1}^4 \left(\frac{X_i - 10}{\sqrt{3}} \right)^2 = \chi_4^2 \right| > k \rightarrow \text{Se rechaza } H_0$



Los valores críticos k se determina a partir del nivel de significación $\alpha = 0,05$

$$\begin{aligned} \alpha &= P[\text{Rechazar } H_0 / H_0 \text{ Cierta}] = P\left[\left|\sum_{i=1}^4 \left(\frac{X_i - 10}{\sqrt{3}}\right)^2\right| > k / H_0 \text{ Cierta}\right] = P[|\chi_4^2| > k] = \\ &= P[(\chi_4^2 < k_1) \cup (\chi_4^2 > k_2)] = P[\chi_4^2 < k_1] + P[\chi_4^2 > k_2] = \alpha_1 + \alpha_2 = 0,025 + 0,025 = 0,05 \end{aligned}$$

$$P[\chi_4^2 < k_1] = 1 - P[\chi_4^2 > k_1] = 1 - 0,025 = 0,975 \rightarrow k_1 = \chi_{0,975, 4}^2 = 0,484$$

$$P[\chi_4^2 > k_2] = 0,025 \rightarrow k_2 = \chi_{0,025, 4}^2 = 11,143$$

Región de aceptación $[0,484, 11,143]$

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

La varianza muestral observada $\sigma_x^2 = \frac{\sum_{i=1}^4 (x_i - \bar{x})^2}{4} = \frac{\sum_{i=1}^4 (x_i - 11)^2}{4} = \frac{20}{4} = 5$ se encuentra dentro de la región de aceptación $0,484 < \sigma_x^2 = 5 < 11,143$, aceptando la hipótesis nula con varianza poblacional 3.

✓ Se puede plantear que bajo la hipótesis nula la distribución en el muestreo:

$$\chi_{n-1}^2 = \frac{\text{Varianza muestral}}{\text{Varianza teórica}} = \frac{\sigma_x^2}{(\sigma / \sqrt{n})^2} \rightarrow \sigma_x^2 = \frac{\sigma^2 \cdot \chi_{n-1}^2}{n} \rightarrow \sigma_x^2 = \frac{3 \cdot \chi_3^2}{4}$$

Regla de decisión: $\begin{cases} \sigma_x^2 \leq k & \text{se acepta } H_0 \\ \sigma_x^2 > k & \text{se rechaza } H_0 \end{cases}$

✓ Análogamente, se podría haber resuelto considerando la región de rechazo:

$$R = \left[\frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma_0^2} \notin [\chi_{1-\alpha/2, n-1}^2; \chi_{\alpha/2, n-1}^2] \right]$$

El valor del estadístico muestral, bajo la hipótesis nula, resulta: $\chi_4^2 = \frac{\sum_{i=1}^4 (x_i - 10)^2}{3} = \frac{24}{3} = 8$.

Los valores tabulares son: $\chi_{\alpha/2, n-1}^2 = \chi_{0,025, 3}^2 = 9,348$ $\chi_{1-\alpha/2, n-1}^2 = \chi_{0,975, 3}^2 = 0,216$

Región de aceptación de la hipótesis nula es el intervalo $[0,216, 9,348]$

El valor muestral del estadístico $\sigma_x^2 = \frac{\sum_{i=1}^4 (x_i - 10)^2}{3} = 8$ no verifica la región de rechazo, se concluye que la varianza poblacional de la distribución es 3.

✓ Se podría haber elaborado un intervalo de confianza para la varianza σ^2 de una población normal, verificando si se acepta la hipótesis nula al encontrarse la varianza $\sigma^2 = 3$ cubierta en el intervalo.

$$I(\sigma^2) = \left[\frac{(n-1) \cdot s_x^2}{\chi_{\alpha/2, n-1}^2}, \frac{(n-1) \cdot s_x^2}{\chi_{1-\alpha/2, n-1}^2} \right]$$

$$(n-1) \cdot s_x^2 = n \cdot \sigma_x^2 = 4 \cdot 5 = 20$$

$$I(\sigma^2) = \left[\frac{20}{9,348}, \frac{20}{0,216} \right] = [2,139, 92,592]$$

Como $3 \in [2,139, 92,592] \rightarrow$ Se acepta H_0

129. En una población $N(\mu, 5)$ se formula la hipótesis nula $H_0: \mu = 18$ frente a la hipótesis alternativa $H_1: \mu \neq 18$, con un nivel de significación $\alpha = 0,01$, con una muestra de tamaño 10 con los datos recogidos: 16 12 15 16 20 25 14 18 17 22

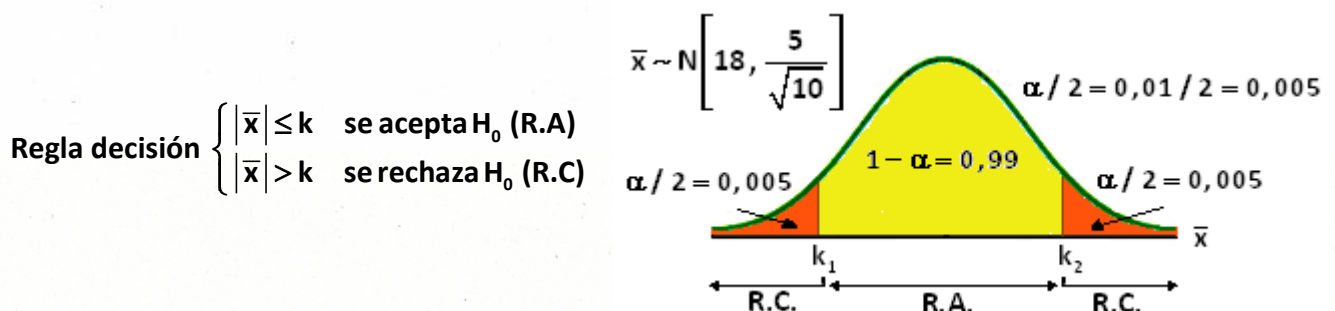
- Contrastar la hipótesis nula
- Realizar el contraste con el p_valor
- Realizar el contraste con métodos afines al p_valor

Solución:

a) Se establecen las hipótesis: $H_0: \mu = 18$ $H_1: \mu \neq 18$

Como la hipótesis alternativa es $\mu \neq 18$ en la decisión deberán ser válidos valores de μ tanto mayores o menores que 18, por lo cual el contraste debe ser bilateral o de dos colas.

La muestra bajo la hipótesis nula sigue una distribución $N\left[18, \frac{5}{\sqrt{10}}\right] = N(18, 1,58)$



Los valores críticos k se calculan mediante el nivel de significación $\alpha = 0,01$

$$\alpha = P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P[|\bar{x}| > k \mid N(18, 1,58)] = P[(\bar{x} < k_1) \cup (\bar{x} > k_2)] = \\ = P(\bar{x} < k_1) + P(\bar{x} > k_2) = 0,005 + 0,005 = 0,01$$

$$P(\bar{x} < k_1) = P\left[\frac{\bar{x} - 18}{1,58} < \frac{k_1 - 18}{1,58}\right] = P\left[z < \frac{k_1 - 18}{1,58}\right] = P\left[z > -\frac{k_1 - 18}{1,58}\right] = 0,005 \\ -\frac{k_1 - 18}{1,58} = 2,575 \rightarrow k_1 = 13,9315$$

$$P(\bar{x} > k_2) = P\left[\frac{\bar{x} - 18}{1,58} > \frac{k_2 - 18}{1,58}\right] = P\left[z > \frac{k_2 - 18}{1,58}\right] = 0,005 \\ \frac{k_2 - 18}{1,58} = 2,575 \mapsto k_2 = 22,0685$$

En consecuencia, la región de aceptación: $13,9315 < \bar{x} < 22,0685$

La media muestral observada (evidencia empírica) $\bar{x} = \frac{\sum_{i=1}^{10} x_i}{10} = 17,5$ se encuentra dentro de la región de aceptación, afirmando con un nivel de confianza del 99%, que se verifica la hipótesis nula.

b) El $p_{\text{valor}} (\alpha_p)$ es el nivel de significación más pequeño posible que puede escogerse para el que todavía se aceptaría la hipótesis alternativa con las observaciones actuales.

$$\alpha_p = p_{\text{valor}} = P[\text{Rechazar el estadístico muestral observado} / H_0 \text{ es Cierta}]$$

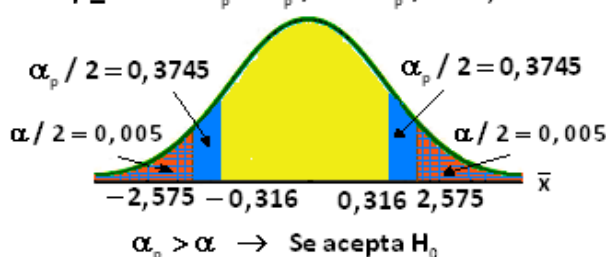
$$\frac{|\bar{x} - \mu_0|}{\sigma / \sqrt{n}}$$

$$\alpha_p = P[|\bar{x}| > 17,5 \mid N(18, 1,58)] = P\left[\left|\frac{\bar{x} - 18}{1,58}\right| > \frac{17,5 - 18}{1,58}\right] = P[|z| > -0,316] = \\ = P[|z| > 0,316] = P[z < -0,316] + P[z > 0,316] = 2P[z > 0,316] = 2(0,3745) = 0,749$$

Tabla $N(0,1)$ $\frac{0,31 - 0,32}{0,3783 - 0,3745} = \frac{0,316 - 0,32}{x - 0,3745} \rightarrow x = 0,3745$

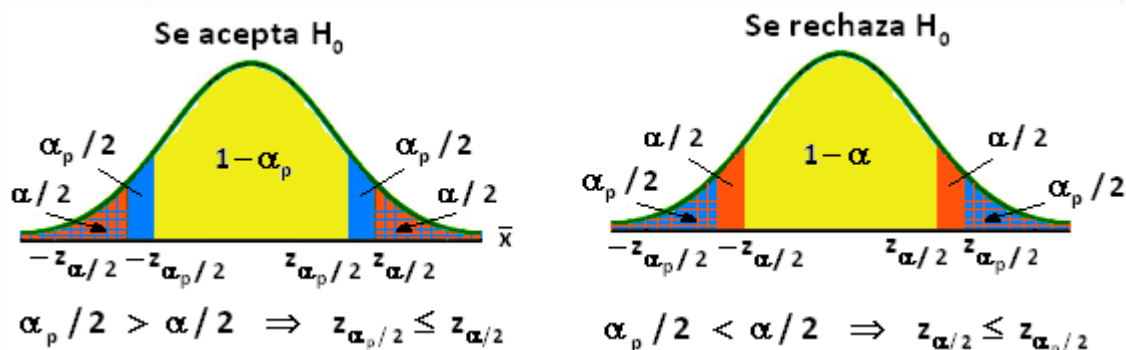
$$p_{\text{valor}} = \alpha_p = \alpha_p / 2 + \alpha_p / 2 = 0,749$$

Como $\alpha_p = 0,749 > 0,01 = \alpha$ se acepta la hipótesis nula.



c) Métodos consecuencia del p_valor:

⇒ El $p_valor = \alpha_p$ es la abscisa de la distribución normal $N(0,1)$ que origina el estadístico de contraste (estadístico empírico u observado), de modo que en un contraste bilateral (o de dos colas) para la media poblacional se acepta la hipótesis nula cuando $z_{\alpha_p/2} \leq z_{\alpha/2}$



$$z_{\alpha_p/2} = \frac{|\bar{x} - \mu_0|}{\sigma / \sqrt{n}} \leq z_{\alpha/2} \rightarrow \alpha_p/2 > \alpha/2 \Rightarrow z_{\alpha_p/2} \leq z_{\alpha/2}$$

$$z_{\alpha_p/2} = \frac{|\bar{x} - \mu_0|}{\sigma / \sqrt{n}} = \frac{|17,5 - 18|}{1,58} = 0,316 < 2,575 = z_{0,005} \rightarrow \text{Se acepta } H_0$$

$$\text{Señalar que } \alpha_p/2 = P[z_{\alpha_p/2} > 0,316] = 0,3745 \rightarrow \alpha_p = 0,749$$

$$\Rightarrow \text{Región de Rechazo: } R = \left\{ |\bar{x} - \mu_0| > z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right\}$$

$$R = \left\{ |17,5 - 18| > 2,575 \cdot \frac{5}{\sqrt{10}} \right\} = \{0,5 > 4,07\} \rightarrow$$

No se verifica la región de rechazo, aceptando la hipótesis nula $\mu = 18$

⇒ Observando si la media de la hipótesis nula $\mu = 18$ se encuentra dentro del intervalo de confianza para la media poblacional μ : $I.C(\mu) = \left[\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$

$$I.C(\mu) = \left[17,5 \pm 2,575 \frac{5}{\sqrt{10}} \right] = [13,43, 21,57] \rightarrow \mu = 18 \in [13,43, 21,57]$$

Se acepta la hipótesis nula.

130. Una empresa de neumáticos afirma que una nueva gama sigue una ley normal y en promedio duran más de 28.000 km. Las pruebas con 64 neumáticos dan como resultado una duración media de 27.800 km, con una desviación estándar de 1.000 km.

- a) Si se usa un nivel de significación del 5%, comprobar si hay evidencia suficiente para rechazar la afirmación de la empresa.
b) ¿Cuál es el p_valor?

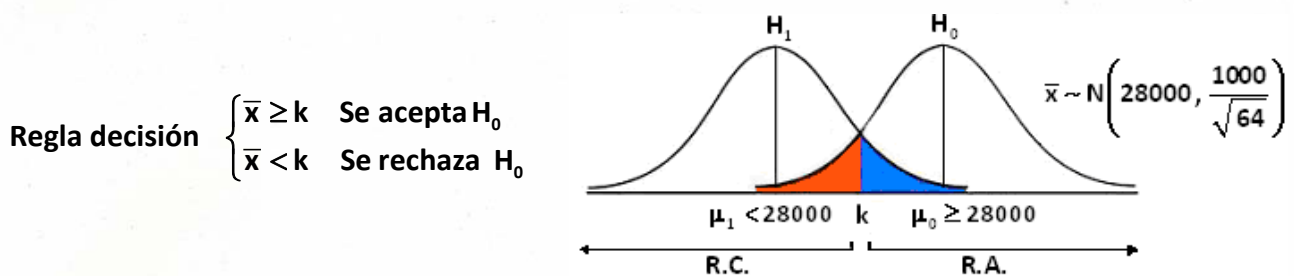
Solución:

a) Sobre la población de los neumáticos se define la variable aleatoria X = "Duración en kilómetros", donde $X \sim N(28000, \sigma)$

Hipótesis sobre la media poblacional μ con varianza σ^2 desconocida:

$$H_0: \mu_0 \geq 28000 \quad H_1: \mu_1 < 28000$$

Se trata de un contraste unilateral por la izquierda (cola por la izquierda)



En el muestreo bajo la hipótesis nula: $\bar{x} \sim N\left(28000, \frac{1000}{\sqrt{64}}\right) \equiv N(28000, 125)$

El valor crítico k bajo la hipótesis nula se determina con el nivel de significación $\alpha = 0,05$:

$$\begin{aligned} \alpha &= P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P[\bar{x} < k \mid N(28000, 125)] = P\left(\frac{\bar{x} - 28000}{125} < \frac{k - 28000}{125}\right) = \\ &= P\left[z < \frac{k - 28000}{125}\right] = P\left[z > \frac{28000 - k}{125}\right] = 0,05 \end{aligned}$$

$$\frac{28000 - k}{125} = 1,645 \rightarrow k = 28000 - 125 \times 1,645 = 27794,375$$

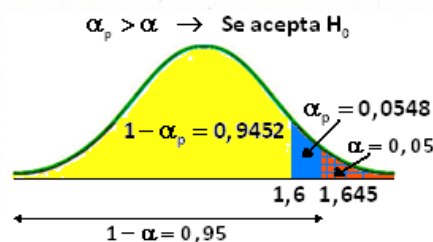
Siendo $\bar{x} = 27800 > 27794,375$ se acepta la hipótesis nula, por tanto, es buena la afirmación de la empresa con un nivel de confianza del 95%.

b) El p_valor (α_p) es el nivel de significación más pequeño posible que puede escogerse para el que todavía se aceptaría la hipótesis alternativa con las observaciones actuales.

$\alpha_p = p_valor = P[\text{Rechazar el estadístico muestral observado} / H_0 \text{ es Cierta}]$

$$\begin{aligned}\alpha_p &= P[\bar{x} < 27800 \mid H_0 \text{ Cierta}] = P[\bar{x} < 27800 \mid N(28000, 125)] = P\left[\frac{\bar{x} - 28000}{125} < \frac{27800 - 28000}{125}\right] = \\ &= P[z < -1,6] = P[z > 1,6] = 0,0548\end{aligned}$$

Siendo $\alpha_p = 0,0548 > \alpha = 0,05$ se acepta la hipótesis nula. Es decir, con una fiabilidad del 95% se acepta que la duración media de los neumáticos es de 28.000 km.



131. Una línea de producción sigue una distribución normal, funciona con un peso de llenado de 16 gramos por envase. El exceso o defecto de peso en el llenado presenta graves problemas, debiendo parar la línea de producción. Un inspector de calidad toma una muestra de 25 artículos, y de acuerdo con los resultados obtenidos, ¿qué decisión debe tomar?.

- a) Para un nivel de significación de 0,05, en la muestra se obtiene $\bar{x} = 16,32$ gramos y una desviación $s_x = 0,8$ gramos.
b) ¿Cuál es el p-valor?.

Solución:

- a) Sea la variable aleatoria $X = \text{"Peso de llenado"}$, $X \sim N(16, \sigma)$

Hipótesis sobre la media poblacional μ con varianza σ^2 desconocida:

$$H_0: \mu_0 = 16 \quad H_1: \mu_1 \neq 16$$

Adviértase que como la hipótesis alternativa $H_1: \mu \neq 16$, en la decisión deberán ser válidos valores de μ tanto mayores o menores que 16, por lo que el contraste debe ser bilateral o de dos colas.

En el muestreo de una población normal $N(\mu, \sigma)$ con varianza poblacional desconocida, con un tamaño muestral pequeño $n = 16 \leq 30$, la media muestral bajo la hipótesis nula sigue una

$$\text{distribución } \bar{x} \sim t_{n-1} \left(\mu, \frac{s_x}{\sqrt{n}} \right) \equiv t_{24} \left(16, \frac{0,8}{\sqrt{25}} \right) \equiv t_{24}(16, 0,16)$$

Para realizar el contraste se compara el estadístico observado (abscisa del p_valor) con el

$$\text{estadístico teórico, aceptándose la hipótesis nula cuando } t_{\alpha_p/2} = \frac{|\bar{x} - \mu_0|}{s_x / \sqrt{n}} \leq t_{\alpha/2, n-1}$$

$$t_{\alpha_p/2} = \frac{|\bar{x} - \mu_0|}{s_x / \sqrt{n}} = \frac{|16,32 - 16|}{0,16} = 2 \leq 2,064 = t_{0,025, 24} \rightarrow \text{Se acepta } H_0$$

✓ Señalar que una forma análoga de proceder sería aceptar la hipótesis nula si no verifica la

$$\text{región de rechazo: } R = \left\{ |\bar{x} - \mu_0| > t_{\alpha/2, n-1} \frac{s_x}{\sqrt{n}} \right\}$$

Se acepta H_0 porque no se verifica $R = \{ |16,32 - 16| > 2,064 \cdot 0,16 \} = \{0,32 \not> 0,33\}$
afirmando que el peso de llenado son 16 gramos, con un nivel de significación de 0,05.

b) El $p_{\text{valor}} (\alpha_p)$ es el nivel de significación más pequeño posible que puede escogerse para el que todavía se aceptaría la hipótesis alternativa con las observaciones actuales.

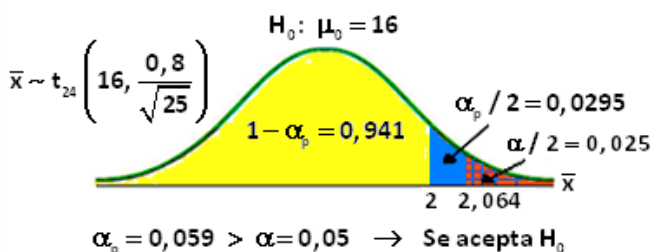
$$t_{\alpha_p/2} = \frac{|\bar{x} - \mu_0|}{s_x / \sqrt{n}} = \frac{|16,32 - 16|}{0,16} = 2$$

Al ser el contraste bilateral, α_p viene dado por la expresión $P(t_{\alpha_p/2} \geq 2) = \frac{\alpha_p}{2}$

En la tabla de la t de Student: $P(t_{24} \geq 1,711) = 0,05$ y $P(t_{24} \geq 2,064) = 0,025$, interpolando se tiene:

0,05	$\alpha_p/2$	0,025		0,05 - 0,025	$\longleftrightarrow$	1,711 - 2,064		0,025	$\longleftrightarrow$	-0,353
1,711	2	2,064		$\alpha_p/2 - 0,025$	$\longleftrightarrow$	2 - 2,064		$\alpha_p/2 - 0,025$	$\longleftrightarrow$	-0,064

$$(\alpha_p/2 - 0,025) \times 0,353 = 0,025 \times 0,064 \rightarrow \alpha_p/2 = 0,025 + \frac{0,025 \times 0,064}{0,353} = 0,0295 \rightarrow \alpha_p = 0,059$$



Siendo $\alpha_p = 0,059 > \alpha = 0,05$ se acepta la hipótesis nula.

132. El análisis laboral que la U.E ha realizado para toda Europa, señala que en España, el salario mensual de los varones, en algunos sectores económicos, supera en más de 100 euros el salario de las mujeres que desempeñan las mismas tareas.

El Ministerio de Trabajo español decide considerar el salario mensual como una variable aleatoria normalmente distribuida con desviación típica de 39,6 euros para los trabajadores masculinos y de 36 euros para las trabajadoras de dichos sectores, siendo el salario de cada población independiente del de la otra. Para tratar de verificar lo publicado, se elige una muestra aleatoria simple de 500 trabajadores y de 700 trabajadoras, obteniéndose unos salarios medios mensuales de 1.500 y 1.370 euros respectivamente.

¿Están fundamentadas las conclusiones de la U.E al 1% de significación?

Solución:

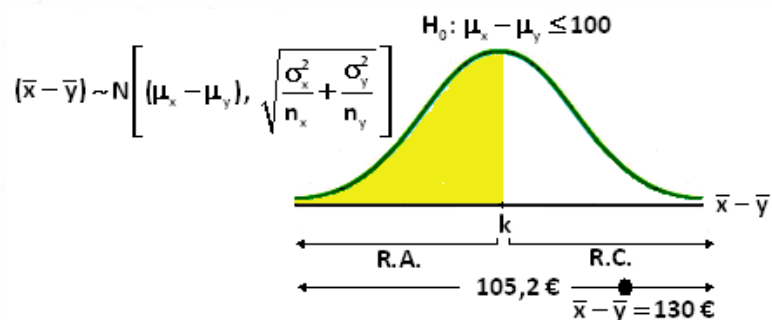
Sean las variables aleatorias, respectivamente, $X =$ 'Salario mensual de los varones' e $Y =$ 'Salario mensual de las mujeres', donde $X \sim N(\mu_x, 39,6)$ e $Y \sim N(\mu_y, 36)$

En las muestras se obtuvieron los resultados: $\begin{cases} \bar{x} = 1.500 \text{ €} & n_x = 500 \\ \bar{y} = 1.370 \text{ €} & n_y = 700 \end{cases}$

En el contraste se establecen las hipótesis: $H_0: \mu_x - \mu_y \leq 100$ $H_1: \mu_x - \mu_y > 100$

La regla de decisión será: $\begin{cases} (\bar{x} - \bar{y}) \leq k & \text{se acepta } H_0 \text{ (R.A)} \\ (\bar{x} - \bar{y}) > k & \text{se rechaza } H_0 \text{ (R.C)} \end{cases}$

La diferencia de medias muestrales $(\bar{x} - \bar{y})$, siendo las varianzas muestrales conocidas, bajo la hipótesis nula, sigue una distribución:



$$N\left(\mu_x - \mu_y, \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}\right) \equiv N\left[100, \sqrt{\frac{39,6^2}{500} + \frac{36^2}{700}}\right] = N(100, 2,23)$$

El valor crítico k se determina mediante el nivel de significación $\alpha = 0,01$:

$$\alpha = P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P[\bar{x} - \bar{y} > k \mid H_0: \mu_x - \mu_y = 100] = P\left[z > \frac{k - 100}{2,23}\right] = 0,01$$

$$\frac{k - 100}{2,23} = 2,32 = z_{0,01} \rightarrow k = 105,2 \text{ €}$$

La evidencia empírica $(\bar{x} - \bar{y}) = 1.500 - 1.370 = 130 \text{ €}$.

Se advierte que $(\bar{x} - \bar{y}) = 130 > 105,2$, esto es, la diferencia de medias muestrales cae en la región de rechazo, con lo cual se rechaza la hipótesis nula, con un nivel de significación del 1%.

Afirmando que con el mismo trabajo, las diferencias salariales entre hombres y mujeres en algunos sectores económicos españoles, son superiores a 100 euros mensuales.

133. Con un nivel de significación del 4,72%, se desea contrastar la hipótesis nula de igualdad de medias de dos poblaciones $N(\mu_1, 4)$ y $N(\mu_2, 4,5)$. Para ello, se han tomado dos muestras aleatorias simples e independientes, respectivamente, obteniéndose los siguientes valores:

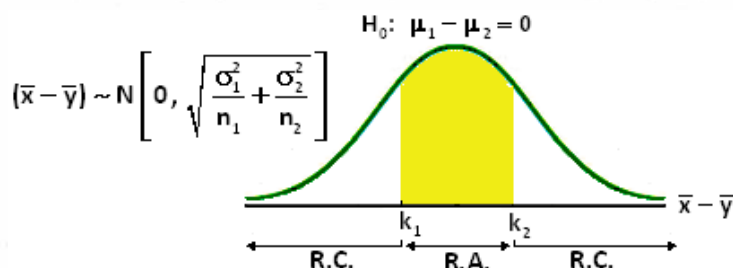
x_i	20,4	10,2	7,3	12,8	13,4	9,4
y_j	19,8	9,7	14,6	15,7	8,4	

Solución:

En el contraste bilateral se establecen las hipótesis: $H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 \neq 0$

Regla de decisión $\begin{cases} |\bar{x} - \bar{y}| \leq k & \text{se acepta } H_0 \\ |\bar{x} - \bar{y}| > k & \text{se rechaza } H_0 \end{cases}$

Para analizar el contraste se calculan las medias muestrales $\bar{x}$ e $\bar{y}$



$$\bar{x} = \sum_{i=1}^6 \frac{x_i}{6} = 12,25 \quad \bar{y} = \sum_{j=1}^5 \frac{y_j}{5} = 13,64 \quad n_1 = 6 \quad n_2 = 5$$

La diferencia de medias muestrales $(\bar{x} - \bar{y})$ bajo la hipótesis nula sigue una distribución:

$$(\bar{x} - \bar{y}) \sim N\left[0, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right] \equiv N\left[0, \sqrt{\frac{4^2}{6} + \frac{4,5^2}{5}}\right] = N[0, 2,59]$$

Los valores críticos k se determinan bajo el nivel de significación $\alpha = 0,0472$, considerando que se trata de un contraste bilateral (o de dos colas):

$$\begin{aligned}\alpha &= P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P\left[|\bar{x} - \bar{y}| > k \mid N(0, 2,59)\right] = P\left[\left|\frac{(\bar{x} - \bar{y}) - 0}{2,59}\right| > \frac{k - 0}{2,59}\right] = \\ &= P\left[|z| > \frac{k - 0}{2,59}\right] = P\left[\left(z \leq \frac{k_1}{2,59}\right) \cup \left(z \geq \frac{k_2}{2,59}\right)\right] = P\left[\left(z \geq \frac{-k_1}{2,59}\right) \cup \left(z \geq \frac{k_2}{2,59}\right)\right] = \\ &= 0,0236 + 0,0236 = 0,0472\end{aligned}$$

$$\frac{-k_1}{2,59} = 1,995 = z_{0,0236} \rightarrow k_1 = -2,59 \times 1,995 = -5,17$$

$$\frac{k_2}{2,59} = 1,995 = z_{0,0236} \rightarrow k_2 = 2,59 \times 1,995 = 5,17$$

En consecuencia, la región de aceptación: $-5,17 \leq \bar{x} - \bar{y} \leq 5,17$

La evidencia empírica $|\bar{x} - \bar{y}| = |12,25 - 13,64| = 1,39$ valor que se encuentra en la región de aceptación por lo que se acepta la hipótesis nula de igualdad de medias con un nivel de significación del 4,72%.

✓ Se puede resolver mediante el p_valor:

$$\begin{aligned}\alpha_p &= p_valor = P[\text{Rechazar el estadístico muestral observado} \mid H_0 \text{ es Cierta}] = \\ &= P\left[|\bar{x} - \bar{y}| > 1,39 \mid N(0, 2,59)\right] = P\left[\left|\frac{(\bar{x} - \bar{y}) - 0}{2,59}\right| > \frac{1,39 - 0}{2,59}\right] = P\left[|z| > \frac{1,39 - 0}{2,59}\right] = \\ &= P\left[|z| > 0,5367\right] = P\left[z < -0,5367\right] + P\left[z > 0,5367\right] = 2 \cdot P\left[z > 0,5367\right] = 2 \cdot 0,095 = 0,19\end{aligned}$$

O bien aplicando consecuencias directas:

⇒ Aceptando la hipótesis nula H_0 cuando el estadístico observado o de contraste es menor que el estadístico teórico:

$$z_{\alpha_p/2} = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2} = z_{0,0236} \rightarrow z_{\alpha_p/2} = \frac{1,39}{2,59} = 0,5367 \leq 1,995 = z_{0,0236}$$

$$\alpha_p / 2 = P\left[z_{\alpha_p/2} \geq 0,5367\right] = 0,095 \rightarrow \alpha_p = 0,19 > 0,0472 = \alpha \text{ Se acepta } H_0$$

⇒ Aceptando la hipótesis nula H_0 cuando no se verifica la región de rechazo:

$$R = \left\{ |\bar{x} - \bar{y}| > z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right\} \rightarrow R = \left\{ |12,25 - 13,64| > 1,995 \cdot \sqrt{\frac{4^2}{6} + \frac{4,5^2}{5}} \right\}$$

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

La región de rechazo de la hipótesis nula no se cumple $R = \{1,39 \neq 5,17\}$, concluyendo que existe igualdad entre las medias poblacionales.

⇒ Verificando que el intervalo de confianza cubre el valor 0, en este caso existe igualdad de medias poblacionales.

$$I(\mu_1 - \mu_2) = \left[(\bar{x} - \bar{y}) \pm z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right] \longrightarrow I(\mu_1 - \mu_2) = \left[(12,25 - 13,64) \pm 1,995 \cdot \sqrt{\frac{4^2}{6} + \frac{4,5^2}{5}} \right]$$

$$I(\mu_1 - \mu_2) = [-1,39 \pm 5,17] = [-6,56, 3,78] \longrightarrow 0 \in [-6,56, 3,78] \text{ Se acepta } H_0$$

	Abcisas	Áreas			
Cálculo de $z_{0,0236}$:	1,98 - 1,99	0,0239 - 0,0233	0,01	0,006	$\rightarrow x = 1,99 + \frac{0,01 \cdot 0,0003}{0,0006} = 1,995$
	x - 1,99	0,0236 - 0,0233	x - 1,99	0,0003	

134. La directora del departamento de personal de una corporación está buscando empleados para un puesto en el extranjero. Durante el proceso de selección, la administración le pregunta cómo va la incorporación de empleados, y ella contesta que sigue una ley normal y que la puntuación promedio en la prueba de aptitudes será de 90 puntos.

Cuando la administración revisa 19 de los resultados de la prueba, encuentra que la puntuación media es de 86,25 puntos con una desviación estándar de 11.

Con un nivel de confianza del 90%, ¿lleva razón la directora?. Calcula su p_{valor} .

Solución:

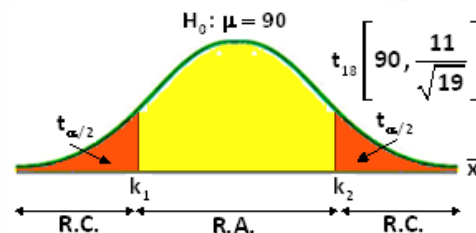
Se establecen las hipótesis: $H_0: \mu = 90$ $H_1: \mu \neq 90$

En el muestreo de la población normal con varianza desconocida, con muestras pequeñas

$n = 19 < 30$, la media muestral $\bar{x} \sim t_{10} \left(\mu, \frac{s_x}{\sqrt{19}} \right) \equiv \left(\mu, \frac{\sigma_x}{\sqrt{18}} \right)$

Como la hipótesis alternativa es $\mu \neq 90$ en la decisión deberán ser válidos valores de μ tanto mayores o menores que 90, por lo cual el contraste debe ser bilateral o de dos colas.

Regla decisión $\begin{cases} |\bar{x}| \leq k & \text{se acepta } H_0 \text{ (R.A.)} \\ |\bar{x}| > k & \text{se rechaza } H_0 \text{ (R.C.)} \end{cases}$



En el muestreo, bajo la hipótesis nula $\bar{x} \sim t_{18} \left[90, \frac{11}{\sqrt{19}} \right] \equiv t_{18} (90, 2,524)$

Los valores críticos k se determinan con el nivel de significación $\alpha = 0,10$

$$\alpha = P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P[|\bar{x}| > k \mid t_{18}(90, 2,524)] = P[(\bar{x} < k_1) \cup (\bar{x} > k_2)] = P(\bar{x} < k_1) + P(\bar{x} > k_2) = 0,05 + 0,05 = 0,10$$

$$P(\bar{x} < k_1) = P\left[\frac{\bar{x} - 90}{2,524} < \frac{k_1 - 90}{2,524}\right] = P\left[t_{18} < \frac{k_1 - 90}{2,524}\right] = P\left[t_{18} \geq -\frac{k_1 - 90}{2,524}\right] = 0,05$$

$$-\frac{k_1 - 90}{2,524} = 1,734 = t_{0,05, 18} \rightarrow k_1 = 85,62$$

$$P(\bar{x} > k_2) = P\left[\frac{\bar{x} - 90}{2,524} > \frac{k_2 - 90}{2,524}\right] = P\left[t_{18} > \frac{k_2 - 90}{2,524}\right] = 0,05$$

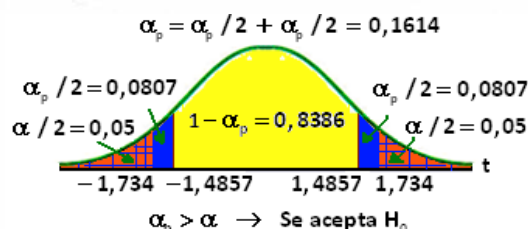
$$\frac{k_2 - 90}{2,524} = 1,734 = t_{0,05, 18} \rightarrow k_2 = 94,37$$

La media muestral observada $\bar{x} = 86,25$ se encuentra en la región de aceptación $86,25 \in [85,62, 94,37]$, con un nivel de significación de 0,10 se acepta la manifestación de la directora de la corporación.

$$\begin{aligned}\alpha_p = p_valor &= P[\text{Rechazar el estadístico muestral observado} / H_0 \text{ es Cierta}] = \\ &= P\left[\left|\bar{x}\right| > 86,25 \mid t_{18}(90, 2,524)\right] = P\left[\left|\frac{\bar{x} - 90}{2,524}\right| > \frac{86,25 - 90}{2,524}\right] = P\left[\left|t_{18}\right| > -1,4857\right] = \\ &= P\left[\left|t_{18}\right| > 2,6774\right] = P\left[t_{18} < -1,4857\right] + P\left[t_{18} > 1,4857\right] = 2P\left[t_{18} > 1,4857\right] = 2(0,0807) = 0,1614\end{aligned}$$

Tabla t-Student: $\frac{1,330 - 1,734}{0,1 - 0,05} = \frac{1,4857 - 1,734}{x - 0,05} \rightarrow x = 0,05 + \frac{0,05 \cdot 0,2483}{0,404} = 0,0807$

Como $\alpha_p = 0,1614 > 0,10 = \alpha$ se acepta la hipótesis nula y en consecuencia es cierta la manifestación de la directora de la corporación.



O bien aplicando consecuencias directas:

⇒ Aceptando la hipótesis nula H_0 cuando el estadístico observado o de contraste es menor que el estadístico teórico:

$$t_{\alpha_p/2} = \frac{|\bar{x} - \mu_0|}{s_x / \sqrt{n}} \leq t_{\alpha/2, n-1} = t_{0,05, 18} \rightarrow t_{\alpha_p/2} = \frac{|86,25 - 90|}{2,524} = 1,4857 < 1,734 = t_{0,05, 18} \rightarrow$$

→ Se acepta la hipótesis nula H_0

Adviértase que $\alpha_p/2 = P[t_{\alpha_p/2} > 1,4857] = 0,0807$

En la tabla de la t de Student: $P[t_{18} \geq 1,330] = 0,1$ y $P[t_{18} \geq 1,734] = 0,05$, interpolando se tiene:

0,10	$\alpha_p/2$	0,05		0,10 - 0,05	↔	1,330 - 1,734		0,05	↔	- 0,404
1,330	1,4857	1,734		$\alpha_p/2 - 0,05$	↔	1,4857 - 1,734		$\alpha_p/2 - 0,05$	↔	- 0,2483

$$(\alpha_p/2 - 0,05) \times 0,404 = 0,05 \times 0,2483 \rightarrow \alpha_p/2 = 0,05 + \frac{0,05 \times 0,2483}{0,404} = 0,0807 \rightarrow \alpha_p = 0,1614$$

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

⇒ Alternativamente, el problema se puede plantear desde la región de rechazo:

$$R = \left\{ \bar{x} - \mu_0 > t_{\alpha, n-1} \frac{s_x}{\sqrt{n}} \right\} \equiv \left\{ \bar{x} - \mu_0 > t_{\alpha, n-1} \frac{\sigma_x}{\sqrt{n-1}} \right\}$$

$$\text{Región de rechazo: } R = \left\{ |86,25 - 90| > \overbrace{1,734}^{t_{0,05,18}} \frac{11}{\sqrt{19}} \right\} = [3,75 \not> 4,3758]$$

Como no se verifica la región de rechazo, no existe evidencia significativa para rechazar la hipótesis H_0 con un nivel de confianza del 90%.

⇒ Si el intervalo de confianza para la media poblacional μ cubre el valor $\mu = 90$ se acepta la

$$\text{hipótesis nula: } \text{I.C.} = \left[\bar{x} \pm t_{\alpha/2, n-1} \frac{s_x}{\sqrt{n}} \right] = \left[86,25 \pm 1,734 \frac{11}{\sqrt{19}} \right] = [81,873, 90,627]$$

$$90 \in [81,873, 90,627] \rightarrow 81,873 < 90 < 90,627 \rightarrow \text{Se acepta } H_0$$

135. Una empresa ubicada en Madrid tiene dos conductores para trasladar a los empleados a Segovia. Los conductores deben anotar la duración de cada trayecto. En una muestra aleatoria simple de 50 partes de incidencias por conductor, el conductor A registra un tiempo medio de trayecto de 62,30 minutos con una desviación típica de 10,325 minutos, mientras que el conductor B tiene un tiempo medio de trayecto de 60,02 minutos con una desviación típica de 8,625 minutos.

El tiempo medio empleado por el conductor A en el trayecto sigue una ley normal $N(\mu_1, 9)$ y el empleado por el conductor B se distribuye según una ley $N(\mu_2, 8)$.

Con un nivel de significación del 5%, se pide contrastar:

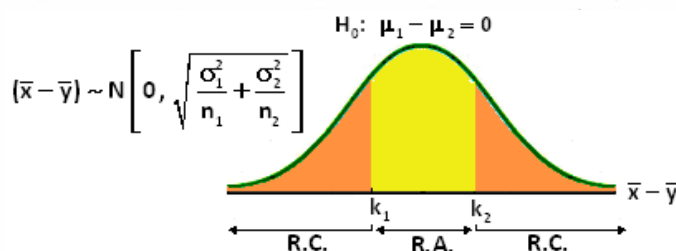
- a) $H_0: \mu_1 = \mu_2$ frente a $H_1: \mu_1 \neq \mu_2$
b) $H_0: \mu_1 - \mu_2 \geq 2$ frente a $H_1: \mu_1 - \mu_2 < 2$

Solución:

a) Para realizar el contraste bilateral planteado $H_0: \mu_1 = \mu_2$ se recurre al test razón de verosimilitud, en este caso el lema de Neyman-Pearson no proporciona una región crítica óptima.

La regla de decisión que proporciona el test de razón de verosimilitud es:

$$\begin{cases} |\bar{x} - \bar{y}| \leq k & \text{se acepta } H_0 \text{ (R.A.)} \\ |\bar{x} - \bar{y}| > k & \text{se rechaza } H_0 \text{ (R.C.)} \end{cases}$$



La región crítica de dos colas $|\bar{x} - \bar{y}| > k$ es función de la diferencia de las medias muestrales. En esta línea, las distribuciones en el muestreo de las medias son:

$$\bar{x} \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \quad \bar{y} \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Con lo cual, la diferencia de medias muestrales, bajo la hipótesis nula, se distribuye:

$$(\bar{x} - \bar{y}) \sim N\left[0, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right] \equiv N\left[0, \sqrt{\frac{9^2}{50} + \frac{8^2}{50}}\right] = N[0, 1,7]$$

Los valores críticos k se determina a partir del nivel de significación $\alpha = 0,05$

$$\begin{aligned} \alpha &= P(\text{Rechazar } H_0 \mid H_0 \text{ Cierta}) = P\left[|\bar{x} - \bar{y}| > k \mid N[0, 1,7]\right] = P\left[\left|\frac{\bar{x} - \bar{y} - 0}{1,7}\right| > \frac{k - 0}{1,7}\right] = \\ &= P\left[|z| > \frac{k}{1,7}\right] = P\left[\left(z < \frac{k_1}{1,7}\right) \cup \left(z > \frac{k_2}{1,7}\right)\right] = P\left[z > -\frac{k_1}{1,7}\right] + P\left[z > \frac{k_2}{1,7}\right] = 0,025 + 0,025 = 0,05 \end{aligned}$$

Por simetría: $\frac{-k_1}{1,7} = 1,96 = z_{0,025} \rightarrow k_1 = -3,33$ $\frac{k_2}{1,7} = 1,96 = z_{0,025} \rightarrow k_2 = 3,33$

La región de aceptación: $-3,33 \leq \bar{x} - \bar{y} \leq 3,33$

La evidencia empírica $|\bar{x} - \bar{y}| = |62,30 - 60,02| = 2,28$ valor que se encuentra en la región de aceptación, por lo que se acepta la hipótesis nula de igualdad de medias.

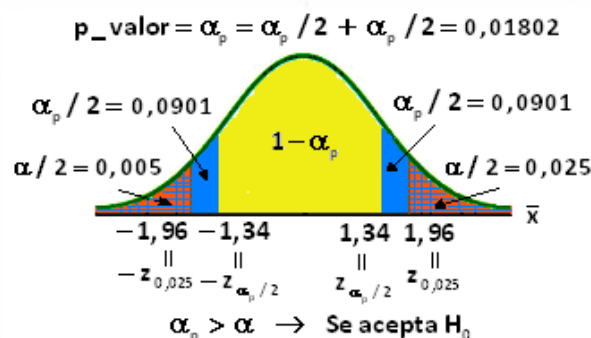
Los dos conductores emplean en promedio el mismo tiempo en el trayecto Madrid – Segovia, con una fiabilidad del 95%.

⇒ Se podía haber decidido con el p_valor o métodos aplicando sus consecuencias directas:

$$\begin{aligned} \alpha_p &= p_valor = P[\text{Rechazar el estadístico muestral observado} / H_0 \text{ es Cierta}] = \\ &= P[|\bar{x} - \bar{y}| > 2,28 \mid N[0, 1,7]] = P\left[\left|\frac{\bar{x} - \bar{y} - 0}{1,7}\right| > \frac{2,28 - 0}{1,7}\right] = P[|z| > 1,34] = \\ &= P[z < -1,34] + P[z > 1,34] = 2P[z > 1,34] = 2(0,0901) = 0,1802 \end{aligned}$$

Siendo $\alpha_p = 0,1802 > 0,05 = \alpha \rightarrow$ Se acepta H_0

El p_valor es el nivel de significación más pequeño posible que puede elegirse para el que se aceptaría la hipótesis alternativa.



⇒ Aceptando la hipótesis nula H_0 cuando el estadístico observado o de contraste es menor que el estadístico teórico:

$$z_{\alpha_p/2} = \frac{|\bar{x} - \bar{y}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2} \rightarrow z_{\alpha_p/2} = \frac{|62,30 - 60,02|}{1,7} = 1,34 < 1,96 = z_{0,025} \rightarrow$$

→ Se acepta la hipótesis nula $H_0 \Rightarrow$ Los dos conductores emplean en promedio el mismo tiempo en el trayecto Madrid – Segovia.

Adviértase que $\alpha_p / 2 = P[z_{\alpha_p/2} > 1,34] = 0,0901 \rightarrow \alpha_p = 0,1802$

⇒ Alternativamente, el problema se puede plantear desde la región de rechazo:

$$R = \left\{ |\bar{x} - \bar{y}| > z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right\} \rightarrow R = \left\{ |62,30 - 60,02| > 1,96 \sqrt{\frac{9^2}{50} + \frac{8^2}{50}} \right\} = \{2,28 \nless 3,332\}$$

No se verifica la región de rechazo y se admite la hipótesis nula.

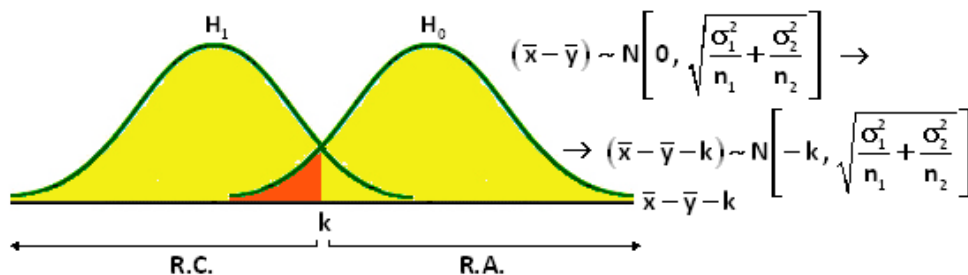
⇒ Si el intervalo de confianza para la diferencia de medias poblacionales $\mu_1 - \mu_2$ cubre el

valor 0 se acepta la hipótesis nula: $I_{1-\alpha}(\mu_1 - \mu_2) = \left[(\bar{x} - \bar{y}) \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$

$$I_{0,95}(\mu_1 - \mu_2) = [2,28 \pm 1,96 \times 1,7] = [-1,05, 5,61]$$

Como el intervalo cubre el cero $-1,05 < 0 < 5,61$ no existe diferencia significativa entre el tiempo empleado por los dos conductores en el trayecto Madrid – Segovia.

b) Hipótesis nula planteada $H_0: \mu_1 - \mu_2 - 2 \geq 0$. Se trata de un contraste unilateral con cola a la izquierda $H_1: \mu_1 - \mu_2 - 2 < 0$



Bajo la hipótesis nula $(\bar{x} - \bar{y}) \sim N[0, 1,7] \rightarrow (\bar{x} - \bar{y} - 2) \sim N[-2, 1,7]$

Regla decisión $\begin{cases} \bar{x} - \bar{y} - 2 \geq k & \text{se acepta } H_0 \text{ (R.A.)} \\ \bar{x} - \bar{y} - 2 < k & \text{se rechaza } H_0 \text{ (R.C.)} \end{cases} \rightarrow \begin{cases} \bar{x} - \bar{y} \geq 2 + k & \text{se acepta } H_0 \text{ (R.A.)} \\ \bar{x} - \bar{y} < 2 + k & \text{se rechaza } H_0 \text{ (R.C.)} \end{cases}$

$$\alpha = P(\text{Rechazar } H_0 \mid H_0 \text{ Cierta}) = P[\bar{x} - \bar{y} - 2 < k \mid N[0, 1,7]] = P[\bar{x} - \bar{y} < k + 2 \mid N[0, 1,7]] =$$

$$= P\left[\frac{\bar{x} - \bar{y} - 0}{1,7} < \frac{k + 2 - 0}{1,7}\right] = P\left[z < \frac{k + 2}{1,7}\right] = P\left[z > \frac{-k - 2}{1,7}\right] = 0,05 \rightarrow \frac{-k - 2}{1,7} = 1,645$$

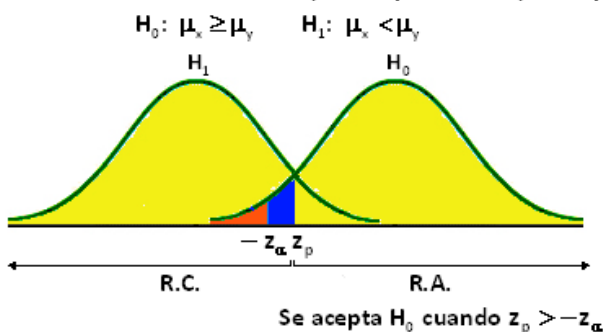
$$\rightarrow k + 2 = -2,7965$$

Regla decisión $\begin{cases} \bar{x} - \bar{y} \geq -2,7965 & \text{se acepta } H_0 \\ \bar{x} - \bar{y} < -2,7965 & \text{se rechaza } H_0 \end{cases}$ siendo $\bar{x} - \bar{y} = 2,28$ se acepta H_0

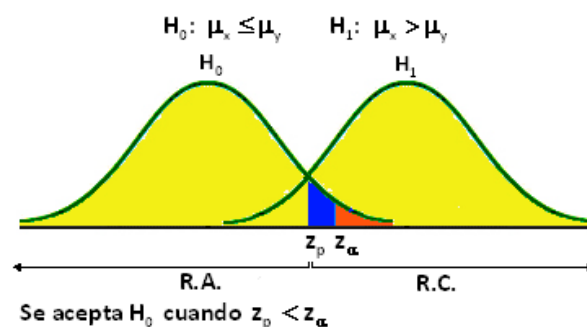
⇒ Se acepta la hipótesis nula si $z_p = z_{\text{observado}} > -z_{\alpha}$

$$z_p = \frac{62,30 - 60,02 - 2}{\sqrt{\frac{9^2}{50} + \frac{8^2}{50}}} = 0,164 > -1,645 = z_{0,05} \rightarrow \text{Se acepta } H_0$$

Contraste unilateral a la izquierda (cola a la izquierda)



Contraste unilateral a la derecha (cola a la derecha)



136. Una compañía de refrescos presenta un nuevo producto en el mercado afirmando que posee menos calorías que su homólogo más antiguo y conserva el resto de propiedades. Para verificar la afirmación de la compañía se eligieron al azar 14 botes del refresco nuevo y se calculó su media, 20 calorías por bote, y su desviación estándar muestral, tres calorías. De modo independiente, se tomó otra muestra aleatoria de 16 botes del refresco antiguo, obteniéndose una media de 28 calorías por bote con una desviación estándar muestral de 5 calorías. Suponiendo que la cantidad de calorías por bote sigue una distribución normal en ambos refrescos, pero con desviaciones típicas diferentes, ¿existe alguna razón para no creer en la afirmación de la compañía con un nivel de significación del 2,5%?

Solución:

Denotando por X e Y las variables aleatorias que representan la cantidad de calorías por bote en el nuevo producto y en el antiguo, respectivamente: $X \sim N(\mu_x, \sigma_x)$ e $Y \sim N(\mu_y, \sigma_y)$ siendo X e Y independientes y con $\sigma_x \neq \sigma_y$

Se pretenden contrastar las hipótesis:

$$H_0: \mu_x - \mu_y \geq 0$$

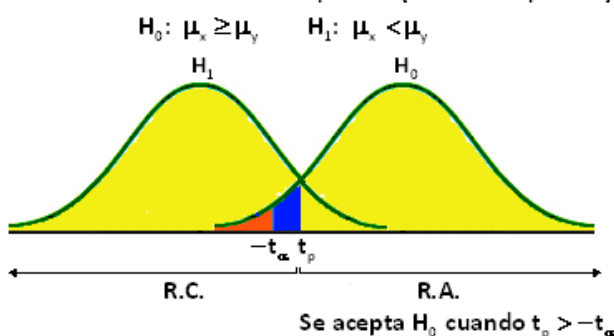
$$H_1: \mu_x - \mu_y < 0 \quad (\text{el nuevo producto posee menos calorías que el antiguo})$$

Se trata de un contraste unilateral a la izquierda. Se acepta H_0 si $t_p = t_{\text{observado}} > -t_\alpha$

Estadístico de contraste:

$$t_p = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} = \frac{20 - 28}{\sqrt{\frac{9}{14} + \frac{25}{16}}} = -5,387$$

Contraste unilateral a la izquierda (cola a la izquierda)



$$f = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\frac{\left(\frac{s_x^2}{n_x}\right)^2}{n_x + 1} + \frac{\left(\frac{s_y^2}{n_y}\right)^2}{n_y + 1}} - 2 = \frac{\left(\frac{9}{14} + \frac{25}{16}\right)^2}{\frac{\left(\frac{9}{14}\right)^2}{15} + \frac{\left(\frac{25}{16}\right)^2}{17}} - 2 \approx 26$$

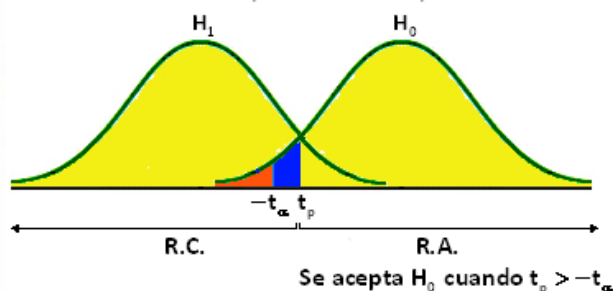
El p_valor es el nivel de significación más pequeño para el que se acepta la hipótesis alternativa.

Estadístico teórico: $-t_{0,025, 26} = -2,056$

Como $t_{\text{observado}} = -5,387 < -2,056 = -t_{0,025, 26}$ se rechaza la hipótesis nula concluyendo que no existen razones para no creer en la afirmación de la compañía.

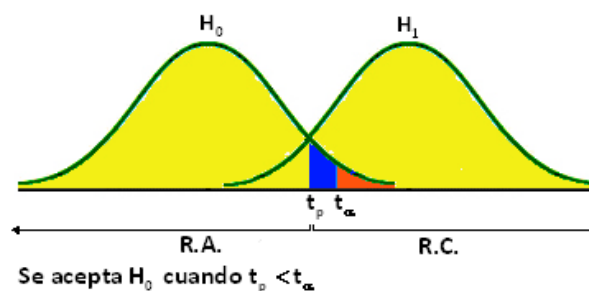
Contraste unilateral a la izquierda (cola a la izquierda)

$$H_0: \mu_x \geq \mu_y \quad H_1: \mu_x < \mu_y$$



Contraste unilateral a la derecha (cola a la derecha)

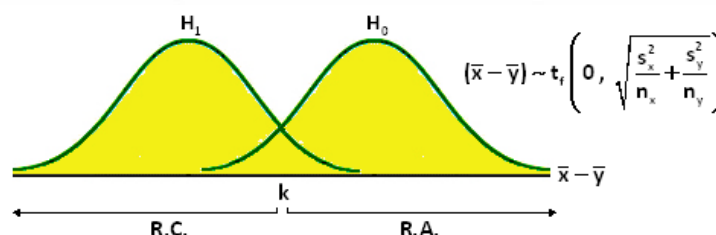
$$H_0: \mu_x \leq \mu_y \quad H_1: \mu_x > \mu_y$$



Otro procedimiento

Regla de decisión:

$$\begin{cases} \bar{x} - \bar{y} \geq k & \text{Se acepta } H_0 \\ \bar{x} - \bar{y} < k & \text{Se rechaza } H_0 \end{cases}$$



Bajo la hipótesis nula $(\bar{x} - \bar{y}) \sim t_f\left(0, \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}\right) = t_f\left(0, \sqrt{\frac{9}{14} + \frac{25}{16}}\right) \equiv t_{26}(0, 1,485)$

$$\begin{aligned} \alpha &= P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P[(\bar{x} - \bar{y}) < k \mid t_{26}(0, 1,485)] = P\left[t_{26} < \frac{k}{1,485}\right] = \\ &= P\left[t_{26} > \frac{-k}{1,485}\right] = 0,025 \rightarrow \frac{-k}{1,485} = 2,056 \rightarrow k = -3,053 \end{aligned}$$

Siendo $\bar{x} - \bar{y} = 20 - 28 = -8 < -3,053$ se rechaza la hipótesis nula, con lo que no se puede corroborar lo que dice la compañía con un nivel de significación del 2,5%.

137. El estacionamiento de corta estancia en el Aeropuerto Adolfo Suárez está cerca de la terminal, de modo que al recoger a un pasajero solo hay que caminar una corta distancia al área de recuperación de equipajes. Para decidir si el estacionamiento tiene suficientes lugares, el gerente del aeropuerto necesita saber si el tiempo medio de permanencia es superior a 40 minutos. Una muestra aleatoria de una población normal mostró que 12 clientes estuvieron en el estacionamiento los siguientes lapsos de tiempo, en minutos:

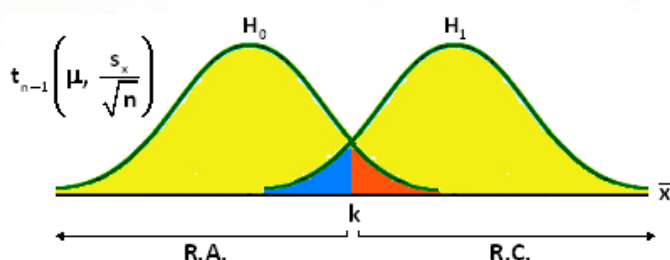
55 47 27 64 53 56
49 53 39 48 48 37

Con un nivel de significación 0,05, ¿se puede afirmar que el tiempo medio en el estacionamiento es superior a 40 minutos?

Solución:

Se establecen las hipótesis nula y alternativa: $H_0 : \mu \leq 40$ $H_1 : \mu > 40$

Se trata de un contraste unilateral (cola a la derecha)



Regla de decisión:

$$\begin{cases} \text{Si } \bar{x} \leq k & \text{Se acepta } H_0 \text{ (R.A.)} \\ \text{Si } \bar{x} > k & \text{Se rechaza } H_0 \text{ (R.C.)} \end{cases}$$

En el muestreo de una población normal con varianza desconocida, y desviación típica muestral σ_x , la variable $t_{n-1} = \frac{\bar{x} - \mu}{\frac{s_x}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\frac{\sigma_x}{\sqrt{n-1}}}$

En la muestra: $\bar{x} = \frac{1}{12} \sum_{i=1}^{12} x_i = \frac{576}{12} = 48$ $s_x^2 = \frac{1}{11} \sum_{i=1}^{12} (x_i - \bar{x})^2 = \frac{1064}{11} = 96,727$

$$s_x = \sqrt{96,73} = 9,835 \rightarrow t_{11} = \frac{\bar{x} - \mu}{\frac{s_x}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\frac{9,835}{\sqrt{12}}}$$

Bajo la hipótesis nula, la muestra sigue una distribución $t_{11} = \frac{\bar{x} - 40}{2,839}$

A partir del nivel de significación α se determina el valor crítico k:

$$\alpha = P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P\left[\bar{x} > k \mid t_{n-1}\left(40, \frac{s_x}{\sqrt{n}}\right)\right] =$$

$$= P\left[\frac{\bar{x} - 40}{9,835 / \sqrt{12}} > \frac{k - 40}{9,835 / \sqrt{12}}\right] = P\left[t_{11} > \frac{k - 40}{2,839}\right] = 0,05 \rightarrow \frac{k - 40}{2,839} = 1,796 \Rightarrow k = 45,098$$

$$\begin{cases} \text{Si } \bar{x} \leq 45,098 & \text{Se acepta } H_0 \text{ (R.A)} \\ \text{Si } \bar{x} > 45,098 & \text{Se rechaza } H_0 \text{ (R.C)} \end{cases}$$

Siendo $\bar{x} = 48 > 45,098$ se rechaza la hipótesis nula, concluyendo que el tiempo medio que los clientes pasan en el estacionamiento es superior a 40 minutos. El aeropuerto necesita ampliar su estacionamiento.

138. Muebles Quintana manifiesta que la satisfacción de sus empleados sigue una ley normal y es idéntica en las distintas secciones. El jefe de recursos humanos para decidir si la afirmación es correcta toma una muestra aleatoria de 100 trabajadores de barnizado encontrando 80 satisfechos y otra de 200 trabajadores de cortado encontrando 140 empleados satisfechos. Con un nivel de significación del 5% se puede admitir la afirmación de Muebles Quintana.

Solución:

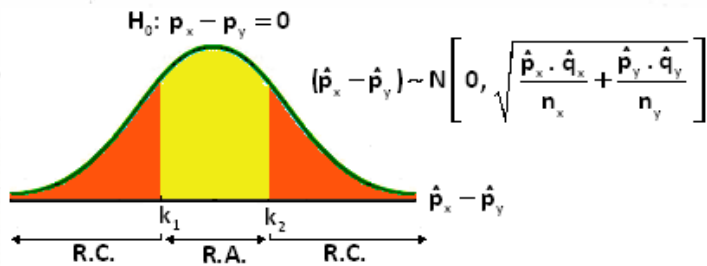
Se contrasta la hipótesis nula de que no existen diferencias entre las proporciones de trabajadores satisfechos en ambas secciones: $H_0: p_x = p_y$ frente a la hipótesis alternativa de que si existen diferencias $H_1: p_x \neq p_y$. El contraste es bilateral o de dos colas.

Las variables aleatorias $X \equiv$ "Trabajador de la sección de barnizado" e $Y \equiv$ "Trabajador de la sección de cortado", respectivamente, siguen una distribución de Bernouilli que toman el valor uno si el trabajador está satisfecho, y el valor cero en caso contrario.

Proporciones muestrales obtenidas son: $\hat{p}_x = \frac{80}{100} = 0,8$ $\hat{p}_y = \frac{140}{200} = 0,7$

Bajo la hipótesis nula la diferencia de las proporciones muestrales $(\hat{p}_x - \hat{p}_y)$, teniendo en cuenta el tamaño de las muestras (TCL) siguen una distribución

$$(\hat{p}_x - \hat{p}_y) \sim N\left[0, \sqrt{\frac{\hat{p}_x \cdot \hat{q}_x}{n_x} + \frac{\hat{p}_y \cdot \hat{q}_y}{n_y}}\right] \equiv N\left[0, \sqrt{\frac{0,8 \cdot 0,2}{100} + \frac{0,7 \cdot 0,3}{200}}\right] \equiv N[0, 0,05]$$



Regla de decisión:

$$\begin{cases} |\hat{p}_x - \hat{p}_y| \leq k & \text{Se acepta } H_0 \text{ (RA)} \\ |\hat{p}_x - \hat{p}_y| > k & \text{Se rechaza } H_0 \text{ (RC)} \end{cases} \rightarrow \begin{cases} k_1 \leq \hat{p}_x - \hat{p}_y \leq k_2 & \text{Se acepta } H_0 \text{ (RA)} \\ \hat{p}_x - \hat{p}_y < k_1 \text{ ó } \hat{p}_x - \hat{p}_y > k_2 & \text{Se rechaza } H_0 \text{ (RC)} \end{cases}$$

Los valores críticos k se determinan mediante el nivel de significación $\alpha = 0,05$

$$\begin{aligned} \alpha &= P(\text{Rechazar } H_0 \mid H_0 \text{ Cierta}) = P\left[|\hat{p}_x - \hat{p}_y| > k \mid N[0, 0,05]\right] = P\left[\left|\frac{\hat{p}_x - \hat{p}_y - 0}{0,05}\right| > \frac{k - 0}{0,05}\right] = \\ &= P\left[|z| > \frac{k}{0,05}\right] = P\left[\left(z < \frac{k_1}{0,05}\right) \cup \left(z > \frac{k_2}{0,05}\right)\right] = P\left[z > -\frac{k_1}{0,05}\right] + P\left[z > \frac{k_2}{0,05}\right] = 0,025 + 0,025 = 0,05 \end{aligned}$$

$$P\left[z > -\frac{k_1}{0,05}\right] = 0,025 \rightarrow -\frac{k_1}{0,05} = 1,96 \rightarrow k_1 = -0,098$$

$$P\left[z > \frac{k_2}{0,05}\right] = 0,025 \rightarrow \frac{k_2}{0,05} = 1,96 \rightarrow k_2 = 0,098$$

Región de aceptación: $-0,098 < \hat{p}_x - \hat{p}_y < 0,098$

La evidencia empírica (estadístico observado) $|\hat{p}_x - \hat{p}_y| = |0,8 - 0,7| = 0,1$ no se encuentra en la región de aceptación, por lo que no se admite la hipótesis nula. La satisfacción entre los trabajadores de las distintas secciones no es la misma.

⇒ Aceptando la hipótesis nula H_0 cuando el estadístico observado o de contraste es menor que el estadístico teórico:

$$z_{\alpha_p/2} = \frac{|\hat{p}_x - \hat{p}_y|}{\sqrt{\frac{\hat{p}_x \cdot \hat{q}_x}{n_x} + \frac{\hat{p}_y \cdot \hat{q}_y}{n_y}}} \leq z_{\alpha/2} \rightarrow z_{\alpha_p/2} = \frac{|0,8 - 0,7|}{0,05} = 2 \not\leq 1,96 = z_{0,025} \rightarrow$$

→ Se rechaza la hipótesis nula H_0 ⇒ La satisfacción de los trabajadores no es la misma en las secciones de Muebles Quintana

⇒ Rechazando la hipótesis nula H_0 al verificarse la región de rechazo:

$$R.C = \left\{ |\hat{p}_x - \hat{p}_y| > z_{\alpha/2} \sqrt{\frac{\hat{p}_x \cdot \hat{q}_x}{n_x} + \frac{\hat{p}_y \cdot \hat{q}_y}{n_y}} \right\} \rightarrow \left\{ (0,8 - 0,7) > 1,96 \sqrt{\frac{0,8 \cdot 0,2}{100} + \frac{0,7 \cdot 0,3}{200}} \right\} =$$

$$= (0,1 > 0,098) \rightarrow \text{Se acepta la región de rechazo de } H_0$$

⇒ Elaborando un intervalo de confianza para la diferencia de parámetros $(\hat{p}_x - \hat{p}_y)$ de dos distribuciones binomiales

$$I(p_x - p_y) = \left[\hat{p}_x - \hat{p}_y \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_x \cdot \hat{q}_x}{n_x} + \frac{\hat{p}_y \cdot \hat{q}_y}{n_y}} \right] = \left[0,1 \pm 1,96 \sqrt{\frac{0,8 \cdot 0,2}{100} + \frac{0,7 \cdot 0,3}{200}} \right] =$$

$$= [0,002, 0,198] \rightarrow 0 \notin [0,002, 0,198] \rightarrow \text{Se rechaza } H_0$$

139. Se quiere comparar las puntuaciones en estadística teórica de dos Universidades que siguen, respectivamente, una distribución normal, con varianzas desconocidas y se supone que estadísticamente iguales. Se toman al azar dos muestras iguales de estudiantes de ambas Universidades. Los resultados obtenidos en SPSS son:

Prueba de muestras independientes

	Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias						
	F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Error típ de la diferencia	95% Intervalo de confianza para la diferencia	
								Inferior	Superior
Calificación Se han asumido varianzas iguales	,152	,699	,443	28	,661	2,200	4,962	-7,964	12,364

Con un nivel de significación de 0,05, se pide:

- ¿Existe diferencia significativa sobre las puntuaciones de estadística teórica en las dos Universidades?
- ¿Cuál es la media ponderada de las cuasivarianzas muestrales?
- Cuál sería el p -valor si se hubieran contrastado las puntuaciones

$$H_0 : \mu_{\text{Autónoma}} \leq \mu_{\text{Complutense}} \quad H_1 : \mu_{\text{Autónoma}} > \mu_{\text{Complutense}}$$

Solución:

- En los resultados obtenidos en SPSS la Prueba de Levene permite decidir si las varianzas poblacionales son iguales. Como la probabilidad asociada al estadístico de Levene (0,699) es mayor que 0,05 se acepta la hipótesis nula de que las varianzas poblacionales son iguales.

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

$$I_{1-\alpha}(\mu_1 - \mu_2) = \left[(\bar{x} - \bar{y}) \pm t_{\alpha/2, (n_1 + n_2 - 2)} \cdot \underbrace{s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}_{\text{error típico diferencia}} \right]$$

Las dos Universidades tienen la misma puntuación en estadística teórica a un nivel de significación del 0,05 ya que el intervalo de confianza cubre el cero.

Un intervalo de confianza sirve para tomar una decisión sobre la hipótesis nula que permite contrastar el estadístico T (t de Student).

Por otra parte,

$p_valor = \text{Sig. (bilateral)} = 0,661 > 0,05 \rightarrow$ Se acepta la hipótesis nula que establece

$H_0: \mu_{\text{Autónoma}} = \mu_{\text{Complutense}}$

b) $s_p^2 \equiv$ Media ponderada cuasivarianzas muestrales: $s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$

Error típico de la diferencia $= s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \rightarrow 4,962 = s_p \cdot \sqrt{\frac{1}{15} + \frac{1}{15}}$

$s_p = \frac{4,962}{\sqrt{\frac{1}{15} + \frac{1}{15}}} = \frac{4,962}{0,3651} = 13,588 \rightarrow s_p^2 = 184,633$

$t \equiv$ valor experimental del estadístico de contraste:

$t = \frac{\text{Diferencia de medias}}{\text{Error típico diferencia}} = \frac{|\bar{x} - \bar{y}|}{13,588 \cdot \sqrt{\frac{1}{15} + \frac{1}{15}}} = \frac{2,200}{4,962} = 0,443$

$t = 0,443 \leq t_{0,025, 28} = 2,048$ se acepta H_0

$p_valor = \text{Sig. (bilateral)} = 0,661$. La Significación bilateral muestra el grado de compatibilidad entre el valor poblacional propuesto y la información muestral disponible.

Si el nivel crítico es pequeño (generalmente menor que 0,05), se concluye que la información recogida en la muestra es incompatible con la hipótesis nula de que las medias poblacionales son iguales.

La Prueba T se basa en el supuesto de normalidad, el cumplimiento de este supuesto sólo es exigible con muestras pequeñas. El supuesto de normalidad carece de relevancia en muestras grandes.

c) SPSS siempre calcula el p_{valor} para un contraste bilateral o de dos colas, si se desea realizar un contraste unilateral o de una cola hay que dividir entre 2 el contraste bilateral ($p_{\text{valor}} = 0,699 / 2 = 0,3495$)

Contraste unilateral (cola a la derecha: $H_1: \mu_{\text{Autónoma}} > \mu_{\text{Complutense}}$)

$p_{\text{valor}} = \text{Sig.}(\text{unilateral derecha}) = 0,3495 > 0,05 \rightarrow$ Se acepta H_0

140. El precio de los productos vendidos por una empresa es una variable aleatoria con función de densidad $f(x) = \theta x^{\theta-1}$, donde $0 < x < 1$, $\theta > 0$, que depende del parámetro desconocido θ .

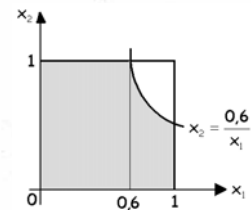
Se quiere contrastar sobre el valor de dicho parámetro la hipótesis nula $H_0: \theta = 1$ frente a la alternativa $H_1: \theta = 2$. Para ello, se toma una muestra aleatoria simple de tamaño dos.

Determinar el nivel de significación y la potencia del contraste, si se toma como región crítica $x_1 x_2 \leq 0,6$

Solución:

El nivel de significación α es la probabilidad de rechazar la hipótesis nula siendo cierta: $\alpha = P(\text{Rechazar } H_0 | H_0 \text{ Cierta})$.

En la región crítica $x_1 x_2 \leq 0,6$ para $\theta = 1$, se tiene $f(x) = 1$, con lo cual:



$$\begin{aligned} P[x_1 x_2 \leq 0,6 | H_0: \theta = 1] &= \int_{x_1=0}^{x_1=0,6} \int_{x_2=0}^{x_2=1} 1 \, dx_1 \, dx_2 + \int_{x_1=0,6}^{x_1=1} \int_{x_2=0}^{x_2=0,6/x_1} 1 \, dx_1 \, dx_2 = \\ &= \int_0^{0,6} (x_2)_0^1 \, dx_1 + \int_{0,6}^1 (x_2)_0^{0,6/x_1} \, dx_1 = \int_0^{0,6} 1 \, dx_1 + \int_{0,6}^1 (0,6/x_1) \, dx_1 = \\ &= (x_1)_0^{0,6} + (\ln x_1)_{0,6}^1 = 0,6 + 0,6 (\ln 1 - \ln 0,6) = 0,6 + 0,6 (0 + 0,51) = 0,906 \end{aligned}$$

La probabilidad de rechazar la hipótesis nula siendo cierta es alta, lo que indica que el contraste es malo.

Potencia $= 1 - \beta = P(\text{Rechazar } H_0 | H_0 \text{ Falsa}) \equiv P(\text{Rechazar } H_0 | H_1 \text{ Cierta})$.

En la región crítica $x_1 x_2 \leq 0,6$, para $\theta = 2$, se tiene $f(x) = 2x$, la potencia será:

$$\begin{aligned} \text{Pot} = P[x_1 x_2 \leq 0,6 | H_0: \theta = 2] &= \int_{x_1=0}^{x_1=0,6} \int_{x_2=0}^{x_2=1} 2x_1 2x_2 \, dx_1 \, dx_2 + \int_{x_1=0,6}^{x_1=1} \int_{x_2=0}^{x_2=0,6/x_1} 2x_1 2x_2 \, dx_1 \, dx_2 = \\ &= \int_0^{0,6} 4x_1 (x_2^2/2)_0^1 \, dx_1 + \int_{0,6}^1 4x_1 (x_2^2/2)_0^{0,6/x_1} \, dx_1 = \int_0^{0,6} 2x_1 \, dx_1 + \int_{0,6}^1 2(0,36/x_1) \, dx_1 = \\ &= (x_1^2)_0^{0,6} + 0,72 (\ln x_1)_{0,6}^1 = 0,36 + 0,72 (\ln 1 - \ln 0,6) = 0,36 + 0,72 (0 + 0,51) = 0,7272 \end{aligned}$$

La potencia del contraste no resulta excesivamente alta, el contraste no es bueno.

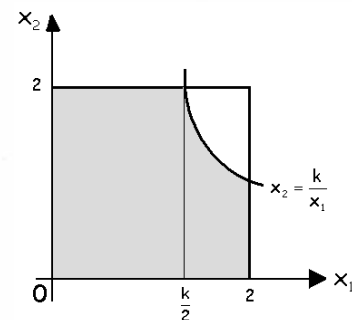
141. El volumen diario de ventas de una empresa, en cien mil euros, es una variable aleatoria X sobre cuya función de densidad se establece la hipótesis nula $H_0 : f(x) = x/2, 0 < x < 2$, frente a la hipótesis alternativa $H_1 : f(x) = 1/2, 0 < x < 2$. Para realizar el contraste se toma una muestra aleatoria simple de dos días en los que los volúmenes de venta fueron de 50.000 euros y 100.000 euros. Con un nivel de significación del 5 por ciento, se pide:

- Hallar la mejor región crítica.
- Calcular la potencia del contraste.
- Qué hipótesis se acepta.

Solución:

a) Para una muestra aleatoria simple de tamaño dos, la mejor región crítica aplicando el teorema de Neyman-Pearson, el cociente de las funciones de verosimilitud es:

$$\frac{L(x_1, x_2, f(x) = x/2)}{L(x_1, x_2, f(x) = 1/2)} = \frac{x_1 x_2}{1/2 \cdot 1/2} = x_1 x_2 \leq k$$



Si la hipótesis nula es cierta $f(x) = x/2$, con un nivel de significación del 5%, se tiene:

$$\begin{aligned} 0,05 &= P[x_1 x_2 \leq k \mid H_0 : f(x) = x/2] = \int_{x_1=0}^{x_1=k/2} \int_{x_2=0}^{x_2=2} \frac{x_1 x_2}{4} dx_1 dx_2 + \int_{x_1=k/2}^{x_1=2} \int_{x_2=0}^{x_2=k/x_1} \frac{x_1 x_2}{4} dx_1 dx_2 = \\ &= \int_{x_1=0}^{x_1=k/2} \frac{x_1}{4} \left(\frac{x_2^2}{2} \right)_0^2 dx_1 + \int_{x_1=k/2}^{x_1=2} \frac{x_1}{4} \left(\frac{x_2^2}{2} \right)_0^{k/x_1} dx_1 = \int_0^{k/2} \frac{x_1}{2} dx_1 + \int_{k/2}^2 \frac{k^2}{8x_1} dx_1 = \\ &= \frac{1}{4} (x_1^2)_0^{k/2} + \frac{k^2}{8} (\ln x_1)_{k/2}^2 = \frac{k^2}{16} + \frac{k^2}{8} [\ln 2 - \ln(k/2)] = \frac{k^2}{8} \left[\frac{1}{2} + 2 \ln 2 - \ln k \right] = \frac{k^2}{8} [1,8633 - \ln k] \rightarrow \\ &\rightarrow \frac{k^2}{8} [1,8633 - \ln k] = 0,05 \rightarrow k^2 [1,8633 - \ln k] = 0,40 \text{ de donde } k = 0,3733 \end{aligned}$$

La regla de decisión es $x_1 x_2 \leq 0,3733$

Es decir, se rechaza la hipótesis nula si durante los días el producto de los dos volúmenes de ventas de la muestra es inferior a 37.330 euros.

b) $\text{Potencia} = 1 - \beta = P[\text{Rechazar } H_0 \mid H_0 \text{ Falsa}] \equiv P[\text{Rechazar}_0 \mid H_1 \text{ Cierta}]$

En la región crítica $x_1 x_2 \leq 0,3733$, con la función de densidad $f(x) = 1/2$, la potencia:

$$\begin{aligned}
 P &= P[x_1 x_2 \leq 0,3733 \mid H_1 : f(x) = 1/2] = \int_{x_1=0}^{x_1=0,3733/2} \int_0^2 \frac{1}{4} dx_1 dx_2 + \int_{x_1=0,3733/2}^{x_1=2} \int_0^{0,3733/x_1} \frac{1}{4} dx_1 dx_2 = \\
 &= \int_0^{0,3733/2} \frac{1}{4} (x_2)_0^2 dx_1 + \int_{0,3733/2}^2 \frac{1}{4} (x_2)_0^{0,3733/x_1} dx_1 = \frac{1}{2} \int_0^{0,3733/2} dx_1 + \frac{0,3733}{4} \int_{0,3733/2}^2 \frac{1}{x_1} dx_1 = \\
 &= \frac{1}{2} (x_1)_0^{0,3733/2} + \frac{0,3733}{4} (\ln x_1)_{0,3733/2}^2 = \frac{0,3733}{4} + \frac{0,3733}{4} [\ln 2 - \ln (0,3733 / 2)] = 0,3146
 \end{aligned}$$

c) En la muestra aleatoria simple, los volúmenes de ventas fueron de 50.000 y 100.000 euros, en consecuencia $x_1 = 0,5$ y $x_2 = 1$, por tanto, $x_1 x_2 = 0,5$

Como la región crítica es $x_1 x_2 \leq 0,3733$, se acepta la hipótesis nula.

142. El peso (en gramos) de las cajas de cereales de una determinada marca sigue una distribución $N(\mu, 5)$. Se han tomado los pesos de 16 cajas seleccionadas aleatoriamente, y los resultados obtenidos han sido:

506, 508, 499, 503, 504, 510, 497, 512, 514, 505, 493, 496, 506, 502, 509, 496.

- Obtener el intervalo de confianza del 95% para la media poblacional.
- Determinar cuál sería el tamaño muestral necesario para conseguir, con un 95% de confianza, un intervalo de longitud igual a 2 gramos.
- Suponiendo ahora que σ es desconocida, calcular el intervalo de confianza para la media al 99%.

Solución:

a) Se trata de construir un intervalo de confianza para la media poblacional μ de varianza conocida $\sigma^2 = 25$. El intervalo de confianza de nivel $(1 - \alpha)$ viene dado por:

$$\begin{aligned}
 I_{1-\alpha}(\mu) &= \left[\overbrace{\bar{x}}^{\text{Media muestral}} \pm \overbrace{z_{\alpha/2} \frac{\sigma}{\sqrt{n}}}^{\text{Error estimación}} \right] \quad L_{1-\alpha} = 2 z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \rightarrow n = \left(\frac{2 z_{\alpha/2} \sigma}{\text{longitud}} \right)^2 \quad L \equiv \text{longitud o amplitud} \\
 \text{Error estimación} &\equiv z_{\alpha/2} \frac{\sigma}{\sqrt{n}}
 \end{aligned}$$

$$\bar{x} = \frac{\sum_{i=1}^{16} x_i}{16} = 503,75 \quad 1 - \alpha = 0,95 \quad \alpha = 0,05 \quad \alpha / 2 = 0,025 \quad z_{\alpha/2} = 1,96$$

Intervalo de confianza:

$$I_{0,95}(\mu) = \left[503,75 \pm 1,96 \frac{5}{\sqrt{16}} \right] = \left[503,75 - 1,96 \frac{5}{\sqrt{16}}, 503,75 + 1,96 \frac{5}{\sqrt{16}} \right] = P[501,30 \leq \mu \leq 506,20]$$

$$b) \left(\begin{array}{c} \text{amplitud o} \\ \text{longitud} \end{array} \right) = \left(\bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) - \left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right) = 2 z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \rightarrow n = \left(\frac{2 z_{\alpha/2} \sigma}{\text{amplitud}} \right)^2$$

$$n = \left(\frac{2 \cdot 1,96 \cdot 5}{2} \right)^2 \approx 96 \text{ cajas de cereales}$$

c) Se trata de construir un intervalo de confianza para la media poblacional μ de varianza poblacional desconocida, con muestras pequeñas $\mu \leq 30$.

$$I_{1-\alpha}(\mu) = \left[\bar{x} \pm t_{\alpha/2, n-1} \frac{s_x}{\sqrt{n}} \right] \quad \alpha = 0,01 \quad t_{0,005, 15} = 2,947$$

$$\text{Cuasivarianza muestral: } s_x^2 = \frac{\sum_{i=1}^{16} (x_i - \bar{x})^2}{15} = 36,037 \rightarrow \text{cuasidesviación típica} \equiv s_x \approx 6$$

$$I_{0,99}(\mu) = \left[503,75 \pm 2,947 \frac{6}{\sqrt{16}} \right] = \left[503,75 - 2,947 \frac{6}{\sqrt{16}}, 503,75 + 2,947 \frac{6}{\sqrt{16}} \right] = [499,33, 508,17]$$

143. El gerente de un centro comercial de Madrid desea estimar la cantidad media que gastan los clientes que visitan el centro comercial. Sabiendo que el gasto sigue una distribución normal, una muestra de 20 clientes señala las siguientes cantidades en euros:

37,9 46,9 41,8 46,8 49,2 52,7 50,8
52,6 61,8 23,7 42,2 61,5 58,8 54,8
48,5 61,9 51,4 48,1 51,3 43,9

Determinar un intervalo de confianza del 95% e interpretar el resultado. ¿Se podría afirmar que la media poblacional es de 50 o 60 euros?

Solución:

Intervalo de confianza para la media μ de una distribución normal $N(\mu, \sigma)$ de varianza

$$\text{desconocida } \sigma^2: I_{1-\alpha}(\mu) = \left[\bar{x} \pm t_{\alpha/2, n-1} \frac{s_x}{\sqrt{n}} \right]$$

$$\bar{x} = \alpha_1 = \frac{1}{20} \sum_{i=1}^{20} x_i = \frac{986,6}{20} = 49,33 \quad \alpha_2 = \frac{1}{20} \sum_{i=1}^{20} x_i^2 = \frac{50.221,66}{20} = 2511,083$$

$$\sigma_x^2 = \alpha_2 - \alpha_1^2 = 2511,083 - 49,33^2 = 77,6341$$

$$n \cdot \sigma_x^2 = (n-1) \cdot s_x^2 \rightarrow s_x^2 = \frac{n \cdot \sigma_x^2}{(n-1)} = \frac{20 \cdot 77,6341}{19} = 81,72$$

$$\text{O bien, } s_x^2 = \frac{1}{19} \sum_{i=1}^{20} (x_i - \bar{x})^2 = \frac{1552,682}{19} = 81,72 \quad s_x = \sqrt{81,72} = 9,04 \quad t_{0,025,19} = 2,093$$

$$I_{0,95}(\mu) = \left[49,33 \pm 2,093 \frac{9,04}{\sqrt{20}} \right] = [49,33 \pm 4,23] = [45,1, 53,56]$$

Como 50 euros se encuentra dentro del intervalo de confianza, resulta razonable que la media poblacional sea de 50 euros.

El valor de 60 euros no se encuentra en el intervalo de confianza, se concluye que no es probable que la media poblacional sea de 50 euros.

144. Un fabricante de televisores está desarrollando un nuevo modelo de televisor en color, y para este fin se pueden utilizar dos tipos de esquemas transistorizados. El fabricante selecciona una muestra de esquemas transistorizados del primer tipo de tamaño 16, y otra del segundo tipo de tamaño 13. Los datos muestrales respecto a la vida media de cada esquema son los

$$\text{siguientes: } \begin{cases} \bar{x}_1 = 1400 \text{ horas} & s_1 = 30 \text{ horas} & n_1 = 16 \\ \bar{x}_2 = 1500 \text{ horas} & s_2 = 17 \text{ horas} & n_2 = 13 \end{cases}$$

Construir un intervalo de confianza del 90% para la diferencia de vida media de cada tipo de esquema.

Solución:

Sea la variable aleatoria $X_1 =$ 'Vida media del primer esquema', que sigue una distribución normal $N(\mu_1, \sigma_1)$. Análogamente, la variable aleatoria $X_2 =$ 'Vida media del segundo esquema', sigue una distribución normal $N(\mu_2, \sigma_2)$.

Se quiere construir un intervalo de confianza para la diferencia de medias poblacionales $(\mu_1 - \mu_2)$ con varianzas poblacionales desconocidas, y no se sabe si distintas o no, siendo las muestras pequeñas $n_1 + n_2 = 29 < 30$.

Para dilucidar si las varianzas poblacionales desconocidas son o no distintas, se construye primero un intervalo de confianza para el cociente de varianzas $(\sigma_1^2 / \sigma_2^2)$, de modo que si el intervalo cubre al punto 1 se puede partir de que las varianzas son desconocidas pero iguales.

- Para construir un intervalo de confianza para el cociente de varianzas se emplea la fórmula:

$$I_{1-\alpha} \left(\frac{\sigma_1^2}{\sigma_2^2} \right) = \left[\frac{s_1^2/s_2^2}{F_{\alpha/2; (n_1-1), (n_2-1)}}, \frac{s_1^2/s_2^2}{F_{1-\alpha/2; (n_1-1), (n_2-1)}} \right]$$

$$\text{siendo } F_{1-\alpha/2; (n_1-1), (n_2-1)} = \frac{1}{F_{\alpha/2; (n_1-1), (n_2-1)}}$$

$$n_1 = 16 \quad n_2 = 13 \quad s_1^2 = 30^2 = 900 \quad s_2^2 = 17^2 = 289 \quad s_1^2/s_2^2 = 900/289 = 3,114$$

$$1-\alpha = 0,90 \quad \alpha/2 = 0,05 \quad F_{0,05; 15, 12} = 2,6169 \quad F_{0,95; 15, 12} = 1/F_{0,05; 12, 15} = 1/2,4753 = 0,404$$

$$\text{de donde, } I_{0,90} \left(\frac{\sigma_1^2}{\sigma_2^2} \right) = \left[\frac{3,114}{2,6169}, \frac{3,114}{0,404} \right] = [1,19, 7,71]$$

El intervalo no cubre el 1, y se concluye que las varianzas poblacionales son desconocidas y distintas, con una fiabilidad del 90%.

Se trata de un intervalo de confianza para la diferencia de medias poblacionales $(\mu_1 - \mu_2)$ con varianzas poblacionales desconocidas y distintas o no, con muestras pequeñas $n_1 + n_2 < 30$.

$$I_{1-\alpha}(\mu_1 - \mu_2) = \left[(\bar{x} - \bar{y}) \pm t_{\alpha/2, f} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

$$\text{donde, } f \text{ es la aproximación de Welch } \equiv f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1 + 1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2 + 1}} - 2$$

$$n_1 = 16 \quad n_2 = 13 \quad s_1^2 = 30^2 = 900 \quad s_2^2 = 17^2 = 289$$

$$s_1^2/n_1 = 900/16 = 56,25 \quad s_2^2/n_2 = 289/13 = 22,23$$

se tiene:

$$\left(s_1^2/n_1 \right)^2 = 3164,06 \quad \left(s_2^2/n_2 \right)^2 = 494,17 \quad \left(s_1^2/n_1 \right)^2 / 17 = 186,12 \quad \left(s_2^2/n_2 \right)^2 / 14 = 13,294$$

$$\left(s_1^2/n_1 + s_2^2/n_2 \right)^2 = 6159,11$$

$$f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right)^2}{\frac{\left(\frac{s_1^2}{n_1} \right)^2}{n_1 + 1} + \frac{\left(\frac{s_2^2}{n_2} \right)^2}{n_2 + 1}} - 2 = \frac{6159,11}{186,12 + 13,294} - 2 = 28,89 \approx 29$$

$$1-\alpha = 0,90 \quad \alpha/2 = 0,05 \quad t_{\alpha/2, f} = t_{0,05, 29} = 1,699 \quad \bar{x}_1 = 1400 \text{ horas} \quad \bar{x}_2 = 1500 \text{ horas}$$

$$I_{0,90}(\mu_1 - \mu_2) = \left[(1400 - 1500) \pm 1,699 \sqrt{\frac{900}{16} + \frac{289}{13}} \right] = [-115,05, -84,95]$$

El intervalo no cubre el 0, concluyendo que existe diferencia significativa entre la vida media de cada esquema, siendo mayor la vida media del segundo esquema con una fiabilidad del 90%.

145. Un instituto de investigaciones agronómicas siembra, en cinco parcelas diferentes, dos tipos de maíz híbrido. Las producciones en quintales métricos por hectárea son:

Híbrido I	90	85	95	76	80
Híbrido II	84	87	90	92	90

Construir un intervalo de confianza del 90% para la diferencia entre las producciones medias.

Solución:

Sea la variable aleatoria X_1 = 'Producción de maíz del híbrido I', que sigue una distribución normal $N(\mu_1, \sigma_1)$. Análogamente, la variable aleatoria X_2 = 'Producción de maíz del híbrido II', sigue una distribución normal $N(\mu_2, \sigma_2)$.

Se construye un intervalo de confianza para el cociente de varianzas para concluir si las varianzas poblacionales desconocidas son o no distintas.

De modo que, si el intervalo de confianza para el cociente de varianzas $(\sigma_1^2 / \sigma_2^2)$ cubre al punto 1 se puede partir de que las varianzas son desconocidas pero iguales.

$$I_{1-\alpha} \left(\frac{\sigma_1^2}{\sigma_2^2} \right) = \left[\frac{s_1^2 / s_2^2}{F_{\alpha/2; (n_1-1), (n_2-1)}}, \frac{s_1^2 / s_2^2}{F_{1-\alpha/2; (n_1-1), (n_2-1)}} \right]$$

$$\text{siendo } F_{1-\alpha/2; (n_1-1), (n_2-1)} = \frac{1}{F_{\alpha/2; (n_1-1), (n_2-1)}}$$

$$\bar{x}_1 = 85,20 \quad s_1^2 = 57,7 \quad n_1 = 5$$

$$\bar{x}_2 = 88,6 \quad s_2^2 = 9,8 \quad n_2 = 5$$

En este caso,

$$s_1^2 / s_2^2 = 57,7 / 9,8 = 5,89 \quad 1 - \alpha = 0,90 \quad \alpha / 2 = 0,05$$

$$F_{0,05; 4, 4} = 6,3883 \quad F_{0,95; 4, 4} = 1 / F_{0,05; 4, 4} = 1 / 6,3883 = 0,1565$$

$$I_{0,90} \left(\frac{\sigma_1^2}{\sigma_2^2} \right) = \left[\frac{5,89}{6,3883}, \frac{5,89}{0,1565} \right] = [0,92, 37,64]$$

El intervalo cubre el 1, y se concluye que las varianzas poblacionales son desconocidas e iguales, con una fiabilidad del 90%.

En consecuencia, se trata de un intervalo de confianza para la diferencia de medias poblacionales $(\mu_1 - \mu_2)$ con varianzas poblacionales desconocidas pero iguales, con muestras pequeñas $n_1 + n_2 = 10 < 30$

$$I_{1-\alpha}(\mu_1 - \mu_2) = \left[(\bar{x} - \bar{y}) \pm t_{\alpha/2, (n_1 + n_2 - 2)} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

donde, $s_p^2 \equiv$ media ponderada de las cuasivarianzas muestrales:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \rightarrow s_p^2 = \frac{4(57,7) + 4(9,8)}{5 + 5 - 2} = 33,75 \rightarrow s_p = 5,81$$

$$\bar{x}_1 = 85,20 \quad \bar{x}_2 = 88,6 \quad n_1 = n_2 = 5$$

$$1 - \alpha = 0,90 \quad \alpha/2 = 0,05 \quad t_{\alpha/2, (n_1 + n_2 - 2)} = t_{0,05, 8} = 1,860$$

$$I_{0,90}(\mu_1 - \mu_2) = \left[(85,20 - 88,6) \pm 1,860 \cdot 5,81 \sqrt{\frac{1}{5} + \frac{1}{5}} \right] = [-10,23, 3,43]$$

El intervalo de confianza cubre el 0, por lo que no existe diferencia significativa entre las producciones medias, con una fiabilidad del 90%.

146. Un equipo de investigación biológica está interesado en ver si una nueva droga reduce el colesterol en la sangre. Con tal fin toma una muestra de diez pacientes y determina el contenido de colesterol en la sangre antes y después del tratamiento. Los datos muestrales expresados en miligramos por 100 mililitros son los siguientes:

Antes	217	252	229	200	209	213	215	260	232	216
Después	209	241	230	208	206	211	209	228	224	203

Construir un intervalo de confianza del 95 por 100 para la diferencia del contenido medio de colesterol en la sangre antes y después del tratamiento.

Solución:

Se trata de datos apareados, en los que no existe independencia entre las muestras.

En este caso, como la muestra es pequeña ($n = 10 < 30$) el intervalo de confianza es:

$$I_{1-\alpha}(\mu_1 - \mu_2) = \left[\bar{d} \pm t_{\alpha/2, (n-1)} \frac{s_d}{\sqrt{n}} \right]$$

donde $d_i = x_i - y_i$ $\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$ $s_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$

siendo $\bar{d}$ la media de las diferencias y s_d la desviación estándar de estas diferencias.

X = 'Antes'	217	252	229	200	209	213	215	260	232	216
Y = 'Después'	209	241	230	208	206	211	209	228	224	203
$d_i = x_i - y_i$	8	11	-1	-8	3	2	6	32	8	13

$\bar{d} = 7,40$ $s_d^2 = 112,1481$ $s_d = 10,59$ $n = 10$
 $1 - \alpha = 0,95$ $\alpha / 2 = 0,025$ $t_{\alpha/2, (n-1)} = t_{0,025, 9} = 2,262$

$$I_{0,95}(\mu_1 - \mu_2) = \left[7,40 \pm 2,262 \frac{10,59}{\sqrt{10}} \right] = [-0,17, 14,97]$$

El intervalo abarca el 0, por lo que no existe diferencia significativa en la diferencia del contenido medio del colesterol antes y después del tratamiento, con una fiabilidad del 95%.

147. Según los dirigentes del partido A, la intención de voto del partido rival B, en Andalucía, es la misma que la que tiene en Madrid. Se realiza una encuesta a 100 personas en Andalucía de los que 25 mostraron su apoyo al partido B, y a otras 100 personas en Madrid de las que 30 se inclinaron por el partido B.

- Construir un intervalo de confianza del 90% para la proporción de personas que votarían al partido B en Andalucía.
- ¿A cuántas personas habría que encuestar para obtener un margen de error o error de estimación $\pm 2\%$, al nivel de confianza anterior?.
- Construir un intervalo de confianza al 90% para la diferencia de proporciones en la estimación del voto del partido B en las dos comunidades. ¿Podemos afirmar que los dirigentes del partido A tienen razón?.

Solución:

a) La característica en estudio en ambas comunidades es dicotómica, tenemos que construir un intervalo de confianza para el parámetro p_1 (proporción) de la variable aleatoria binomial asociada al estudio de la característica en la comunidad de Andalucía.

Como el tamaño de la muestra es suficientemente grande, $n_1 = 100$, se puede utilizar la aproximación normal.

$$I_{1-\alpha}(p) = \left[\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right]$$

donde, $1 - \alpha = 0,90$ $\alpha/2 = 0,05$ $z_{\alpha/2} = z_{0,05} = 1,645$
 $\hat{p}_1 = 25 / 100 = 0,25$ $\hat{q}_1 = 1 - \hat{p}_1 = 0,75$ $n_1 = 100$

$$I_{0,90}(p_1) = \left[0,25 \pm 1,645 \sqrt{\frac{0,25 \cdot 0,75}{100}} \right] = [0,179 ; 0,321]$$

En Andalucía la intención de voto del partido B se encuentra entre el 17,9% y 32,1%, con un nivel de confianza del 90%.

b) La amplitud o longitud vendrá dado por la fórmula:

$$I = \left[\left(\begin{array}{c} \text{proporción} \\ \text{muestral} \end{array} \right) \pm \left(\begin{array}{c} \text{error} \\ \text{estimación} \end{array} \right) \right] \equiv \left[\hat{p}_x + z_{\alpha/2} \sqrt{\frac{\hat{p}_x \cdot \hat{q}_x}{n}} \right] \rightarrow \epsilon^2 = \left(z_{\alpha/2} \sqrt{\frac{\hat{p}_x \cdot \hat{q}_x}{n}} \right)^2$$

de donde, $n = \frac{(z_{\alpha/2})^2 (\hat{p}_x \cdot \hat{q}_x)}{\epsilon^2}$

El caso más desfavorable será cuando $\hat{p}_x = \hat{q}_x = 0,5$

Siendo $\epsilon^2 = (\pm 0,02)^2 = 0,0004 \rightarrow n = \frac{(1,645)^2 (0,5 \cdot 0,5)}{0,0004} \approx 1691$

c) Se trata de un intervalo de confianza para la diferencia de parámetros poblacionales ($p_1 - p_2$) de dos distribuciones binomiales, con el tamaño de las muestras suficientemente grandes, $n_1 = n_2 = 100$, para utilizar la aproximación normal.

$$I_{1-\alpha}(p_1 - p_2) = \left[(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_1 (1 - \hat{p}_1)}{n_1} + \frac{\hat{p}_2 (1 - \hat{p}_2)}{n_2}} \right]$$

$z_{\alpha/2} = z_{0,05} = 1,645$

$n_1 = 100$ $\hat{p}_1 = 25/100 = 0,25$ $\hat{q}_1 = 1 - \hat{p}_1 = 0,75$

$n_2 = 100$ $\hat{p}_2 = 30/100 = 0,3$ $\hat{q}_2 = 1 - \hat{p}_2 = 0,70$

$$I_{0,90}(p_1 - p_2) = \left[(0,25 - 0,3) \pm 1,645 \sqrt{\frac{0,25 \cdot 0,75}{100} + \frac{0,3 \cdot 0,70}{100}} \right] = [-0,153 , 0,053]$$

El intervalo de confianza cubre el 0, lo que indica que no existe diferencia significativa entre la intención de voto del partido B en ambas comunidades, con lo cual los dirigentes del partido A tienen razón con una fiabilidad del 90%.

148. Las puntuaciones de estadística teórica se distribuyen según una ley normal, con μ y σ desconocidas. Se desea contrastar $H_0: \mu = 74$ frente a $H_1: \mu \neq 74$. Para ello, se introducen en SPSS las puntuaciones de quince alumnos escogidos aleatoriamente. Los resultados han sido:

Prueba para una muestra

	Valor de prueba = 74					
	t	gl	Sig. (bilateral)	Diferencia de medias	95% Intervalo de confianza para la diferencia	
					Inferior	Superior
Autónoma	1,923	14	,075	7,200	-,83	15,23

Con un nivel de significación de 0,05, se pide:

- ¿Qué se puede decir del contraste?
- Media y cuasidesviación típica muestral
- Define el p_valor y como calcular la significación bilateral
- Calcula el p_valor al contrastar $H_0: \mu \leq 74$ frente a $H_1: \mu > 74$
- Calcula el p_valor al contrastar $H_0: \mu \geq 74$ frente a $H_1: \mu < 74$

Solución:

- El intervalo de confianza $[-0,83, 15,23]$ cubre el cero, se puede afirmar que la muestra procede de una población $N(74, \sigma)$, con un nivel de significación de 0,05

Por otro lado,

p_valor = Sig.(bilateral) = 0,075 > 0,05 y se acepta la hipótesis nula $H_0: \mu = 74$

$$b) I_{1-\alpha}(\mu) = \left[\bar{x} \pm t_{\alpha/2, (n-1)} \frac{s_x}{\sqrt{n}} \right]$$

Diferencia de medias: $\bar{x} - 74 = 7,2 \rightarrow \bar{x} = 81,2$

t $\equiv$ valor experimental del estadístico de contraste:

$$t = \frac{|\bar{x} - 74|}{s / \sqrt{15}} \rightarrow \frac{s_x}{\sqrt{15}} = \frac{|\bar{x} - 74|}{t} = \frac{7,2}{1,923} = 3,743$$

Error típico de la media: $\frac{s_x}{\sqrt{15}} = 3,743 \rightarrow s_x = 14,496$

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

c) La Significación bilateral (p_{valor}) muestra el grado de compatibilidad entre el valor poblacional propuesto y la información muestral disponible.

Si el nivel crítico es pequeño (generalmente menor que 0,05) se concluye que la información recogida en la muestra es incompatible con la hipótesis nula de que la muestra procede de la población.

$$\alpha_p = p_{\text{valor}} = P[\text{Rechazar } \bar{x} \mid H_0 \text{ es Cierta}]$$

$$\alpha_p = p_{\text{valor}} = P[|t_{14}| > 1,923] = 2 \cdot P[t_{14} > 1,923] = 0,075$$

d) SPSS siempre calcula el p_{valor} para un contraste bilateral o de dos colas, si se desea realizar un contraste unilateral o de una cola hay que dividir entre 2 el contraste bilateral ($p_{\text{valor}} = 0,075 / 2 = 0,0375$)

$$\alpha_p = p_{\text{valor}} = P[t_{14} > 1,923] = 0,0375$$

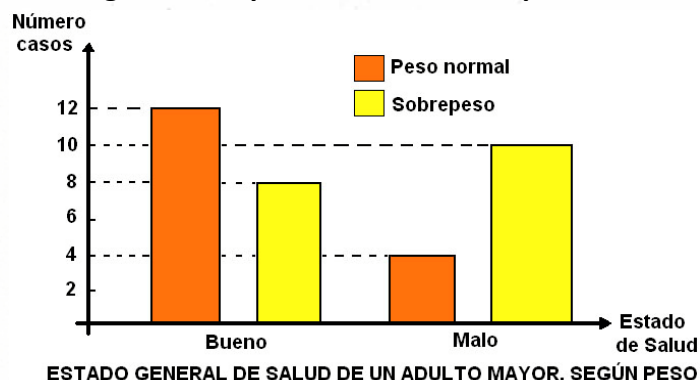
Se trata de un contraste unilateral (cola a la derecha: $H_1: \mu > 74$)

$$p_{\text{valor}} = \text{Sig. (Unilateral derecha)} = 0,0375 < 0,05 \rightarrow \text{Se rechaza } H_0$$

e) Se trata de un contraste unilateral (cola a la izquierda: $H_1: \mu < 74$)

$$p_{\text{valor}} = \text{Sig. (Unilateral izquierda)} = 1 - 0,0375 = 0,9625 > 0,05 \rightarrow \text{Se acepta } H_0$$

149. En el gráfico se presenta la evaluación del estado general de salud de una muestra de personas adultas mayores, según sea su peso normal o sobrepeso.



Analizar la existencia de una relación significativa entre el peso y el estado general de salud en el adulto mayor, con un nivel de significación del 5%.

Solución:

Se trata de dos variables dicotómicas con datos de frecuencia, pudiéndose aplicar una prueba de contraste de asociación con la Chi-cuadrado.

La hipótesis nula H_0 : El estado de salud y el peso son independientes

Llevando la información a una tabla de contingencia de 2x2

Estado de Salud	Peso		$n_{i\cdot}$
	Normal	Sobrepeso	
Buena	12 $e_{11} = 9,41$	8 $e_{12} = 10,59$	20 20
Mala	4 $e_{21} = 6,59$	10 $e_{22} = 7,41$	14 14
$n_{\cdot j}$	16	18	34

La frecuencia observada $n_{21} = 4$ es menor que lo aconsejable en cada celda (≥ 5), lo que podría hacer pensar en una inestabilidad del cálculo.

Como la frecuencia esperada $e_{21} = 6,59$, todas las celdas cumplen con el mínimo aconsejable de 5 en su valor esperado. En la práctica se acepta hasta un 20% de las celdas que no cumplen con el requisito de que la frecuencia esperada sea ≥ 5

Se calculan los valores de χ^2 correspondientes a las dos observaciones, siendo la frecuencia

$$\text{esperada } e_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$$

$$e_{11} = \frac{20 \cdot 16}{34} = 9,41 \quad e_{21} = \frac{14 \cdot 16}{34} = 6,59$$

$$e_{12} = \frac{20 \cdot 18}{34} = 10,59 \quad e_{22} = \frac{14 \cdot 18}{34} = 7,41$$

Estadístico de contraste:

$$\chi^2_{(2-1) \cdot (2-1)} = \chi^2_1 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{n_{ij}^2}{e_{ij}} - n = \frac{12^2}{9,41} + \frac{8^2}{10,59} + \frac{4^2}{6,59} + \frac{10^2}{7,41} - 34 = 3,265$$

Estadístico teórico: $\chi^2_{0,05, 1} = 3,841$

Como $\chi^2_1 = 3,265 < 3,841 = \chi^2_{0,05, 1}$ se acepta la hipótesis nula, concluyendo que el estado general de salud del adulto mayor no está asociado a su peso.

⇒ Adviértase que como la muestra $n < 40$ se hace aconsejable el uso de la Chi-cuadrado con el factor de corrección de continuidad de Yates:

Factor corrección: $n_{ij} < e_{ij} \rightarrow n_{ij} + 0,5 \quad n_{ij} > e_{ij} \rightarrow n_{ij} - 0,5$

Para una tabla de contingencia de 2×2 la corrección de Yates: $\chi^2_1 = \frac{n \left(|n_{11} \cdot n_{22} - n_{12} \cdot n_{21}| - \frac{n}{2} \right)^2}{n_{1\cdot} \cdot n_{2\cdot} \cdot n_{\cdot 1} \cdot n_{\cdot 2}}$

La corrección no es válida cuando $|n_{11} \cdot n_{22} - n_{12} \cdot n_{21}| \leq \frac{n}{2}$

En general, la corrección de Yates se hace cuando el número de grados de libertad es 1.

$$\text{En este caso, } \chi^2_1 = \frac{34 \times \left(|12 \times 10 - 8 \times 4| - \frac{34}{2} \right)^2}{20 \times 14 \times 16 \times 18} = 2,13$$

Como $\chi^2_1 = 2,13 < 3,841 = \chi^2_{0,05, 1}$ se acepta la hipótesis nula.

La validez del contraste también se puede hacer con el p_valor: $\alpha_p = P(\chi^2_{p, 1} > 2,13) = 0,271$

0,90	α_p	0,10
0,0158	2,13	2,706

$$0,90 - 0,10 \longrightarrow 0,0158 - 2,706$$

$$\alpha_p - 0,10 \longrightarrow 2,13 - 2,706$$

$$(\alpha_p - 0,10) \times (0,0158 - 2,706) = (0,90 - 0,10) \times (2,13 - 2,706) \rightarrow \alpha_p = 0,271$$

Al ser $\alpha_p = 0,271 > 0,05 = \alpha$ se acepta la hipótesis nula, afirmando que el estado general de salud del adulto mayor es independiente de su peso.

150. El número de defectos en las tarjetas de circuito impreso sigue una distribución Poisson. Se reúne una muestra aleatoria de 60 tarjetas de circuito impreso y se observa el número de defectos. Los resultados obtenidos son los siguientes:

Número de defectos	0	1	2	3 o más
Frecuencia observada	32	15	9	4

¿Muestran los datos suficiente evidencia para decir que provienen de una distribución de Poisson con una fiabilidad del 95%?

Solución:

H_0 : Los datos provienen de una distribución de Poisson $P(\lambda = \bar{x})$

H_1 : Los datos no provienen de una distribución de Poisson

La distribución de Poisson se caracteriza porque sólo depende de un parámetro λ que coincide con la media. En este sentido, para ajustar la distribución observada a la de Poisson es necesario calcular la media.

Denotando por X = 'Número de defectos en tarjetas impresas' y n_i = 'Frecuencia observada', se tiene:

x_i	0	1	2	3 o más
n_i	32	15	9	4
$x_i \cdot n_i$	0	15	18	12

$$\bar{x} = \frac{1}{n} \sum_{i=1}^4 x_i \cdot n_i = \frac{45}{60} = 0,75 \rightarrow \lambda = 0,75: P(x=k) = \frac{0,75^k}{k!} e^{-0,75} \quad k = 0, 1, 2, 3$$

$$P(x=0) = \frac{0,75^0}{0!} e^{-0,75} = 0,472 \quad P(x=1) = \frac{0,75}{1!} e^{-0,75} = 0,354$$

$$P(x=2) = \frac{0,75^2}{2!} e^{-0,75} = 0,133 \quad P(x \geq 3) = 1 - P(x < 3) = 1 - [0,472 + 0,354 + 0,133] = 0,041$$

x_i	n_i	p_i	$e_i = n \cdot p_i$
0	32	0,472	28,32
1	15	0,354	21,24
2	9	0,133	7,98
3 o más	4	0,041	2,46
	60		

Es necesario que las frecuencias esperadas de las distintas modalidades $e_i \geq 5 \quad \forall i$, teniendo que agrupar las dos últimas modalidades.

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

x_i	n_i	$e_i = n \cdot p_i$	n_i^2	$\frac{n_i^2}{e_i}$
0	32	28,32	1024	36,1582
1	15	21,24	225	10,5932
2 o más	13	10,44	169	16,1877
	60			62,9392

x_i	0	1	2 o más
n_i	32 ($e_1 = 28,32$)	15 ($e_2 = 21,24$)	13 ($e_3 = 10,44$)

Los grados de libertad son $(k - p - 1) = 3 - 1 - 1 = 1$

Estadístico observado: $\chi^2_{3-1-1} = \sum_{i=1}^3 \frac{n_i^2}{e_i} - n = \frac{1024}{28,32} + \frac{225}{21,24} + \frac{169}{10,44} - 60 = 2,9392$

Estadístico teórico: $\chi^2_{0,05, 1} = 3,841$

Como $\chi^2_1 = 2,9392 < \chi^2_{0,05, 1} = 3,841$ se acepta la hipótesis nula, concluyendo que los datos proceden de una población de Poisson de parámetro $\lambda = 0,75$, a un nivel de significación de 0,05

151. Para comprobar si los operarios encontraban dificultades con una prensa manual de imprimir, se hizo una prueba a cuatro operarios anotando el número de atascos sufridos al introducir el mismo número de hojas, dando lugar a la siguiente tabla:

Operario	A	B	C	D	Total
Obstrucciones	6	7	9	18	40

Con un nivel de significación del 5%, ¿existe diferencia entre los operarios?

Solución:

Estableciendo la hipótesis nula $H_0 \equiv$ 'No existe diferencia entre los operarios'

La probabilidad de que se atascase una hoja sería $1/4$ para todos los operarios.

De este modo, el número de atascos esperados para cada uno de ellos sería $(e_i = 10)_{i=1, \dots, 4}$

Sea la tabla de contingencia 1×4 :

Operario	A	B	C	D	Total
Obstrucciones	6 (10)	7 (10)	9 (10)	18 (10)	40 (40)

A un nivel de significación $\alpha = 0,05$ se acepta la hipótesis nula sí

$$\chi^2_{k-1} = \underbrace{\sum_{i=1}^k \frac{(n_i - e_i)^2}{e_i}}_{\text{estadístico contraste}} = \sum_{i=1}^k \frac{n_i^2}{e_i} - n < \underbrace{\chi^2_{\alpha, k-1}}_{\text{estadístico teórico}} \quad \begin{array}{l} k \equiv \text{Número intervalos} \\ p \equiv \text{Número parámetros a estimar} \end{array}$$

$$\text{O bien, la región de rechazo de la hipótesis nula: } R = \left\{ \sum_{i=1}^k \frac{(n_i - e_i)^2}{e_i} \geq \chi^2_{\alpha, k-1} \right\}$$

$$\text{con lo cual, } \chi^2_3 = \sum_{i=1}^4 \frac{n_i^2}{e_i} - n = \frac{6^2}{10} + \frac{7^2}{10} + \frac{9^2}{10} + \frac{18^2}{10} - 40 = 9$$

Siendo $\chi^2_3 = 9 > 7,815 = \chi^2_{0,05, 3}$ se verifica la región de rechazo.

En consecuencia, se concluye que existe diferencia significativa entre los operarios respecto al número de atascos en la prensa manual de imprimir.

152. La tabla refleja el número de accidentes mortales de tráfico que se producen en una carretera a lo largo de un período de tiempo.

Accidentes mortales por día	0	1	2	3	4	5
Número de días	132	195	120	60	24	9

¿Se ajustan los datos a una distribución de Poisson?. Utilizar un nivel de significación 0,05

Solución:

Hipótesis nula $H_0 \equiv$ La distribución empírica se ajusta a la distribución de Poisson

La hipótesis nula se acepta a un nivel de significación α sí

$$\chi^2_{k-1} = \underbrace{\sum_{i=1}^k \frac{(n_i - e_i)^2}{e_i}}_{\text{estadístico contraste}} = \sum_{i=1}^k \frac{n_i^2}{e_i} - n < \underbrace{\chi^2_{\alpha, k-1}}_{\text{estadístico teórico}}$$

$k \equiv$ Número intervalos
 $p \equiv$ Número parámetros a estimar

La distribución de Poisson se caracteriza porque sólo depende del parámetro λ que coincide con la media.

Sea la variable aleatoria $X =$ "Número de accidentes mortales por día" y $n_i \equiv$ Número de días

x_i	n_i	$x_i n_i$	$P(x_i = k) = p_i$
0	132	0	0,2465
1	195	195	0,3451
2	120	240	0,2415
3	60	180	0,1127
4	24	96	0,0394
5	9	45	0,0110
$n = 540$		756	

$$\bar{x} = \lambda = \frac{1}{6} \sum_{i=1}^6 x_i n_i = \frac{756}{540} = 1,4$$

$$P(x_i = k) = \frac{1,4^k}{k!} e^{-1,4} \quad k = 0, \dots, 5$$

Las probabilidades con que llegan las partículas $k = 0, \dots, 5$ se obtienen sustituyendo los

valores de k en $P(x_i = k) = \frac{1,4^k}{k!} e^{-1,4}$

Para verificar si el ajuste de los datos a una distribución de Poisson se acepta o no, mediante una χ^2 , hay que calcular las frecuencias esperadas ($e_i = n \cdot p_i$)

x_i	0	1	2	3	4	5
Frecuencias	132 $e_1 = 133,1$	195 $e_2 = 186,3$	120 $e_3 = 130,4$	60 $e_4 = 60,8$	24 $e_5 = 21,3$	9 $e_6 = 5,9$

$$e_1 = 540 \cdot 0,2465 = 133,1 \quad e_2 = 540 \cdot 0,3451 = 186,3 \quad e_3 = 540 \cdot 0,2415 = 130,4$$

$$e_4 = 540 \cdot 0,1127 = 60,8 \quad e_5 = 540 \cdot 0,0394 = 21,3 \quad e_6 = 540 \cdot 0,0110 = 5,9$$

Dando lugar a una tabla de contingencia 1 x 6, no teniendo que agrupar columnas contiguas al no aparecer frecuencias esperadas menor que cinco.

Los grados de libertad son cuatro: $k - p - 1 = 6 - 1 - 1 = 4$

Estadístico de contraste:
$$\chi^2_3 = \sum_{i=1}^6 \frac{(n_i - e_i)^2}{e_i} = \sum_{i=1}^6 \frac{n_i^2}{e_i} - n =$$

$$= \frac{132^2}{133,1} + \frac{195^2}{186,3} + \frac{120^2}{130,4} + \frac{60^2}{60,8} + \frac{24^2}{21,3} + \frac{9^2}{5,9} - 540 = 5,41$$

Estadístico teórico: $\chi^2_{0,05,4} = 9,488$

El estadístico de contraste (bondad de ajuste) $\chi^2_3 = 5,41$ es menor que el estadístico teórico

$\chi^2_{0,05,4} = 9,488$, por lo que se acepta la hipótesis nula.

Con un nivel de significación 0,05, se puede afirmar que los accidentes mortales de tráfico en la carretera se ajustan a una distribución de Poisson.

153. Para curar cierta enfermedad se sabe que existen cuatro tratamientos diferentes. Aplicados por separado a un grupo distinto de enfermos, se han observado los siguientes resultados:

Enfermo Tratamientos	Curados	No curados	Total
A	60	23	83
B	46	10	56
C	70	17	87
D	54	30	84

¿Se puede considerar que la eficacia de los cuatro tratamientos es la misma con un nivel de confianza del 95 por 100?

Solución 1:

Se trata de un contraste de homogeneidad de cuatro muestras, con 83, 56, 87 y 84 personas, de las cuales hay, respectivamente, 60, 46, 70 y 54 personas curadas.

Se establece la hipótesis nula

H_0 : Los cuatro tratamientos muestrales son de la misma eficacia

Con lo que la proporción de personas curadas en cada muestra es $p = 230 / 310 = 0,742$ y las no curadas $q = 1 - 0,742 = 0,258$, donde $e_{i\cdot} = n_{i\cdot} \times p$

Enfermo Tratamientos	Curados n_{i1}	No curados	$n_{i\cdot}$
A	60	23	83 $e_{1\cdot} = 61,58$
B	46	10	56 $e_{2\cdot} = 41,55$
C	70	17	87 $e_{3\cdot} = 64,55$
D	54	30	84 $e_{4\cdot} = 62,32$

$$e_{1\cdot} = 83 \cdot 0,742 = 61,58 \quad e_{2\cdot} = 56 \cdot 0,742 = 41,55 \quad e_{3\cdot} = 87 \cdot 0,742 = 64,55 \quad e_{4\cdot} = 84 \cdot 0,742 = 62,32$$

$$\sum_{i=1}^4 \frac{(n_{i1} - e_{i\cdot})^2}{n_{i\cdot}} = \frac{(60 - 61,58)^2}{83} + \frac{(46 - 41,55)^2}{56} + \frac{(70 - 64,55)^2}{87} + \frac{(54 - 62,32)^2}{84} = 1,55$$

$$\chi^2_{4-1} = \left(\frac{1}{p \times q} \right) \sum_{i=1}^4 \frac{(n_{i1} - e_{i\cdot})^2}{n_{i\cdot}} = \left(\frac{1}{0,742 \times 0,258} \right) \times 1,55 = 8,09$$

Siendo $\chi^2_3 = 8,09 > 7,815 = \chi^2_{0,05, 3}$ se rechaza la hipótesis nula. Los tratamientos a efectos de curar a los pacientes tienen diferente eficacia, con una fiabilidad del 95%.

Solución 2:

Se establece la hipótesis nula

H_0 : Los cuatro tratamientos muestrales son de la misma eficacia

Se elabora una tabla de contingencia 4×2 , con una frecuencia teórica $e_{ij} = p_{ij} \cdot n = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$

Enfermo Tratamientos	Curados n_{i1}	No curados	$n_{i\cdot}$
A	60 $e_{11} = 61,58$	23 $e_{12} = 21,42$	83
B	46 $e_{21} = 41,55$	10 $e_{22} = 14,45$	56
C	70 $e_{31} = 64,55$	17 $e_{32} = 22,45$	87
D	54 $e_{41} = 62,32$	30 $e_{42} = 21,68$	84
$n_{\cdot j}$	230	80	310

$$e_{11} = \frac{83 \times 230}{310} = 61,58 \quad e_{21} = \frac{56 \times 230}{310} = 41,55 \quad e_{31} = \frac{87 \times 230}{310} = 64,55 \quad e_{41} = \frac{84 \times 230}{310} = 62,32$$

$$e_{12} = \frac{83 \times 80}{310} = 21,42 \quad e_{22} = \frac{56 \times 80}{310} = 14,45 \quad e_{32} = \frac{87 \times 80}{310} = 22,45 \quad e_{42} = \frac{84 \times 80}{310} = 21,68$$

$$\chi^2_{(4-1) \cdot (2-1)} = \sum_{i=1}^4 \sum_{j=1}^2 \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \frac{(60 - 61,58)^2}{61,58} + \frac{(23 - 21,42)^2}{21,42} + \frac{(46 - 41,55)^2}{41,55} +$$

$$+ \frac{(10 - 14,45)^2}{14,45} + \frac{(70 - 64,55)^2}{64,55} + \frac{(17 - 22,45)^2}{22,45} + \frac{(54 - 62,32)^2}{62,32} + \frac{(30 - 21,68)^2}{21,68} = 8,09$$

La expresión anterior se podía haber realizado de forma más sencilla con la igualdad:

$$\chi^2_{(4-1) \cdot (2-1)} = \sum_{i=1}^4 \sum_{j=1}^2 \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^4 \sum_{j=1}^2 \frac{n_{ij}^2}{e_{ij}} - n = \frac{60^2}{61,58} + \frac{23^2}{21,42} + \frac{46^2}{41,55} + \frac{10^2}{14,45} +$$

$$+ \frac{70^2}{64,55} + \frac{17^2}{22,45} + \frac{54^2}{62,32} + \frac{30^2}{21,68} - 310 = 8,09$$

Siendo $\chi^2_3 = 8,09 > 7,815 = \chi^2_{0,05,3}$ se rechaza la hipótesis nula, es decir, los tratamientos a efectos de curar a los pacientes tienen diferente eficacia, a un nivel de significación $\alpha = 0,05$

154. Un experimento para investigar el efecto de vacunación de animales de laboratorio refleja la siguiente tabla:

Vacuna	Animal Laboratorio	
	Enfermo	No Enfermo
Vacunado	9	42
No Vacunado	18	28

Con un nivel de significación de 0,05, ¿Es conveniente vacunar?.

Solución:

Hipótesis nula H_0 : Es independiente la vacuna de los animales enfermos

Vacuna	Animal laboratorio		Total
	Enfermo	No Enfermo	
Vacunado	9	42	51
No Vacunado	18	28	46
Total	27	70	97

En una tabla de contingencia de 2×2 se puede calcular la χ^2 de una forma sencilla recurriendo a las frecuencias observadas.

$$\text{Estadístico observado: } \chi^2_1 = \frac{n \cdot (n_{11} \cdot n_{22} - n_{12} \cdot n_{21})^2}{n_{1\cdot} \cdot n_{2\cdot} \cdot n_{\cdot 1} \cdot n_{\cdot 2}} = \frac{97 \cdot (9 \cdot 28 - 42 \cdot 18)^2}{51 \cdot 46 \cdot 27 \cdot 70} = 5,5570$$

El número de grados de libertad es $(2 - 1) \times (2 - 1) = 1$

$$\text{Estadístico teórico: } \chi^2_{0,05,1} = 3,841$$

Siendo $\chi^2_1 = 5,5570 > 3,841 = \chi^2_{0,05,1}$ se rechaza la hipótesis nula, es decir, la vacuna afecta a la enfermedad con un nivel de significación $\alpha = 0,05$

155. Para analizar la repercusión que tienen los debates televisivos en la intención de voto, un equipo de investigación recogió datos entre 240 individuos antes y después del debate, resultando la siguiente tabla:

Antes del debate (candidatos)	Después del debate (candidatos)		Total
	A	B	
A	46 (a)	50 (b)	96 (a + b)
B	85 (c)	59 (d)	144 (c + d)
Total	131 (a + c)	109 (b + d)	240 (n)

Se desea saber si el debate televisivo cambió la intención de voto, con un nivel de significación del 5%.

Solución:

Se trata de una muestra pareada en una situación antes-después, con lo que es idóneo un contraste estadístico Chi-cuadrado de McNemar.

Sea la hipótesis nula H_0 : La intención de voto es la misma antes y después del debate

$$\text{Estadístico muestral: } \chi^2_{\text{McNemar}} = \frac{(c - b)^2}{c + b} = \frac{(85 - 50)^2}{85 + 50} = 9,074$$

$$\text{Estadístico teórico: } \chi^2_{\alpha/2, 1} = \chi^2_{0,025, 1} = 5,024$$

Como $\chi^2_{\text{McNemar}} = 9,074 > 5,024 = \chi^2_{0,025, 1}$ se rechaza la hipótesis nula con un nivel de significación del 5%, concluyendo que la intención de voto cambió significativamente después del debate.

156. Se desea analizar si los estudiantes de universidades privadas preferentemente son de los estratos económicos altos del país. Para ello, se ha tomado la siguiente muestra:

Universidades	Grupos socioeconómicos			
	Alto	Medio alto	Medio bajo	Bajo
Estado	13	17	4	3
Privadas	38	19	2	2

a) Para validar el análisis con un nivel de confianza del 95%, realizar un contraste por la razón de verosimilitud (test G).

b) Estudiar el grado de dependencia entre el tipo de universidad y el estrato socioeconómico.

Solución

a) El test de contraste de independencias por la razón de verosimilitudes (test G) es una prueba de hipótesis de la Chi-cuadrado que presenta mejores resultados que el de Pearson.

Se distribuye asintóticamente como una variable aleatoria χ^2 con $(k-1) \cdot (m-1)$ grados de libertad.

Se define el estadístico: $G = 2 \sum_{i=1}^k \sum_{j=1}^m n_{ij} \ln \left(\frac{n_{ij}}{e_{ij}} \right)$

Se acepta la hipótesis nula H_0 si $G = 2 \sum_{i=1}^k \sum_{j=1}^m n_{ij} \ln \left(\frac{n_{ij}}{e_{ij}} \right) < \chi_{\alpha, (k-1) \cdot (m-1)}^2$

Hipótesis nula H_0 : El tipo de universidad es independiente del grupo socioeconómico

En un principio, la tabla presenta un 50% de celdas que no verifican que las frecuencias sean mayores que 5, teniendo que agrupar modalidades contiguas en una sola hasta lograr que la nueva frecuencia sea mayor que cinco.

Universidades	Grupos socioeconómicos		
	Alto	Medio alto	Medio bajo - Bajo
Estado	13	17	7
Privadas	38	19	4

Se calculan los valores esperados de cada celda, donde $e_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$

Universidades	Grupos socioeconómicos			$n_{i\cdot}$
	Alto	Medio alto	Medio bajo - Bajo	
Estado	13 $e_{11} = 19,26$	17 $e_{12} = 13,59$	7 $e_{13} = 4,15$	37 37
Privadas	38 $e_{21} = 31,74$	19 $e_{22} = 22,41$	4 $e_{23} = 6,85$	61 61
$n_{\cdot j}$	51	36	11	$n = 98$

$$e_{11} = \frac{37 \cdot 51}{98} = 19,26$$

$$e_{12} = \frac{37 \cdot 36}{98} = 13,59$$

$$e_{13} = \frac{37 \cdot 11}{98} = 4,15$$

$$e_{21} = \frac{61 \cdot 51}{98} = 31,74$$

$$e_{22} = \frac{61 \cdot 36}{98} = 22,41$$

$$e_{23} = \frac{61 \cdot 11}{98} = 6,85$$

La frecuencia esperada $e_{13} = 4,15 < 5$, valor mínimo recomendado para la prueba. En un caso práctico se admite hasta un 20% de las celdas que no verifican este requisito, como ocurre en este caso.

En cada celda se calcula el valor de $n_{ij} \times \ln\left(\frac{n_{ij}}{e_{ij}}\right)$

Universidades	Grupos socioeconómicos			Total
	Alto	Medio alto	Medio bajo - Bajo	
Estado	-5,11	3,80	3,66	2,35
Privadas	6,84	-3,14	-2,15	1,55
Total	1,73	0,66	1,51	3,9

$$13 \times \ln\left(\frac{13}{19,26}\right) = -5,11 \quad 17 \times \ln\left(\frac{17}{13,59}\right) = 3,80 \quad 7 \times \ln\left(\frac{7}{4,15}\right) = 3,66$$

$$38 \times \ln\left(\frac{38}{31,74}\right) = 6,84 \quad 19 \times \ln\left(\frac{19}{22,41}\right) = -3,14 \quad 4 \times \ln\left(\frac{4}{6,85}\right) = -2,15$$

Estadístico observado $G = 2 \sum_{i=1}^2 \sum_{j=1}^3 n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right) = 2 \times 3,9 = 7,8$

El número de grados de libertad es $(2-1) \cdot (3-1) = 2$

Estadístico teórico $\chi_{0,05,2}^2 = 5,991$

Como $G = 7,8 > 5,991 = \chi_{0,05,2}^2$ se rechaza la hipótesis nula de independencia, concluyendo que el tipo de Universidad está asociado al grupo socioeconómico.

⇒ La validez del contraste también se puede hacer con el p_valor: $\alpha_p = P(\chi_{p,1}^2 > 7,8) = 0,271$

0,025	α_p	0,02	$0,025 - 0,02 \longrightarrow 7,378 - 7,824$
7,378	7,8	7,824	$\alpha_p - 0,02 \longrightarrow 7,8 - 7,824$

$$(\alpha_p - 0,02) \times (7,378 - 7,824) = (0,025 - 0,02) \times (7,8 - 7,824) \rightarrow \alpha_p = 0,02026$$

Al ser $\alpha_p = 0,02026 < 0,05 = \alpha$ se rechaza la hipótesis nula, afirmando que el tipo de Universidad depende del estrato socioeconómico.

b) El grado de contingencia mide el grado de relación o dependencia: $C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{7,8}{7,8 + 98}} = 0,2715 \rightarrow \text{Existe una dependencia del 27,15\%}.$$

157. En un estudio sobre la opinión de fumar en lugares públicos se realiza una encuesta a 350 personas, obteniendo los siguientes resultados:

	Muy en contra	En contra	A Favor	Muy a favor	$n_{i\cdot}$
Fumador	60	50	20	10	140
No Fumador	10	30	70	100	210
$n_{\cdot j}$	70	80	90	110	350

Con un nivel de significación de 0,05 se desea conocer si existe diferencia de opinión entre fumadores y no fumadores.

Solución:

Se establecen las hipótesis:

H_0 : La opinión es independiente de su condición de fumador o no fumador

H_1 : La opinión no es independiente de su condición de fumador o no fumador

Se acepta H_0 sí: $\chi_c^2 = \overbrace{\sum_{i=1}^2 \sum_{j=1}^4 \frac{(n_{ij} - e_{ij})^2}{e_{ij}}}$ estadístico observado $< \overbrace{\chi_{\alpha, (2-1) \cdot (4-1)}^2}^{\text{estadístico teórico}} = \chi_{0,05,3}^2$

	Opinión				
	Muy en contra	En contra	A Favor	Muy a favor	$n_{i\cdot}$
Fumador	60 $e_{11} = 28$	50 $e_{12} = 32$	20 $e_{13} = 36$	10 $e_{14} = 44$	140 140
No Fumador	10 $e_{21} = 42$	30 $e_{22} = 48$	70 $e_{23} = 54$	100 $e_{24} = 66$	210 210
$n_{\cdot j}$	70	80	90	110	350

$$e_{11} = \frac{140 \cdot 70}{350} = 28 \quad e_{12} = \frac{140 \cdot 80}{350} = 32 \quad e_{13} = \frac{140 \cdot 90}{350} = 36 \quad e_{14} = \frac{140 \cdot 110}{350} = 44$$

$$e_{21} = \frac{210 \cdot 70}{350} = 42 \quad e_{22} = \frac{210 \cdot 80}{350} = 48 \quad e_{23} = \frac{210 \cdot 90}{350} = 54 \quad e_{24} = \frac{210 \cdot 110}{350} = 66$$

$$\chi_c^2 = \sum_{i=1}^2 \sum_{j=1}^4 \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^2 \sum_{j=1}^4 \frac{n_{ij}^2}{e_{ij}} - n =$$

$$= \frac{60^2}{28} + \frac{50^2}{32} + \frac{20^2}{36} + \frac{10^2}{44} + \frac{10^2}{42} + \frac{30^2}{48} + \frac{70^2}{54} + \frac{100^2}{66} - 350 = 133,46$$

Estadístico teórico: $\chi_{0,05,3}^2 = 7,815$

Siendo $\chi_c^2 = 133,46 > \chi_{0,05,3}^2 = 7,815$ se rechaza la hipótesis nula, se acepta por tanto la hipótesis alternativa, pudiendo afirmar con una significación 0,05 que la opinión sobre el tabaco depende que sea fumador o no.

158. Se ha medido el perímetro craneal a niños de edad comprendida entre los dos y tres años, obteniéndose los siguientes datos en centímetros:

Intervalos	38 – 40	40 – 42	42 – 44	44 – 46	46 – 48
Frecuencia observada	11	7	7	6	4
Frecuencia esperada	5,348		9,2855	5,9255	

Comprobar si las observaciones craneales son o no distribuidas según una ley normal con un nivel de confianza del 99%.

Solución:

El método de aplicación de la prueba de ajuste para la normalidad comienza por hallar la media y desviación típica de los datos observados:

Intervalos	Marca clase x_i	n_i	$x_i n_i$	$x_i^2 n_i$	p_i	$e_i = n p_i$
38 – 40	39	11	429	16731	0,1528	5,348
40 – 42	41	7	287	11767	0,4093	14,3255
42 – 44	43	7	301	12943	0,2653	9,2855
44 – 46	45	6	270	12150	0,1693	5,9255
46 – 48	47	4	188	8836	0,067	2,345
		35	1475	62427		37,2295

$$\alpha_1 = \mu = \frac{1}{n} \sum_{i=1}^5 x_i n_i = \frac{1475}{35} = 42,14 \quad \alpha_2 = \frac{1}{n} \sum_{i=1}^5 x_i^2 n_i = \frac{62427}{35} = 1783,62$$

$$\sigma^2 = \alpha_2 - \alpha_1^2 = 1783,62 - (42,14)^2 = 7,84 \quad \sigma = \sqrt{7,84} = 2,8$$

Considerando una ley normal $N(42,14, 2,8)$ se hallan las probabilidades de cada uno de los intervalos:

$$\begin{aligned} P[38 \leq x < 40] &= P\left[\frac{38 - 42,14}{2,8} \leq \frac{x - 42,14}{2,8} < \frac{40 - 42,14}{2,8}\right] = P[-1,47 \leq z < -0,76] = \\ &= P[0,76 \leq z < 1,47] = 0,2236 - 0,0708 = 0,1528 \end{aligned}$$

$$P[40 \leq x < 42) = P\left[\frac{40 - 42,14}{2,8} \leq \frac{x - 42,14}{2,8} < \frac{42 - 42,14}{2,8}\right] = P[-0,76 \leq z < -0,05) = \\ = P[0,05 \leq z < 0,76) = 0,4801 - 0,0708 = 0,4093$$

$$P[42 \leq x < 44) = P\left[\frac{42 - 42,14}{2,8} \leq \frac{x - 42,14}{2,8} < \frac{44 - 42,14}{2,8}\right] = P[-0,05 \leq z < 0,66) = \\ = P(z \geq -0,05) - P(z \geq 0,66) = P(z \leq 0,05) - P(z \geq 0,66) = 1 - P(z \geq 0,05) - P(z \geq 0,66) = \\ = 1 - 0,4801 - 0,2546 = 0,2653$$

$$P[44 \leq x < 46) = P\left[\frac{44 - 42,14}{2,8} \leq \frac{x - 42,14}{2,8} < \frac{46 - 42,14}{2,8}\right] = P[0,66 \leq z < 1,37) = \\ = 0,2546 - 0,0853 = 0,1693$$

$$P[46 \leq x < 48) = P\left[\frac{46 - 42,14}{2,8} \leq \frac{x - 42,14}{2,8} < \frac{48 - 42,14}{2,8}\right] = P[1,37 \leq z < 2,09) = \\ = 0,0853 - 0,0183 = 0,067$$

Para comprobar si el ajuste se acepta o no mediante una χ^2 hay que encontrar las frecuencias esperadas $e_i = n p_i$. Las frecuencias esperadas de las distintas modalidades no debe ser inferior a cinco.

El intervalo $[46, 48)$ presenta una frecuencia esperada $e_5 = 35 \times 0,067 = 2,345 < 5$, por lo que hay que agrupar dos intervalos contiguos hasta lograr que la frecuencia esperada sea mayor que cinco.

Intervalos	n_i	p_i	$e_i = n p_i$	n_i^2	n_i^2 / e_i
38 - 40	11	0,1528	5,348	121	22,6253
40 - 42	7	0,4093	14,3255	49	3,4205
42 - 44	7	0,2653	9,2855	49	5,2770
44 - 48	10	0,2363	8,2705	100	12,0912
	$n = 35$	1	37,2295		43,4140

Se establece la hipótesis nula H_0 : El perímetro craneal sigue una distribución normal

Que se aceptará cuando el estadístico observado $\chi_c^2 = \sum_{i=1}^5 \frac{n_i^2}{e_i} - n$ es menor que el estadístico teórico o

esperado $\chi_{\alpha, k-p-1}^2$

$k \equiv$ Número de clases o intervalos $p \equiv$ Número de parámetros que se necesitan calcular

El número de intervalos $k = 4$ y el número de parámetros $p = 3$ que ha habido que calcular (n, μ, σ) .

Al ser independientes las modalidades se tiene $\chi_{\alpha, k-p}^2 = \chi_{0,01, 4-3}^2 = \chi_{0,01, 1}^2 = 6,635$

Estadístico observado $\chi_c^2 = \sum_{i=1}^5 \frac{n_i^2}{e_i} - n = 43,4140 - 35 = 8,414$

Se observa que el estadístico de contraste $\chi_c^2 = 8,414 > 6,635 = \chi_{0,01,1}^2$ concluyendo que el perímetro craneal de los niños no sigue una distribución normal, con una fiabilidad del 99%.

159. Para conocer la opinión de los ciudadanos sobre la actuación del alcalde de una determinada ciudad, se realiza una encuesta a 404 personas, cuyos resultados se recogen en la siguiente tabla:

	Desacuerdo	De acuerdo	No contestan
Mujeres	84	78	37
Varones	118	62	25

Contrastar, con un nivel de significación del 5%, que no existen diferencias de opinión entre hombres y mujeres ante la actuación del alcalde.

Solución:

Se trata de un contraste de homogeneidad en el que se desea comprobar si las muestras proceden de poblaciones distintas.

Se tienen dos muestras clasificadas en tres niveles, donde se desea conocer si los hombres y mujeres proceden de la misma población, es decir, si se comportan de manera semejante respecto a la opinión de la actuación del alcalde.

Hipótesis nula: H_0 : 'No existe diferencia entre hombres y mujeres respecto a la opinión'

La hipótesis nula se acepta a un nivel de significación α sí

$$\chi_{(k-1) \cdot (m-1)}^2 = \underbrace{\sum_{i=1}^k \sum_{j=1}^m \frac{(n_{ij} - e_{ij})^2}{e_{ij}}}_{\text{estadístico contraste}} = \sum_{i=1}^k \sum_{j=1}^m \frac{n_{ij}^2}{e_{ij}} - n < \underbrace{\chi_{\alpha, (k-1) \cdot (m-1)}^2}_{\text{estadístico teórico}}$$

$$\text{Región de rechazo: } R_{\text{rechazo}} = \left\{ \chi_{(k-1) \cdot (m-1)}^2 \geq \chi_{\alpha, (k-1) \cdot (m-1)}^2 \right\}$$

Se forma una tabla de contingencia 2 x 3: En cada frecuencia observada $(n_{ij})_{i=1, \dots, k; j=1, \dots, m}$ en la tabla de contingencia se tiene una frecuencia teórica o esperada e_{ij} que se calcula mediante la

expresión: $e_{ij} = p_{ij} \cdot n = \frac{n_{i \cdot} \times n_{\cdot j}}{n}$, donde p_{ij} son las probabilidades de que un elemento tomado de la muestra presente las modalidades x_i de X e y_j de Y.

	Desacuerdo	De acuerdo	No contestan	$n_{i\cdot}$
Mujeres	84 $e_{11} = 99,5$	78 $e_{12} = 68,96$	37 $e_{13} = 30,53$	$199 = n_{x_1}$ 199
Varones	118 $e_{21} = 102,5$	62 $e_{22} = 71,03$	25 $e_{23} = 31,46$	$205 = n_{x_2}$ 205
$n_{\cdot j}$	$202 = n_{y_1}$	$140 = n_{y_2}$	$62 = n_{y_3}$	$n = 404$

$$e_{11} = \frac{199 \cdot 202}{404} = 99,5 \quad e_{12} = \frac{199 \cdot 140}{404} = 68,96 \quad e_{13} = \frac{199 \cdot 62}{404} = 30,53$$

$$e_{21} = \frac{205 \cdot 202}{404} = 102,5 \quad e_{22} = \frac{205 \cdot 140}{404} = 71,03 \quad e_{23} = \frac{205 \cdot 62}{404} = 31,46$$

Estadístico de contraste:

$$\chi^2_{(2-1) \cdot (3-1)} = \chi^2_2 = \sum_{i=1}^2 \sum_{j=1}^3 \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \frac{(84 - 99,5)^2}{99,5} + \frac{(78 - 68,96)^2}{68,96} + \frac{(37 - 30,53)^2}{30,53} + \frac{(118 - 102,5)^2}{102,5} + \frac{(62 - 71,03)^2}{71,03} + \frac{(25 - 31,46)^2}{31,46} = 9,76$$

O bien,

$$\chi^2_2 = \sum_{i=1}^2 \sum_{j=1}^3 \frac{n_{ij}^2}{e_{ij}} - n = \frac{84^2}{99,5} + \frac{78^2}{68,96} + \frac{37^2}{30,53} + \frac{118^2}{102,5} + \frac{62^2}{71,03} + \frac{25^2}{31,46} - 404 = 9,76$$

En caso de que $e_{ij} < 5 \quad \forall i, j$ se agruparían filas o columnas contiguas.

En tablas de contingencia 2×3 se puede recurrir a una fórmula sencilla:

$$\chi^2_2 = \left[\frac{n}{n_{x_1}} \right] \left[\frac{n_{11}^2}{n_{y_1}} + \frac{n_{12}^2}{n_{y_2}} + \frac{n_{13}^2}{n_{y_3}} \right] + \left[\frac{n}{n_{x_2}} \right] \left[\frac{n_{21}^2}{n_{y_1}} + \frac{n_{22}^2}{n_{y_2}} + \frac{n_{23}^2}{n_{y_3}} \right] - n$$

$$\chi^2_2 = \left[\frac{404}{199} \right] \left[\frac{84^2}{202} + \frac{78^2}{140} + \frac{37^2}{62} \right] + \left[\frac{404}{205} \right] \left[\frac{118^2}{202} + \frac{62^2}{140} + \frac{25^2}{62} \right] - 404 = 9,76$$

Estadístico teórico $\chi^2_{0,05,2} = 5,991$

Como $\chi^2_2 = 9,76 > \chi^2_{0,05,2} = 5,991$ se cumple la región de rechazo, concluyendo que las muestras no son homogéneas, es decir, no proceden de la misma población, hombres y mujeres no opinan lo mismo.

160. En dos ciudades, A y B, se observó el color del pelo y de los ojos de sus habitantes, encontrándose las siguientes tablas:

Ciudad A		
Pelo \ Ojos	Rubio	No Rubio
Azul	47	23
No azul	31	93

Ciudad B		
Pelo \ Ojos	Rubio	No Rubio
Azul	54	30
No azul	42	80

- a) Hallar los coeficientes de contingencia de las dos ciudades.
b) ¿En cuál de las dos ciudades se puede afirmar que hay mayor dependencia entre el color del pelo y de los ojos?

Solución:

- a) Se calculan los valores de la χ^2 correspondientes a las dos observaciones, siendo la

frecuencia esperada $e_{ij} = \frac{n_{i \cdot} \cdot n_{\cdot j}}{n}$

Ciudad A			
Pelo \ Ojos	Rubio	No Rubio	Total
Azul	47 $e_{11} = 28,14$	23 $e_{12} = 41,85$	70
No azul	31 $e_{21} = 49,85$	93 $e_{22} = 74,14$	124
Total	78	116	194

$$e_{11} = \frac{70 \cdot 78}{194} = 28,14 \quad e_{12} = \frac{70 \cdot 116}{194} = 41,85 \quad e_{21} = \frac{124 \cdot 78}{194} = 49,85 \quad e_{22} = \frac{124 \cdot 116}{194} = 74,14$$

Estadístico contraste:

$$\chi^2_{(2-1) \cdot (2-1)} = \chi^2_1 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{n_{ij}^2}{e_{ij}} - n = \frac{47^2}{28,14} + \frac{23^2}{41,85} + \frac{31^2}{49,85} + \frac{93^2}{74,14} - 194 = 33,07$$

$$\text{O bien, } \chi^2_1 = \frac{n \cdot (n_{11} \cdot n_{22} - n_{12} \cdot n_{21})^2}{n_{x_1} \cdot n_{x_2} \cdot n_{y_1} \cdot n_{y_2}} = \frac{194 \cdot (47 \cdot 93 - 23 \cdot 31)^2}{70 \cdot 124 \cdot 78 \cdot 116} = 33,07$$

$$\text{Coeficiente de contingencia: } C_A = \sqrt{\frac{\chi^2_1}{\chi^2_1 + n}} = \sqrt{\frac{33,07}{33,07 + 194}} = 0,3816$$

En la población B la tabla de contingencia 2 x 2:

Ciudad B

Pelo \ Ojos	Rubio	No Rubio	Total
Azul	54 $e_{11} = 39,15$	30 $e_{12} = 44,85$	84 84
No azul	42 $e_{21} = 56,85$	80 $e_{22} = 65,15$	122 122
Total	96	110	206

$$e_{11} = \frac{84 \cdot 96}{206} = 39,15 \quad e_{12} = \frac{84 \cdot 110}{206} = 44,85 \quad e_{21} = \frac{96 \cdot 122}{206} = 56,85 \quad e_{22} = \frac{110 \cdot 122}{206} = 65,15$$

Estadístico contraste:

$$\chi^2_{(2-1) \cdot (2-1)} = \chi^2_1 = \sum_{i=1}^2 \sum_{j=1}^2 \frac{n_{ij}^2}{e_{ij}} - n = \frac{54^2}{39,15} + \frac{30^2}{44,85} + \frac{42^2}{56,85} + \frac{80^2}{65,15} - 206 = 17,82$$

Coefficiente de contingencia: $C_B = \sqrt{\frac{\chi^2_1}{\chi^2_1 + n}} = \sqrt{\frac{17,82}{17,82 + 206}} = 0,282$

b) Como el coeficiente de contingencia mide el grado de relación o dependencia entre las variables, en la población A hay mayor dependencia entre el color de los ojos y del pelo.

161. Para una muestra aleatoria simple de 350 días, el número de urgencias tratadas diariamente en un hospital queda reflejado en la siguiente tabla:

Número urgencias	0 – 5	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	Total días
Número de días	20	65	100	95	60	10	350

Contrastar con un nivel de significación del 5% si la distribución del número de urgencias tratadas diariamente en el hospital se ajusta a una distribución normal.

Solución:

Sea la variable aleatoria $X =$ " Número de urgencias diarias".

Se establece la hipótesis nula H_0 : 'La distribución empírica se ajusta a la normal'

Se acepta la hipótesis nula, a un nivel de significación α sí

$$\chi^2_{k-p-1} = \underbrace{\sum_{i=1}^k \frac{(n_i - e_i)^2}{e_i}}_{\text{estadístico contraste}} = \sum_{i=1}^k \frac{n_i^2}{e_i} - n < \underbrace{\chi^2_{\alpha, k-p-1}}_{\text{estadístico teórico}} \quad \text{donde} \quad \begin{array}{l} k \equiv \text{número intervalos} \\ p \equiv \text{número parámetros a estimar} \end{array}$$

1) Se obtiene la media y la desviación típica:

Intervalos	x_i	n_i	$x_i n_i$	$x_i^2 n_i$
0 – 5	2,5	20	50	125
5 – 10	7,5	65	487,5	3656,25
10 – 15	12,5	100	1250	15625
15 – 20	17,5	95	1662,5	29093,75
20 – 25	22,5	60	1350	30375
25 – 30	27,5	10	275	7562,5
		$n = \sum_{i=1}^6 n_i = 350$	$\sum_{i=1}^6 x_i n_i = 5075$	$\sum_{i=1}^6 x_i^2 n_i = 86437,5$

$$\bar{x} = \frac{1}{350} \sum_{i=1}^6 x_i n_i = 14,5 \quad \sigma_x^2 = \frac{1}{350} \sum_{i=1}^6 (x_i - \bar{x})^2 n_i = \frac{1}{350} \sum_{i=1}^6 x_i^2 n_i - (\bar{x})^2 = 36,71 \quad \sigma_x = 6,06$$

2) Se procede al ajuste de una distribución normal $N(14,5, 6,06)$ hallando las probabilidades de cada uno de los intervalos:

Intervalos	n_i	p_i	$e_i = p_i \cdot n$	$(n_i - e_i)^2$	$(n_i - e_i)^2 / e_i$
0 – 5	20	0,0498	17,43	6,6	0,38
5 – 10	65	0,1714	59,99	25,1	0,42
10 – 15	100	0,3023	105,81	33,76	0,32
15 – 20	95	0,2867	100,35	28,62	0,29
20 – 25	60	0,1396	48,86	124,1	2,54
25 – 30	10	0,0366	12,81	7,9	0,62
$n = 350$				$\sum_{i=1}^6 (n_i - e_i)^2 / e_i = 4,57$	

$$P(0 < x < 5) = P\left[\frac{0 - 14,5}{6,06} < \frac{x - 14,5}{6,06} < \frac{5 - 14,5}{6,06}\right] = P(-2,39 < z < -1,57) =$$

$$= P(1,57 < z < 2,39) = P(z > 1,57) - P(z > 2,39) = 0,0582 - 0,00842 = 0,04978$$

$$P(5 < x < 10) = P\left[\frac{5 - 14,5}{6,06} < \frac{x - 14,5}{6,06} < \frac{10 - 14,5}{6,06}\right] = P(-1,57 < z < -0,74) =$$

$$= P(0,74 < z < 1,57) = P(z > 0,74) - P(z > 1,57) = 0,2296 - 0,0582 = 0,1714$$

$$P(10 < x < 15) = P\left[\frac{10 - 14,5}{6,06} < \frac{x - 14,5}{6,06} < \frac{15 - 14,5}{6,06}\right] = P(-0,74 < z < 0,08) =$$

$$= P(0,08 < z < 0,74) = 1 - P(z > 0,74) - P(z > 0,08) = 1 - 0,4681 - 0,2296 = 0,3023$$

$$P(15 < x < 20) = P\left[\frac{15 - 14,5}{6,06} < \frac{x - 14,5}{6,06} < \frac{20 - 14,5}{6,06}\right] = P(0,08 < z < 0,91) =$$

$$= P(z > 0,08) - P(z > 0,91) = 0,4681 - 0,1814 = 0,2867$$

$$P(20 < x < 25) = P\left[\frac{20 - 14,5}{6,06} < \frac{x - 14,5}{6,06} < \frac{25 - 14,5}{6,06}\right] = P(0,91 < z < 1,73) =$$

$$= P(z > 0,91) - P(z > 1,73) = 0,1814 - 0,0418 = 0,1396$$

$$P(25 < x < 30) = P\left[\frac{25 - 14,5}{6,06} < \frac{x - 14,5}{6,06} < \frac{30 - 14,5}{6,06}\right] = P(1,73 < z < 2,56) =$$

$$= P(z > 1,73) - P(z > 2,56) = 0,0418 - 0,0052 = 0,0366$$

3) Se calculan las frecuencias esperadas, multiplicando las probabilidades por el número total de observaciones en el muestreo $e_i = p_i \cdot n$

Cuando un intervalo de clase presenta una frecuencia esperada menor que 5 es necesario agrupar dos o más intervalos contiguos hasta lograr que la frecuencia esperada sea mayor que 5. En un caso práctico se admite hasta un 20% de las celdas que no verifican este requisito.

4) Se comparan los estadístico de contraste empírico y teórico, donde el número de grados de libertad de Chi-cuadrado es:

$$k - p - 1 = (\text{número intervalos}) - (\text{número parámetros a estimar}) - 1 = 6 - 2 - 1 = 3$$

$$\text{Estadístico empírico: } \chi^2_3 = \sum_{i=1}^6 \frac{(n_i - e_i)^2}{e_i} = 4,57 \quad \text{Estadístico teórico } \chi^2_{0,05,3} = 7,815$$

Como $\chi^2_3 = 4,57 < \chi^2_{0,05,3} = 7,815$ se acepta la hipótesis nula a un nivel de significación del 5%.

La variable aleatoria número de urgencias en el hospital sigue una distribución normal.

162. La tabla adjunta refleja un análisis de la obesidad en 14 sujetos. Con un nivel de significación de 0,05, se desea analizar si existen diferencias en la prevalencia de obesidad entre hombres y mujeres o si, por el contrario, el porcentaje de obesos no varía entre sexos.

Sexo	Obesidad		Total
	Sí	No	
Mujeres	1	4	5
Hombres	7	2	9
Total	8	6	14

Solución:

El *test exacto de Fisher* permite analizar si dos variables dicotómicas están asociadas cuando la muestra a estudiar es demasiado pequeña y no cumple las condiciones necesarias para que la aplicación del test de la Chi-cuadrado sea idónea.

Las condiciones necesarias para aplicar el test de la Chi-cuadrado exige que al menos el 80% de los valores esperados de las celdas sean mayores que 5.

De este modo, en una tabla de contingencia de 2×2 será necesario que todas las celdas verifiquen esta condición, si bien en la práctica suele permitirse que una de ellas tenga frecuencias esperadas ligeramente por debajo de 5.

Si las dos variables que se están analizando son dicotómicas, y la frecuencia esperada es menor que 5 en más de una celda, no resulta adecuado aplicar el test de la χ^2 , aunque sí el *test exacto de Fisher*.

El test exacto de Fisher se basa en evaluar la probabilidad asociada a cada una de las tablas 2×2 que se pueden formar manteniendo los mismos totales de filas y columnas que los de la tabla

observada. Cada uno de estas probabilidades se obtiene bajo la hipótesis de independencia de las dos variables que se están analizando.

La probabilidad asociada a los datos que han sido observados viene dada por:

$$p = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{n! a! b! c! d!}$$

La fórmula general de la probabilidad descrita deberá calcularse para todas las tablas de contingencia que puedan formarse con los mismos totales de filas y columnas de la tabla observada.

El p_valor asociado al test exacto de Fisher puede calcularse sumando las probabilidades de las tablas que resulten menores o iguales a la probabilidad de la tabla que ha sido observada.

Cuando el planteamiento es bilateral, es decir, cuando la hipótesis alternativa asume la dependencia entre las variables dicotómicas, pero sin especificar de antemano en qué sentido se producen dichas diferencias, el p_valor se multiplica por 2.

En este caso, planteando la hipótesis nula H_0 : El sexo y ser obeso son independientes

Sexo	Obesidad		Total
	Sí	No	
Mujeres	1 (a)	4 (b)	5 (a+b)
Hombres	7 (c)	2 (d)	9 (c+d)
Total	8 (a+c)	6 (b+d)	14 (n)

$$p = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{n! a! b! c! d!} = \frac{5! 9! 8! 6!}{14! 1! 4! 7! 2!} = 0,0599$$

Las siguientes tablas muestran todas las posibles combinaciones de frecuencias que se pueden obtener con los mismos totales de filas y columnas:

Sexo	Obesidad		Total
	Sí	No	
Mujeres	4 (a)	1 (b)	5 (a+b)
Hombres	4 (c)	5 (d)	9 (c+d)
Total	8 (a+c)	6 (b+d)	14 (n)

$$p = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{n! a! b! c! d!} = \frac{5! 9! 8! 6!}{14! 4! 1! 4! 5!} = 0,2098 > 0,0599$$

Sexo	Obesidad		Total
	Sí	No	
Mujeres	2 (a)	3 (b)	5 (a + b)
Hombres	6 (c)	3 (d)	9 (c + d)
Total	8 (a + c)	6 (b + d)	14 (n)

$$p = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{n! a! b! c! d!} = \frac{5! 9! 8! 6!}{14! 2! 3! 6! 3!} = 0,2797 > 0,0599$$

Sexo	Obesidad		Total
	Sí	No	
Mujeres	3 (a)	2 (b)	5 (a + b)
Hombres	5 (c)	4 (d)	9 (c + d)
Total	8 (a + c)	6 (b + d)	14 (n)

$$p = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{n! a! b! c! d!} = \frac{5! 9! 8! 6!}{14! 3! 2! 5! 4!} = 0,4196 > 0,0599$$

Sexo	Obesidad		Total
	Sí	No	
Mujeres	0 (a)	5 (b)	5 (a + b)
Hombres	8 (c)	1 (d)	9 (c + d)
Total	8 (a + c)	6 (b + d)	14 (n)

$$p = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{n! a! b! c! d!} = \frac{5! 9! 8! 6!}{14! 0! 5! 8! 1!} = 0,0030 < 0,0599 \text{ Ninguna mujer obesa}$$

Sexo	Obesidad		Total
	Sí	No	
Mujeres	5 (a)	0 (b)	5 (a + b)
Hombres	3 (c)	6 (d)	9 (c + d)
Total	8 (a + c)	6 (b + d)	14 (n)

$$p = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{n! a! b! c! d!} = \frac{5! 9! 8! 6!}{14! 5! 0! 3! 6!} = 0,0280 < 0,0599$$

El p_valor se obtiene sumando las probabilidades de las tablas que son menores o iguales a la probabilidad de la tabla observada ($p = 0,0599$): $p_valor = 0,0599 + 0,0030 + 0,0280 = 0,0909$



Asignatura..... Grupo.....
 Apellidos..... Nombre.....
 Ejercicio del día.....

Siendo $p_valor = 0,0909 > 0,05$ se acepta la hipótesis nula, concluyendo que el sexo y el hecho de ser obeso son independientes. Es decir, no existe asociación entre las variables en estudio con un nivel de significación $\alpha = 0,05$

⇒ Otro método de calcular el p_valor consiste en sumar las probabilidades asociadas a aquellas tablas que sean más favorables a la hipótesis alternativa de los datos observados. La tabla extrema de los datos observados es la que no se observa ninguna mujer obesa tiene una probabilidad $p = 0,0030$

$$p_valor = 0,0599 + 0,0030 = 0,0629$$

El SPSS para el cómputo del test de Fisher, calcula el p_valor correspondiente a la alternativa bilateral ($2p = 2 \times 0,0909 = 0,1818$) y el p_valor asociado a un planteamiento unilateral ($p = 0,0909$).

163. En un estudio sobre la opinión de fumar en lugares públicos se realiza una encuesta a 350 personas, obteniendo los siguientes resultados:

	Opinión				$n_{i.}$
	Muy en contra	En contra	A Favor	Muy a favor	
Fumador	60	50	20	10	140
No Fumador	10	30	70	100	210
$n_{.j}$	70	80	90	110	350

Con un nivel de significación de 0,05 se pide:

- 1) ¿Existe diferencia de opinión entre fumadores y no fumadores?
- 2) ¿Cuál es el grado de dependencia en la opinión?
- 3) Coeficiente Phi y Coeficiente V de Cramer
- 4) Test G da la razón de verosimilitud
- 5) Coeficiente Lambda de Goodman y Kruskal
- 6) Coeficiente Tau de Goodman y Kruskal

Solución:

- 1) Se establecen las hipótesis:

H_0 : La opinión es independiente de su condición de fumador o no fumador

H_1 : La opinión no es independiente de su condición de fumador o no fumador

Se acepta H_0 sí: $\chi_c^2 = \overbrace{\sum_{i=1}^2 \sum_{j=1}^4 \frac{(n_{ij} - e_{ij})^2}{e_{ij}}}^{\text{estadístico observado}} < \overbrace{\chi_{\alpha, (2-1) \cdot (4-1)}^2}^{\text{estadístico teórico}} = \chi_{0,05,3}^2$

	Opinión				$n_{i.}$
	Muy en contra	En contra	A Favor	Muy a favor	
Fumador	60 $e_{11} = 28$	50 $e_{12} = 32$	20 $e_{13} = 36$	10 $e_{14} = 44$	140 140
No Fumador	10 $e_{21} = 42$	30 $e_{22} = 48$	70 $e_{23} = 54$	100 $e_{24} = 66$	210 210
$n_{.j}$	70	80	90	110	350

$$e_{11} = \frac{140 \cdot 70}{350} = 28 \quad e_{12} = \frac{140 \cdot 80}{350} = 32 \quad e_{13} = \frac{140 \cdot 90}{350} = 36 \quad e_{14} = \frac{140 \cdot 110}{350} = 44$$

$$e_{21} = \frac{210 \cdot 70}{350} = 42 \quad e_{22} = \frac{210 \cdot 80}{350} = 48 \quad e_{23} = \frac{210 \cdot 90}{350} = 54 \quad e_{24} = \frac{210 \cdot 110}{350} = 66$$

$$\chi_c^2 = \sum_{i=1}^2 \sum_{j=1}^4 \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^2 \sum_{j=1}^4 \frac{n_{ij}^2}{e_{ij}} - n =$$

$$= \frac{60^2}{28} + \frac{50^2}{32} + \frac{20^2}{36} + \frac{10^2}{44} + \frac{10^2}{42} + \frac{30^2}{48} + \frac{70^2}{54} + \frac{100^2}{66} - 350 = 133,46$$

Estadístico teórico: $\chi_{0,05,3}^2 = 7,815$

Siendo $\chi_c^2 = 133,46 > \chi_{0,05,3}^2 = 7,815$ se rechaza la hipótesis nula, se acepta por tanto la hipótesis alternativa, pudiendo afirmar con una significación 0,05 que la opinión sobre el tabaco depende de sí es o no fumador.

2) El grado de dependencia en la opinión viene dado por el Coeficiente de Contingencia:

$$C = \sqrt{\frac{\chi_c^2}{\chi_c^2 + n}} = \sqrt{\frac{133,46}{133,46 + 350}} = 0,525$$

El grado de dependencia es del 52,5% por lo que la asociación entre las variables es alta.

En las tablas de contingencia $k \times k$ el valor máximo de C es $C_{\text{máximo}} = \sqrt{\frac{k-1}{k}}$

3) El grado de asociación se calcula mediante el Coeficiente Phi y el Coeficiente V de Cramer, en las tablas de contingencia 2×2 es idéntico al estadístico Phi, presenta el problema de subestimar el grado de asociación entre las variables.

Coeficiente Phi: $\phi = \sqrt{\frac{\chi_c^2}{n}} = \sqrt{\frac{133,46}{350}} = 0,618$

Coeficiente V de Cramer: $V_{\text{Cramer}} = \sqrt{\frac{\chi_c^2}{n \cdot \min(k-1, m-1)}} = \sqrt{\frac{\chi_c^2}{n}} = \sqrt{\frac{133,46}{350}} = 0,618$

4) El Test G da la razón de verosimilitud es una Prueba de hipótesis de la Chi-cuadrado que presenta mejores resultados que el Test de la Chi-cuadrado de Pearson.

Test G de la razón de verosimilitud: $G = 2 \sum_{i=1}^k \sum_{j=1}^m n_{ij} \ln \left(\frac{n_{ij}}{e_{ij}} \right)$

Se acepta la hipótesis nula H_0 sí: $G = 2 \sum_{i=1}^2 \sum_{j=1}^4 n_{ij} \ln \left(\frac{n_{ij}}{e_{ij}} \right) < \chi_{\alpha, (2-1) \cdot (4-1)}^2$

	Opinión				$n_{i\cdot}$
	Muy en contra	En contra	A Favor	Muy a favor	
Fumador	60 $e_{11} = 28$ $g_{11} = 45,7$	50 $e_{12} = 32$ $g_{12} = 22,3$	20 $e_{13} = 36$ $g_{13} = -11,7$	10 $e_{14} = 44$ $g_{14} = -14,8$	140 140
No Fumador	10 $e_{21} = 42$ $g_{21} = -14,3$	30 $e_{22} = 48$ $g_{22} = -14,1$	70 $e_{23} = 54$ $g_{23} = 18,2$	100 $e_{24} = 66$ $g_{24} = 41,6$	210 210
$n_{\cdot j}$	70	80	90	110	350

$$g_{11} = 60 \ln\left(\frac{60}{28}\right) = 45,7 \quad g_{12} = 50 \ln\left(\frac{50}{32}\right) = 22,3 \quad g_{13} = 20 \ln\left(\frac{20}{36}\right) = -11,7 \quad g_{14} = 10 \ln\left(\frac{10}{44}\right) = -14,8$$

$$g_{21} = 10 \ln\left(\frac{10}{42}\right) = -14,3 \quad g_{22} = 30 \ln\left(\frac{30}{48}\right) = -14,1 \quad g_{23} = 70 \ln\left(\frac{70}{54}\right) = 18,2 \quad g_{24} = 100 \ln\left(\frac{100}{66}\right) = 41,6$$

$$G = 2 \sum_{i=1}^2 \sum_{j=1}^4 n_{ij} \ln\left(\frac{n_{ij}}{e_{ij}}\right) = 2[45,7 + 22,3 - 11,7 - 14,8 - 14,3 - 14,1 + 18,2 + 41,6] = 145,475$$

Siendo $G = 145,475 > 7,815 = \chi^2_{0,05,3}$ se rechaza la hipótesis nula pudiendo afirmar que la opinión sobre el tabaco depende de sí es o no fumador.

5) El Coeficiente Lambda (λ) de Goodman y Kruskal, conocido también como coeficiente de Goodman Predicción, se basa en la reducción proporcional del error en la predicción la moda, es decir, del número de aciertos que proporciona el conocer la distribución dividido por el número de errores sin conocerla.

Valores Lambda (λ) próximos a 0 implican baja asociación y valores próximos a 1 denotan fuerte asociación.

Dos variables son independientes cuando $\lambda = 0$

Sin embargo, $\lambda = 0$ no implica independencia estadística.

$$\lambda_{yx} = \frac{\sum m_y - M_y}{n - M_y} \begin{cases} M_y \equiv \text{Frecuencia modal global} \\ \sum m_y \equiv \text{Suma de frecuencias modales} \\ n \equiv \text{Número total de casos} \end{cases} \quad \text{ó} \quad \lambda = \frac{E_1 - E_2}{E_1} \begin{cases} E_1 = n - M_y \\ E_2 = n - \sum m_y \end{cases}$$

	Opinión				$n_{i\cdot}$
	Muy en contra	En contra	A Favor	Muy a favor	
Fumador	60	50	20	10	140
No Fumador	10	30	70	100	210
$n_{\cdot j}$	70	80	90	110	350

$$\lambda_{yx} = \frac{\sum m_y - M_y}{n - M_y} = \frac{280 - 210}{350 - 210} = 0,5 \quad \begin{cases} M_y \equiv 210 \\ \sum m_y \equiv 60 + 50 + 70 + 100 = 280 \\ n \equiv 350 \end{cases}$$

$$\lambda_{yx} = \frac{E_1 - E_2}{E_1} = \frac{140 - 70}{140} = 0,5 \quad \begin{cases} E_1 = n - M_y = 350 - 210 = 140 \\ E_2 = n - \sum m_y = 350 - 280 = 70 \end{cases}$$

$$\lambda_{xy} = \frac{\sum m_x - M_x}{n - M_x} = \frac{160 - 110}{350 - 110} = 0,208 \quad \begin{cases} M_x \equiv 110 \\ \sum m_x \equiv 60 + 100 = 160 \\ n \equiv 350 \end{cases}$$

$$\lambda_{xy} = \frac{E_1 - E_2}{E_1} = \frac{240 - 190}{240} = 0,208 \quad \begin{cases} E_1 = n - M_x = 350 - 110 = 240 \\ E_2 = n - \sum m_x = 350 - 160 = 190 \end{cases}$$

Un Fumador que estuviera *Muy en contra* de fumar en lugares públicos acertaría 60 veces de 70, es decir fallaría en 10 ocasiones. Un fumador que estuviera en contra tendría $80 - 50 = 30$ errores.

6) Coeficiente Tau de Goodman y Kruskal: Al igual que el coeficiente Lambda (λ) es un coeficiente asimétrico, aunque a diferencia del Lambda parte de los errores cometidos al asignar aleatoriamente los casos a las categorías de la variable dependiente.

$$\tau = \frac{E_1 - E_2}{E_1} \quad \text{donde} \quad E_1 = \sum_{i=1}^k \left[\frac{(n - n_{i\cdot}) n_{i\cdot}}{n} \right] \quad \text{y} \quad E_2 = \sum_{j=1}^m \sum_{i=1}^k \left[\frac{(n_{\cdot j} - n_{ij}) n_{ij}}{n_{\cdot j}} \right]$$

✓ Para conocer los errores sin conocer la distribución de la variable independiente:

Se supone que en cada categoría se clasificaran erróneamente por azar un número de casos, que en cada categoría es igual al número de casos que no pertenecen a la misma.

$$E_1 = \sum_{i=1}^k \left[\frac{(n - n_{i\cdot}) n_{i\cdot}}{n} \right] \quad \begin{aligned} n &\equiv \text{Número total de casos} \\ k &\equiv \text{Número de categorías de la variable} \\ n_{i\cdot} &\equiv \text{Frecuencia de la categoría } i\text{-ésima} \end{aligned}$$

	Opinión				$n_{i.}$
	Muy en contra	En contra	A Favor	Muy a favor	
Fumador	60	50	20	10	140
No Fumador	10	30	70	100	210
$n_{.j}$	70	80	90	110	350

⇒ En la categoría de Fumadores de $n_{1.} = 140$ de un total de $n = 350$ se cometerían

$n - n_{1.} = 350 - 140 = 210$ errores.

Intentando designar al azar los $n_{1.} = 140$ casos de Fumadores se cometería un error promedio

$$\text{de: } \frac{(n - n_{1.})}{n} \times n_{1.} = \frac{350 - 140}{350} \times 140 = 84$$

⇒ En la categoría de No Fumadores de $n_{2.} = 210$ de un total de $n = 350$ se cometerían

$n - n_{2.} = 350 - 210 = 140$ errores.

Intentando designar al azar los $n_{2.} = 210$ casos de No Fumadores se cometería un error

$$\text{promedio de: } \frac{(n - n_{2.})}{n} \times n_{2.} = \frac{350 - 210}{350} \times 210 = 84$$

$$E_1 = \sum_{i=1}^2 \left[\frac{(350 - n_{i.}) n_{i.}}{350} \right] = 84 + 84 = 168$$

✓ Para conocer los errores conociendo la distribución de la variable independiente:

$$E_2 = \sum_{j=1}^m \sum_{i=1}^k \left[\frac{(n_{.j} - n_{ij}) n_{ij}}{n_{.j}} \right]$$

$n_{ij} \equiv$ frecuencia de celdilla en la categoría i -ésima variable dependiente
 $m \equiv$ número de categorías de la variable independiente
 $n_{.j} \equiv$ total parcial de las categorías de la variable independiente

⇒ Categoría con la opinión Muy en contra:

$$\text{Fumadores: } \frac{(n_{.1} - n_{11}) n_{11}}{n_{.1}} = \frac{(70 - 60) 60}{70} = 8,57$$

$$\text{No Fumadores: } \frac{(n_{.1} - n_{21}) n_{21}}{n_{.1}} = \frac{(70 - 10) 10}{70} = 8,57$$

Errores en la categoría $E_{21} = 8,57 + 8,57 = 17,14$

⇒ Categoría con la opinión En contra:

$$\text{Fumadores: } \frac{(n_{\cdot 2} - n_{12}) n_{12}}{n_{\cdot 2}} = \frac{(80 - 50) 50}{80} = 18,75$$

$$\text{No Fumadores: } \frac{(n_{\cdot 2} - n_{22}) n_{22}}{n_{\cdot 2}} = \frac{(80 - 30) 30}{80} = 18,75$$

$$\text{Errores en la categoría } E_{22} = 18,75 + 18,75 = 37,5$$

⇒ Categoría con la opinión A favor:

$$\text{Fumadores: } \frac{(n_{\cdot 3} - n_{13}) n_{13}}{n_{\cdot 3}} = \frac{(90 - 20) 20}{90} = 15,56$$

$$\text{No Fumadores: } \frac{(n_{\cdot 3} - n_{23}) n_{23}}{n_{\cdot 3}} = \frac{(90 - 70) 70}{90} = 15,56$$

$$\text{Errores en la categoría } E_{23} = 15,56 + 15,56 = 31,12$$

⇒ Categoría con la opinión Muy a favor:

$$\text{Fumadores: } \frac{(n_{\cdot 4} - n_{14}) n_{14}}{n_{\cdot 4}} = \frac{(110 - 10) 10}{110} = 9,09$$

$$\text{No Fumadores: } \frac{(n_{\cdot 4} - n_{24}) n_{24}}{n_{\cdot 4}} = \frac{(110 - 100) 100}{110} = 9,09$$

$$\text{Errores en la categoría } E_{24} = 9,09 + 9,09 = 18,18$$

$$E_2 = \sum_{j=1}^4 \sum_{i=1}^2 \left[\frac{(n_{\cdot j} - n_{ij}) n_{ij}}{n_{\cdot j}} \right] = 17,14 + 37,5 + 31,12 + 18,18 = 103,94$$

$$\tau = \frac{E_1 - E_2}{E_1} = \frac{168 - 103,94}{168} = 0,381$$

El coeficiente Tau de Goodman y Kruskal $0 \leq \tau \leq 1$ toma valores entre 0 y 1.

0 significa ausencia de reducción del error de clasificación y 1 reducción completa del error de clasificación.

164. En la tabla se refleja la edad de los empleados de una empresa y el grado de satisfacción en el trabajo, con un nivel de significación del 5%, se solicita:

Edad	Satisfacción en el trabajo				
	A	B	C	D	E
< 25	10	10	20	40	70
25 - 36	20	10	15	20	30
> 36	60	50	30	10	5

- Contrastar si el grado de satisfacción en el trabajo no depende de la edad de los empleados mediante el test de Chi-cuadrado y el test G de la razón de verosimilitud.
- Analizar el grado de asociación entre las variables con los estadísticos Phi, V de Cramer y Coeficiente de contingencia.
- Calcular la Gamma de Goodman y Kruskal, Coeficiente Tau-C, Tau-B de Kendall, Coeficiente D de Somers y el Coeficiente Lambda de Goodman y Kruskal.
- Calcular el Coeficiente Tau de Goodman y Kruskal y el Coeficiente de Incertidumbre.

Solución:

- a) Sean las variables $X \equiv$ 'Edad de los empleados' e $Y \equiv$ 'Satisfacción en el trabajo'

Hipótesis nula H_0 : 'El grado de satisfacción en el trabajo no depende de la edad de los empleados'

$$\text{Se acepta } H_0: \chi^2_c = \sum_{i=1}^3 \sum_{j=1}^5 \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^3 \sum_{j=1}^5 \frac{n_{ij}^2}{e_{ij}} - n < \chi^2_{\alpha; (3-1) \cdot (5-1)}$$

Se forma la tabla de contingencia 3 x 5 donde cada frecuencia observada (n_{ij}) $i=1, 2, 3$; $j=1, \dots, 5$ tiene una frecuencia teórica o esperada en caso de independencia $e_{ij} = \frac{n_{i\cdot} \times n_{\cdot j}}{n}$

Edad	Satisfacción en el trabajo					$n_{i\cdot}$
	A	B	C	D	E	
< 25	10 $e_{11} = 33,75$	10 $e_{12} = 26,25$	20 $e_{13} = 24,37$	40 $e_{14} = 26,25$	70 $e_{15} = 39,37$	150 150
25 - 36	20 $e_{21} = 21,37$	10 $e_{22} = 16,62$	15 $e_{23} = 15,44$	20 $e_{24} = 16,62$	30 $e_{25} = 24,94$	95 95
> 36	60 $e_{31} = 34,87$	50 $e_{32} = 27,12$	30 $e_{33} = 25,19$	10 $e_{34} = 27,12$	5 $e_{35} = 40,69$	155 155
$n_{\cdot j}$	90	70	65	70	105	400

$$\begin{aligned}
 e_{11} &= \frac{150 \cdot 90}{400} = 33,75 & e_{21} &= \frac{95 \cdot 90}{400} = 21,37 & e_{31} &= \frac{155 \cdot 90}{400} = 34,87 \\
 e_{12} &= \frac{150 \cdot 70}{400} = 26,25 & e_{22} &= \frac{95 \cdot 70}{400} = 16,62 & e_{32} &= \frac{155 \cdot 70}{400} = 27,12 \\
 e_{13} &= \frac{150 \cdot 65}{400} = 24,37 & e_{23} &= \frac{95 \cdot 65}{400} = 15,44 & e_{33} &= \frac{155 \cdot 65}{400} = 25,19 \\
 e_{14} &= \frac{150 \cdot 70}{400} = 26,25 & e_{24} &= \frac{95 \cdot 70}{400} = 16,62 & e_{34} &= \frac{155 \cdot 70}{400} = 27,12 \\
 e_{15} &= \frac{150 \cdot 105}{400} = 39,37 & e_{25} &= \frac{95 \cdot 105}{400} = 24,94 & e_{35} &= \frac{155 \cdot 105}{400} = 40,69
 \end{aligned}$$

Estadístico observado:

$$\begin{aligned}
 \chi^2_c &= \sum_{i=1}^3 \sum_{j=1}^5 \frac{(n_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^3 \sum_{j=1}^5 \frac{n_{ij}^2}{e_{ij}} - n = \\
 &= \left(\frac{10^2}{33,75} + \frac{10^2}{26,25} + \frac{20^2}{24,37} + \frac{40^2}{26,25} + \frac{70^2}{39,37} \right) + \left(\frac{20^2}{21,37} + \frac{10^2}{16,62} + \frac{15^2}{15,44} + \frac{20^2}{16,62} + \frac{30^2}{24,94} \right) + \\
 &+ \left(\frac{60^2}{34,87} + \frac{50^2}{27,12} + \frac{30^2}{25,19} + \frac{10^2}{27,12} + \frac{5^2}{40,69} \right) - 400 = 143,458
 \end{aligned}$$

Estadístico teórico: $\chi^2_{0,05; (3-1) \cdot (5-1)} = \chi^2_{0,05; 8} = 15,507$

Como $\chi^2_8 = 143,458 > 15,507 = \chi^2_{0,05; 8}$ se rechaza la hipótesis nula de independencia entre la edad y la satisfacción en el trabajo. En consecuencia, la edad influye significativamente en la satisfacción en el trabajo.

▪ Test G de la razón de verosimilitud: $G = 2 \sum_{i=1}^k \sum_{j=1}^m n_{ij} \ln \left(\frac{n_{ij}}{e_{ij}} \right)$

Se acepta la hipótesis nula H_0 sí: $G = 2 \sum_{i=1}^3 \sum_{j=1}^5 n_{ij} \ln \left(\frac{n_{ij}}{e_{ij}} \right) < \chi^2_{\alpha, (3-1) \cdot (5-1)}$

Edad	Satisfacción en el trabajo					$n_{i.}$
	A	B	C	D	E	
< 25	10	10	20	40	70	150
	$e_{11} = 33,75$ $g_{11} = -12,16$	$e_{12} = 26,25$ $g_{12} = -9,65$	$e_{13} = 24,37$ $g_{13} = -3,95$	$e_{14} = 26,25$ $g_{14} = 16,85$	$e_{15} = 39,37$ $g_{15} = 40,28$	150 31,37
25 - 36	20	10	15	20	30	95
	$e_{21} = 21,37$ $g_{21} = -1,33$	$e_{22} = 16,62$ $g_{22} = -5,08$	$e_{23} = 15,44$ $g_{23} = -0,43$	$e_{24} = 16,62$ $g_{24} = 3,7$	$e_{25} = 24,94$ $g_{25} = 5,54$	95 2,4
> 36	60	50	30	10	5	155
	$e_{31} = 34,87$ $g_{31} = 32,56$	$e_{32} = 27,12$ $g_{32} = 30,59$	$e_{33} = 25,19$ $g_{33} = 5,24$	$e_{34} = 27,12$ $g_{34} = -9,98$	$e_{35} = 40,69$ $g_{35} = -10,48$	155 47,93
$n_{.j}$	90	70	65	70	105	400 81,7

$$g_{11} = 10 \ln \left(\frac{10}{33,75} \right) = -12,16 \quad g_{21} = 20 \ln \left(\frac{20}{21,37} \right) = -1,33 \quad g_{31} = 60 \ln \left(\frac{60}{34,87} \right) = 32,56$$

$$g_{12} = 10 \ln \left(\frac{10}{26,25} \right) = -9,65 \quad g_{22} = 10 \ln \left(\frac{10}{16,62} \right) = -5,08 \quad g_{32} = 50 \ln \left(\frac{50}{27,12} \right) = 30,59$$

$$g_{13} = 20 \ln \left(\frac{20}{24,37} \right) = -3,95 \quad g_{23} = 15 \ln \left(\frac{15}{15,44} \right) = -0,43 \quad g_{33} = 30 \ln \left(\frac{30}{25,19} \right) = 5,24$$

$$g_{14} = 40 \ln \left(\frac{40}{26,25} \right) = 16,85 \quad g_{24} = 20 \ln \left(\frac{20}{16,62} \right) = 3,7 \quad g_{34} = 10 \ln \left(\frac{10}{27,12} \right) = -9,98$$

$$g_{15} = 70 \ln \left(\frac{70}{39,37} \right) = 40,28 \quad g_{25} = 30 \ln \left(\frac{30}{24,94} \right) = 5,54 \quad g_{35} = 5 \ln \left(\frac{5}{40,69} \right) = -10,48$$

$$G = 2 \sum_{i=1}^3 \sum_{j=1}^5 n_{ij} \ln \left(\frac{n_{ij}}{e_{ij}} \right) = 2.81,667 = 163,334 > 15,507 = \chi^2_{0,05;8}$$

Se rechaza la hipótesis nula de independencia entre la edad y la satisfacción en el trabajo, concluyendo que la edad influye significativamente en la satisfacción en el trabajo.

El test G da la razón de verosimilitud es una Prueba de hipótesis que presenta mejores resultados que el Test de la Chi-cuadrado de Pearson.

b) Coeficiente Phi: $\phi = \sqrt{\frac{\chi^2_c}{n}} = \sqrt{\frac{143,51}{400}} = 0,599$

El estadístico Phi mide el grado de asociación entre las variables.

Coefficiente V de Cramer:

$$V_{\text{Cramer}} = \sqrt{\frac{\chi_c^2}{n \cdot \min(k-1, m-1)}} = \sqrt{\frac{143,51}{400 \cdot \min(3-1, 5-1)}} = \sqrt{\frac{143,51}{400 \cdot 2}} = 0,423$$

El estadístico V de Cramer es una medida simétrica que cuantifica la relación entre dos o más variables de la escala nominal. Quizás es el estadístico más utilizado.

Un valor del estadístico V de Cramer próximo a 0 indica la falta de asociación de las variables, mientras que próximo a 1 refleja mayor asociación entre las variables en estudio.

Como $V_{\text{Cramer}} = 0,423$ se detecta una relación moderada de las variables.

Coefficiente de contingencia: $C = \sqrt{\frac{\chi_c^2}{\chi_c^2 + n}} = \sqrt{\frac{143,51}{143,51 + 400}} = 0,514$

El grado de dependencia es del 51,4% por lo que la asociación entre las variables es moderada.

c) La Gamma γ de Goodman y Kruskal mide la fuerza de asociación de los datos cuando las variables se miden en el nivel ordinal.

$$\gamma = \frac{C - D}{C + D} \quad -1 \leq \gamma \leq 1 \quad \rightarrow \quad \gamma = 0 \text{ indica ausencia de asociación.}$$

Siendo: C $\equiv$ Pares Concordantes D $\equiv$ Pares Discordantes

Edad	Satisfacción en el trabajo					$n_{i\cdot}$
	A	B	C	D	E	
< 25	10	10	20	40	70	150
25 - 36	20	10	15	20	30	95
> 36	60	50	30	10	5	155
$n_{\cdot j}$	90	70	65	70	105	400

Pares Concordantes:

$$\begin{aligned}
 C &= 10 [10 + 15 + 20 + 30 + 50 + 30 + 10 + 5] \\
 &+ 10 [15 + 20 + 30 + 30 + 10 + 5] \\
 &+ 20 [20 + 30 + 10 + 5] \\
 &+ 40 [30 + 5] \\
 &+ 20 [50 + 30 + 10 + 5] \\
 &+ 10 [30 + 10 + 5] \\
 &+ 15 [10 + 5] \\
 &+ 20 [5] \\
 &= 8175
 \end{aligned}$$

Pares Discordantes:

$$\begin{aligned}
 D &= 70 [20 + 10 + 15 + 20 + 60 + 50 + 30 + 10] \\
 &+ 40 [20 + 10 + 15 + 60 + 50 + 30] \\
 &+ 20 [20 + 10 + 60 + 50] \\
 &+ 10 [20 + 60] \\
 &+ 30 [60 + 50 + 30 + 10] \\
 &+ 20 [60 + 50 + 30] \\
 &+ 15 [60 + 50] \\
 &+ 10 [60] \\
 &= 35600
 \end{aligned}$$

$$\gamma = \frac{C - D}{C + D} = \frac{8175 - 35600}{8175 + 35600} = -0,626$$

▪ El coeficiente Tau-C de Kendall (τ_c) a menudo se utiliza como un estadístico control en una prueba de hipótesis estadística para establecer si dos variables pueden considerarse estadísticamente dependientes.

Es una prueba no paramétrica, ya que no se basa en suposiciones sobre las distribuciones de X o Y o la distribución de (X, Y).

Bajo la hipótesis nula de independencia de X e Y, la distribución muestral de Tau-C (τ_c) tiene un valor esperado de cero.

Para muestras pequeñas: $\tau_c = \frac{2 \cdot \min(k, m) \cdot (C - D)}{\min(k - 1, m - 1) \cdot n^2}$

En muestras grandes, se utiliza una aproximación a $N(0, 1)$: $\tau_c = \frac{2(2n+5)}{9n(n-1)}$

$$\tau_c = \frac{2 \cdot \min(k, m) \cdot (C - D)}{\min(k-1, m-1) \cdot n^2} = \frac{2 \cdot \min(3, 5) \cdot (8175 - 35600)}{\min(3-1, 5-1) \cdot 400^2} = \frac{2 \cdot 3 \cdot (-27425)}{2 \cdot 400^2} = -0,514$$

- El coeficiente Tau-B de Kendall (τ_B) es una medida no paramétrica de la correlación para variables ordinales o de rangos que tiene en consideración los empates.

El signo del coeficiente indica la dirección de la relación y su valor absoluto indica la fuerza de la relación. Varía entre -1 y 1 según sea el sentido de la asociación entre las variables. Los valores mayores indican que la relación es más estrecha.

Cuando la tabla no es cuadrada este coeficiente no puede llegar a valer 1 dado que existirán más pares empatados en la variable que tenga más categorías.

$$\tau_B = \frac{C - D}{\sqrt{\left(\frac{n(n-1)}{2} - T_x\right)\left(\frac{n(n-1)}{2} - T_y\right)}}$$

- Parejas empatadas en X o en Y: $T_x = \sum_{i=1}^k \frac{n_{i\cdot}(n_{i\cdot}-1)}{2}$ $T_y = \sum_{j=1}^m \frac{n_{\cdot j}(n_{\cdot j}-1)}{2}$

Edad	Satisfacción en el trabajo					$n_{i\cdot}$
	A	B	C	D	E	
< 25	10	10	20	40	70	150
25 - 36	20	10	15	20	30	95
> 36	60	50	30	10	5	155
$n_{\cdot j}$	90	70	65	70	105	400

$$T_x = \sum_{i=1}^3 \frac{n_{i\cdot}(n_{i\cdot}-1)}{2} = \frac{1}{2} [150 \cdot 149 + 95 \cdot 94 + 155 \cdot 154] = 27575$$

$$T_y = \sum_{j=1}^5 \frac{n_{\cdot j}(n_{\cdot j}-1)}{2} = \frac{1}{2} [90 \cdot 89 + 70 \cdot 69 + 65 \cdot 64 + 70 \cdot 69 + 105 \cdot 104] = 16375$$

$$\tau_B = \frac{8175 - 35600}{\sqrt{(79800 - 27575)(79800 - 16375)}} = -0,477$$

- El estadístico D de Somers establece si las variables ordinales son dependientes o independientes entre sí.

El coeficiente D de Somers varía entre -1 y 1 , es una medida asimétrica como el coeficiente Lambda, los dos valores que se pueden obtener de la tabla dependen de que se tome como independiente la variable X o Y.

Valores del estadístico D cercanos a 0 indican que no hay ninguna o muy poca asociación entre las variables.

$$D \text{ de Somers: } D_x = \frac{C - D}{\frac{n(n-1)}{2} - T_x} \quad D_y = \frac{C - D}{\frac{n(n-1)}{2} - T_y}$$

$$\text{Número de pares: } \binom{n}{2} = \frac{n(n-1)}{2} = \frac{400(400-1)}{2} = 79800$$

$$D_x = \frac{C - D}{\frac{n(n-1)}{2} - T_x} = \frac{8175 - 35600}{79800 - 27575} = -0,525$$

$$D_y = \frac{C - D}{\frac{n(n-1)}{2} - T_x} = \frac{8175 - 35600}{79800 - 16375} = -0,432$$

- El Coeficiente Lambda λ de Goodman y Kruskal, conocido también como coeficiente de Goodman Predicción, se basa en la reducción proporcional del error en la predicción la moda.

Estadístico utilizado para determinar si usar los resultados de una de las variables puede utilizarse para predecir los resultados de la otra variable.

Valores Lambda λ próximos a 0 implican baja asociación y valores próximos a 1 denotan fuerte asociación.

Dos variables son independientes tienen $\lambda = 0$

Sin embargo $\lambda = 0$ no implica independencia estadística.

$$\lambda_{yx} = \frac{\sum m_y - M_y}{n - M_y} \begin{cases} M_y \equiv \text{Frecuencia modal global} \\ \sum m_y \equiv \text{Suma de frecuencias modales} \\ n \equiv \text{Número total de casos} \end{cases} \quad \text{ó bien} \quad \lambda = \frac{E_1 - E_2}{E_1} \begin{cases} E_1 = n - M_y \\ E_2 = n - \sum m_y \end{cases}$$

Edad	Satisfacción en el trabajo					$n_{i\cdot}$
	A	B	C	D	E	
< 25	10	10	20	40	70	150
25 - 36	20	10	15	20	30	95
> 36	60	50	30	10	5	→ 155
$n_{\cdot j}$	90	70	65	70	105	$n = 400$

$$\lambda_{yx} = \frac{\sum m_y - M_y}{n - M_y} = \frac{250 - 155}{400 - 155} = 0,388 \quad \left\{ \begin{array}{l} M_y = 155 \\ \sum m_y = 60 + 50 + 30 + 40 + 70 = 250 \\ n = 400 \end{array} \right.$$

$$\lambda_{xy} = \frac{\sum m_x - M_x}{n - M_x} = \frac{160 - 105}{400 - 105} = 0,186 \quad \left\{ \begin{array}{l} M_x = 105 \\ \sum m_x = 70 + 30 + 60 = 160 \\ n = 400 \end{array} \right.$$

d) La Tau de Goodman y Kruskal τ considera todas las categorías de respuesta y no únicamente la que contempla más casos entre dos variables nominales (variables cualitativas).

El valor de Tau de Goodman y Kruskal τ se interpreta como el porcentaje que mejora el error al incluir la variable independiente en la predicción de los valores de la variable dependiente.

Se parece a la Lambda λ siendo su cálculo más complejo. Lo mismo que Lambda adopta valores entre 0 y 1, dónde 0 es independencia y 1 el total de dependencia.

$$\tau = \frac{E_1 - E_2}{E_1} \quad \text{donde} \quad \begin{array}{l} E_1 \equiv \text{Errores sin conocer la distribución de la variable independiente} \\ E_2 \equiv \text{Errores conociendo la distribución de la variable independiente} \end{array}$$

$$E_1 = \sum_{i=1}^k \left[\frac{(n - n_{i\cdot}) n_{i\cdot}}{n} \right] \quad \left\{ \begin{array}{l} n \equiv \text{número total de casos} \\ k \equiv \text{número de categorías de la variable} \\ n_{i\cdot} \equiv \text{frecuencia de la categoría i-ésima} \end{array} \right.$$

$$E_2 = \sum_{j=1}^m \sum_{i=1}^k \left[\frac{(n_{\cdot j} - n_{ij}) n_{ij}}{n_{\cdot j}} \right] \quad \begin{array}{l} n_{ij} \equiv \text{frecuencia de celdilla en la categoría i-ésima variable dependiente} \\ m \equiv \text{número de categorías de la variable independiente} \\ n_{\cdot j} \equiv \text{total parcial de las categorías de la variable independiente} \end{array}$$

Edad	Satisfacción en el trabajo					$n_{i\cdot}$
	A	B	C	D	E	
< 25	10	10	20	40	70	150
25 - 36	20	10	15	20	30	95
> 36	60	50	30	10	5	155
$n_{\cdot j}$	90	70	65	70	105	$n = 400$

✓ Cuando la variable dependiente es la edad:

$$E_1 = \sum_{i=1}^3 \left[\frac{(n - n_{i\cdot}) n_{i\cdot}}{n} \right] = \frac{(400 - 150)150}{400} + \frac{(400 - 95)95}{400} + \frac{(400 - 155)155}{400} = 261,125$$

$$E_2 = \sum_{j=1}^5 \sum_{i=1}^3 \left[\frac{(n_{\cdot j} - n_{ij}) n_{ij}}{n_{\cdot j}} \right] = \frac{(90 - 10)10}{90} + \frac{(90 - 20)20}{90} + \frac{(90 - 60)60}{90} +$$

$$+ \frac{(70 - 10)10}{70} + \frac{(70 - 10)10}{70} + \frac{(70 - 50)50}{70} +$$

$$+ \frac{(65 - 20)20}{65} + \frac{(65 - 15)15}{65} + \frac{(65 - 30)30}{65} +$$

$$+ \frac{(70 - 40)40}{70} + \frac{(70 - 20)20}{70} + \frac{(70 - 10)10}{70} +$$

$$+ \frac{(105 - 70)70}{105} + \frac{(105 - 30)30}{105} + \frac{(105 - 5)5}{105} +$$

$$= 206,93$$

$$\tau = \frac{E_1 - E_2}{E_1} = \frac{261,125 - 206,93}{261,125} = 0,208 \text{ edad variable dependiente}$$

✓ Cuando la variable dependiente es la satisfacción en el trabajo:

$$E_1 = \sum_{j=1}^5 \left[\frac{(n - n_{\cdot j}) n_{\cdot j}}{n} \right] = \frac{(400 - 90)90}{400} + \frac{(400 - 70)70}{400} + \frac{(400 - 65)65}{400} + \frac{(400 - 70)70}{400} + \frac{(400 - 105)105}{400} =$$

$$= 317,125$$

$$\begin{aligned}
 E_2 &= \sum_{i=1}^3 \sum_{j=1}^5 \left[\frac{(n_{i\cdot} - n_{ij}) n_{ij}}{n_{i\cdot}} \right] = \\
 &= \frac{(150-10)10}{150} + \frac{(150-10)10}{150} + \frac{(150-20)20}{150} + \frac{(150-40)40}{150} + \frac{(150-70)70}{150} + \\
 &+ \frac{(95-20)20}{95} + \frac{(95-10)10}{95} + \frac{(95-15)15}{95} + \frac{(95-20)20}{95} + \frac{(95-30)30}{95} + \\
 &+ \frac{(155-60)60}{155} + \frac{(155-50)50}{155} + \frac{(155-30)30}{155} + \frac{(155-10)10}{155} + \frac{(155-5)5}{155} = \\
 &= 285,38
 \end{aligned}$$

$$\tau = \frac{E_1 - E_2}{E_1} = \frac{317,125 - 285,38}{317,125} = 0,100 \quad \text{satisfacción variable dependiente}$$

▪ El Coeficiente de Incertidumbre es una medida de asociación basada en la reducción proporcional del error. Es una medida semejante a Lambda en cuanto a su concepción de la asociación de las variables, en relación a la capacidad predictiva y la disminución del error de dicha predicción.

El coeficiente de incertidumbre I depende de toda la distribución y no sólo de los valores modales (caso de Lambda), varía entre 0 y 1, tomando el valor 0 en el caso total de independencia. Es más difícil de interpretar que Lambda.

Tiene versiones asimétricas (dependiendo de cual de las dos variables sea dependiente) y una simétrica (donde no se distingue entre variable dependiente e independiente).

La versión asimétrica se interpreta como la proporción de incertidumbre reducida al predecir los valores de una variable a partir de los de valores de la otra variable.

La versión simétrica se interpreta como la proporción de incertidumbre reducida al predecir los valores de cualquiera de las dos variables mediante la tabla de contingencia.

Se obtiene mediante la fórmula: $I_{Y/X} = \frac{I(X) + I(Y) - I(XY)}{I(Y)}$

Para obtener $I_{X/Y}$ basta con intercambiar los papeles de I(X) e I(Y)

La versión simétrica: $I = \frac{2 [I(X) + I(Y) - I(XY)]}{I(X) + I(Y)}$

$$\text{donde: } I(X) = \sum_{i=1}^k \frac{n_{i\cdot}}{n} \ln \left(\frac{n_{i\cdot}}{n} \right) \quad I(Y) = \sum_{j=1}^m \frac{n_{\cdot j}}{n} \ln \left(\frac{n_{\cdot j}}{n} \right) \quad I(XY) = \sum_{i=1}^k \sum_{j=1}^m \frac{n_{ij}}{n} \ln \left(\frac{n_{ij}}{n} \right)$$

Edad	Satisfacción en el trabajo					$I_{i\cdot}$
	A	B	C	D	E	
< 25	10 $I_{11} = -0,092$	10 $I_{12} = -0,092$	20 $I_{13} = -0,150$	40 $I_{14} = -0,230$	70 $I_{15} = -0,305$	150 $-0,368$
25 - 36	20 $I_{21} = -0,150$	10 $I_{22} = -0,092$	15 $I_{23} = -0,123$	20 $I_{24} = -0,150$	30 $I_{25} = -0,194$	95 $-0,341$
> 36	60 $I_{31} = -0,284$	50 $I_{32} = -0,260$	30 $I_{33} = -0,194$	10 $I_{34} = -0,092$	5 $I_{35} = -0,055$	155 $-0,367$
$I_{\cdot j}$	90 $-0,335$	70 $-0,305$	65 $-0,295$	70 $-0,305$	105 $-0,351$	400

$$I_{11} = \frac{10}{400} \ln \left(\frac{10}{400} \right) = -0,092 \quad I_{21} = \frac{20}{400} \ln \left(\frac{20}{400} \right) = -0,150 \quad I_{31} = \frac{60}{400} \ln \left(\frac{60}{400} \right) = -0,284$$

$$I_{12} = \frac{10}{400} \ln \left(\frac{10}{400} \right) = -0,092 \quad I_{22} = \frac{10}{400} \ln \left(\frac{10}{400} \right) = -0,092 \quad I_{32} = \frac{50}{400} \ln \left(\frac{50}{400} \right) = -0,260$$

$$I_{13} = \frac{20}{400} \ln \left(\frac{20}{400} \right) = -0,150 \quad I_{23} = \frac{15}{400} \ln \left(\frac{15}{400} \right) = -0,123 \quad I_{32} = \frac{30}{400} \ln \left(\frac{30}{400} \right) = -0,194$$

$$I_{14} = \frac{40}{400} \ln \left(\frac{40}{400} \right) = -0,230 \quad I_{24} = \frac{20}{400} \ln \left(\frac{20}{400} \right) = -0,150 \quad I_{34} = \frac{10}{400} \ln \left(\frac{10}{400} \right) = -0,092$$

$$I_{15} = \frac{70}{400} \ln \left(\frac{70}{400} \right) = -0,305 \quad I_{25} = \frac{30}{400} \ln \left(\frac{30}{400} \right) = -0,194 \quad I_{35} = \frac{5}{400} \ln \left(\frac{5}{400} \right) = -0,055$$

$$I_{1\cdot} = \frac{150}{400} \ln \left(\frac{150}{400} \right) = -0,368 \quad I_{2\cdot} = \frac{95}{400} \ln \left(\frac{95}{400} \right) = -0,341 \quad I_{3\cdot} = \frac{155}{400} \ln \left(\frac{155}{400} \right) = -0,367$$

$$I(X) = \sum_{i=1}^3 \frac{n_{i\cdot}}{n} \ln \left(\frac{n_{i\cdot}}{n} \right) = -0,368 - 0,341 - 0,367 = -1,076$$

$$I_{\cdot 1} = \frac{90}{400} \ln \left(\frac{90}{400} \right) = -0,335 \quad I_{\cdot 2} = \frac{70}{400} \ln \left(\frac{70}{400} \right) = -0,305 \quad I_{\cdot 3} = \frac{65}{400} \ln \left(\frac{65}{400} \right) = -0,295$$

$$I_{\cdot 4} = \frac{70}{400} \ln \left(\frac{70}{400} \right) = -0,305 \quad I_{\cdot 5} = \frac{105}{400} \ln \left(\frac{105}{400} \right) = -0,351$$

$$I(Y) = \sum_{j=1}^5 \frac{n_{\cdot j}}{n} \ln \left(\frac{n_{\cdot j}}{n} \right) = -0,335 - 0,305 - 0,295 - 0,305 - 0,351 = -1,591$$

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

$$I(XY) = \sum_{i=1}^3 \sum_{j=1}^5 \frac{n_{ij}}{n} \ln \left(\frac{n_{ij}}{n} \right) = -2,463$$

✓ Coeficiente de Incertidumbre: Satisfacción como variable dependiente:

$$I_{Y/X} = \frac{I(X) + I(Y) - I(XY)}{I(Y)} = \frac{-1,076 - 1,591 + 2,463}{-1,591} = 0,128$$

✓ Coeficiente de Incertidumbre: Edad como variable dependiente:

$$I_{X/Y} = \frac{I(X) + I(Y) - I(XY)}{I(X)} = \frac{-1,076 - 1,591 + 2,463}{-1,076} = 0,190$$

✓ Coeficiente de Incertidumbre simétrico:

$$I = \frac{2[I(X) + I(Y) - I(XY)]}{I(X) + I(Y)} = \frac{2[-1,076 - 1,591 + 2,463]}{-1,076 - 1,591} = 0,153$$

165. Para analizar la repercusión que tienen los debates televisivos en la intención de voto, un equipo de investigación recogió datos entre 240 individuos antes y después del debate, resultando la siguiente tabla:

Antes del debate (candidatos)	Después del debate (candidatos)		Total
	A	B	
A	46	50	96
B	85	59	144
Total	131	109	240

¿El debate televisivo cambió la intención de voto, con un nivel de significación del 5%.

Solución:

Se trata de una muestra pareada en una situación antes – después, con lo que es idóneo un contraste estadístico Chi-cuadrado de McNemar.

Antes del debate (candidatos)	Después del debate (candidatos)		Total
	A	B	
A	46 (a)	50 (b)	96 (a + b)
B	85 (c)	59 (d)	144 (c + d)
Total	131 (a + c)	109 (b + d)	240 (n)

Hipótesis nula H_0 : La intención de voto es la misma antes y después del debate

En esta prueba para la significación de cambios solo interesa conocer las celdas que presentan cambios (celdas b y c) y siendo $(b + c)$ el número de personas que cambiaron, de acuerdo con la hipótesis nula planteada se espera que $\left(\frac{b+c}{2}\right)$ casos cambien en una dirección y $\left(\frac{b+c}{2}\right)$ casos a otra dirección.

- Estadístico de contraste si $b + c < 20$

Se acepta H_0 si $\chi^2_{McNemar} = b < \chi^2_{\alpha/2, 1}$

- Estadístico de contraste si $b + c \geq 20$: $\chi^2_{McNemar} = \chi^2_1 = \frac{[|b - c| - 1]^2}{b + c}$

Se acepta H_0 si $\chi^2_{McNemar} = \chi^2_1 = \frac{[|b - c| - 1]^2}{b + c} < \chi^2_{\alpha/2, 1}$

En este caso, $b + c = 50 + 85 = 135 > 20$

Estadístico muestral: $\chi^2_{\text{McNemar}} = \frac{[|50 - 85| - 1]^2}{50 + 85} = 8,563$

Estadístico teórico: $\chi^2_{\alpha/2, 1} = \chi^2_{0,025, 1} = 5,024$

Como $\chi^2_{\text{McNemar}} = 8,563 > 5,024 = \chi^2_{0,025, 1}$ se rechaza la hipótesis nula, concluyendo que la intención de voto cambió significativamente después del debate.

166. Un agricultor desea saber si existe diferencia entre diez abonos en el cultivo en una determinada zona. Para ello abona seis matas con cada abono, observa el número de kilos y anota el tipo de abono X_i y la varianza de cada abono σ_i^2 , obteniendo los siguientes resultados:

X_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
σ_i^2	9	3	4	3	5	4	2	4	5	3

¿Hay diferencia entre los abonos a un nivel de $\alpha = 0,01$? ¿Y aun nivel de $\alpha = 0,05$?

Solución:

Se desea comprobar si todos los abonos tienen o no la misma variabilidad. Para ello se establece la hipótesis nula H_0 suponiendo que todas las varianzas son iguales, esto es, no hay diferencia entre los abonos.

Aplicando la prueba de Bartlett se acepta H_0 cuando el estadístico de la muestra χ^2_{k-1} de k muestras con $n_1, n_2, \dots, n_k$ elementos es menor que el estadístico teórico $\chi^2_{\alpha, k-1}$

$$\chi^2_{k-1} = \left[\ln(\bar{\sigma}^2) \cdot \sum_{i=1}^k (n_i - 1) \right] - \sum_{i=1}^k [(n_i - 1) \cdot \ln \sigma_i^2] \quad \bar{\sigma}^2 = \frac{1}{k} \sum_{i=1}^k \sigma_i^2$$

Si todas las muestras tuvieran los mismos elementos $n_1 = n_2 = \dots = n_k$ la expresión anterior se transforma en:

$$\chi^2_{k-1} = (n - 1) \cdot \left[k \cdot \ln \bar{\sigma}^2 - \sum_{i=1}^k \ln \sigma_i^2 \right] \quad \text{donde } k \equiv \text{Número de varianzas que se comparan}$$

En este caso, $n_1 = n_2 = \dots = n_{10} = 6$

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

X_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	Total
σ_i^2	9	3	4	3	5	4	2	4	5	3	42
$\ln \sigma_i^2$	2,20	1,10	1,39	1,10	1,61	1,39	0,69	1,39	1,61	1,10	13,58

$$\bar{\sigma}^2 = \frac{1}{10} \sum_{i=1}^{10} \sigma_i^2 = \frac{42}{10} = 4,2$$

$$\chi^2_{10-1} = (n-1) \cdot \left[k \cdot \ln \bar{\sigma}^2 - \sum_{i=1}^{10} \ln \sigma_i^2 \right] = (6-1) \cdot [10 \cdot \ln 4,2 - 13,58] = 3,85$$

Con un nivel de significación $\alpha = 0,01$: $\chi^2_9 = 3,85 < 21,666 = \chi^2_{0,01,9} \rightarrow$ Se acepta H_0

167. Un inversor tiene acciones en diferentes sectores del mercado continuo de valores. Después de quedarse sin asesor, para tener idea sobre la cotización de sus acciones seleccionó nueve sociedades en las que participaba y revisó sus cotizaciones de cierre el pasado viernes:

Sociedades	Cotización compra	Cotización actual
Acerinox	10,07	9,92
Bankinter	7,70	7,85
Bankia	3,98	4,37
Caixabank	2,56	1,44
Mapfre	3,05	3,05
BBVA	4,41	3,32
IAG	6,73	6,46
Telefónica	7,85	8,09
Mercadona	2,95	2,95

A un nivel de significación del 5%, ¿puede afirmarse que no hay diferencia significativa entre el valor de las acciones en las sociedades en las que participa este inversor?

Nota: Utilizar la prueba de rangos con signos de Wilcoxon

Solución:

La *prueba de rangos con signos de Wilcoxon* para dos muestras pareadas y la variable de respuesta ordinal o cuantitativa es la prueba homóloga no paramétrica a la prueba paramétrica *t* de Student para muestras pareadas.

Se utiliza como alternativa de la *prueba t de Student* cuando no se puede suponer la normalidad de las muestras.

Se establecen las hipótesis:

H_0 : No hay diferencias en valor de las acciones en las que participa el inversor

H_1 : El valor de las acciones actuales es inferior al valor de compra



La prueba no paramétrica de Wilcoxon sirve para rechazar o no la diferencia estadísticamente significativa entre dos conjuntos de observaciones que tienen una distribución cualesquiera desconocida. La prueba consiste en lo siguiente:

- Se calculan las diferencias de dos observaciones
- Se jerarquiza en orden ascendente el valor absoluto de las diferencias
- Se suman las jerarquizaciones de las diferencias positivas, denominando a la suma con el símbolo $T_+ = \text{Suma de rangos de las } D_i > 0$
- Se calculan los valores críticos $\bar{T}$ de la tabla para el test de rangos-signos de Wilcoxon
- Si $T_+ < \bar{T}$ no existe diferencia significativa entre las observaciones a un nivel de significación α

La prueba de Wilcoxon supone que la distribución de las diferencias es continua, simétrica, independiente y que las diferencias tienen la misma mediana.

Sociedad	Cotización compra	Cotización actual	Diferencia D_i	Diferencia absoluta $ D_i $	Jerarquía ascendente	Rango Diferencias R_i		R_i^-	R_i^+
1	10,07	9,92	0,15	0,15	4	2	1,5	1,5	
2	7,7	7,85	-0,15	0,15	3	1	1,5		1,5
3	3,98	4,37	-0,39	0,39	7	5	5		5
4	2,56	1,44	1,12	1,12	9	7	7	7	
5	3,05	3,05	0	0	1				
6	4,41	3,32	1,09	1,09	8	6	6	6	
7	6,73	6,46	0,27	0,27	6	4	4	4	
8	7,85	8,09	-0,24	0,24	5	3	3		3
9	2,95	2,95	0	0	1				
								18,5	9,5

Para obtener los rangos, los valores absolutos de las diferencias $|D_i|$ se jerarquizan en orden ascendente (de menor a mayor) y se asignan los rangos o números de orden desde 1 hasta n (en este caso, $n = 1, \dots, 9$). Si existen valores de $|D_i|$ repetidos, el rango correspondiente será el promedio de los que se les asignarían si fueran diferentes.

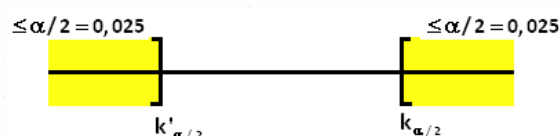
Cuando hay diferencias nulas no se tienen en cuenta, el tamaño muestral se reduce en el número de observaciones menor el número de diferencias nulas.

✓ En este caso, hay dos diferencias nulas y el tamaño muestral es $n = 9 - 7 = 2$, con lo que los rangos o números de orden van de 1 a 7. Por otra parte, la sociedad 1 y 2 presentan el mismo valor $|D_i|$ por lo que el rango de cada una de ellas será el promedio que se les asignarían si fueran diferentes $(1 + 2) / 2 = 1,5$

$$T_+ = \sum R_i^+ = 9,5 \quad T_- = \sum R_i^- = 18,5$$

El valor experimental es $\hat{T}_+ = \sum R_i^+ = 1,5 + 5 + 3 = 9,5$

Al tratarse de un contraste bilateral, la región crítica tendrá dos colas:



$$P(T_+ \leq k'_{\alpha/2, n}) \leq \alpha/2 \equiv P(T_+ \leq k'_{0,025, 7}) \leq 0,025 \rightarrow k'_{0,025, 7} = 2$$

$$P(T_+ \geq k_{\alpha/2, n}) \leq \alpha/2 \equiv P(T_+ \geq k_{0,025, 7}) \leq 0,05 = P(T_+ \leq k_{0,025, 7}) \leq 1 - 0,025 = 0,975 \rightarrow k_{0,025, 7} = 26$$

Así pues, como $k'_{0,025, 7} = 2 < \hat{T}_+ = 18,5 < 26 = k_{0,025, 7}$ se acepta la hipótesis nula, afirmando que no hay diferencias en el valor de las acciones de las sociedades en que participa el inversor.

Valores críticos para el test de rangos-signos de Wilcoxon

$$T_+ = \sum_{i=1}^n z_i \cdot \text{rango}(|D_i|)$$

α^* n	$P(T_+ \leq k_\alpha) \leq \alpha$					$P(T_+ \geq k_\alpha) \leq 1 - \alpha$				
	0,005	0,01	0,025	0,05	0,10	0,90	0,95	0,975	0,99	0,995
4	—	—	—	—	0	10	—	—	—	—
5	—	—	—	0	2	13	15	—	—	—
6	—	—	0	2	3	18	19	21	—	—
7	—	0	2	3	5	23	25	26	28	—
8	0	1	3	5	8	28	31	33	35	36
9	1	3	5	8	10	35	37	40	42	44
10	3	5	8	10	14	41	45	47	50	52
11	5	7	10	13	17	49	53	56	59	61
12	7	9	13	17	21	57	61	65	69	71
13	9	12	17	21	26	65	70	74	79	82
14	12	15	21	25	31	74	80	84	90	93
15	15	19	25	30	36	84	90	95	101	105

α^* no tiene por qué coincidir con el nivel de significación

168. Se emplean dos estrategias de ventas (A y B) de un mismo producto en veinte ciudades. El volumen de ventas anuales del mismo producto en cada ciudad (en miles de euros) por tipo de estrategia, está reflejado en la tabla adjunta. ¿Existe alguna diferencia significativa entre las estrategias de ventas a un nivel de significación del 5%?

Ventas (miles euros)			Ventas (miles euros)		
Ciudad	A	B	Ciudad	A	B
1	12,36	11,32	11	10,15	10,03
2	15,65	10,21	12	13,15	13,02
3	9,83	9,79	13	14,65	12,05
4	18,37	19,21	14	11,47	12,62
5	11,72	11,43	15	13,15	12,96
6	13,04	12,96	16	15,67	14,97
7	10,15	9,81	17	13,12	13,65
8	15,41	14,65	18	10,17	9,87
9	11,07	11,79	19	14,78	14,17
10	12,31	11,05	20	9,63	10,17

Solución:

Se trata de un contraste bilateral o de dos colas donde se establecen las hipótesis:

H_0 : No existe diferencia en la estrategia de ventas

H_1 : La estrategia de ventas son distintas

Se aplica la prueba de Wilcoxon suponiendo que la distribución de las diferencias en las estrategias de ventas es continua, simétrica, independiente y tienen la misma mediana.

Valores críticos $\bar{T}$ para la prueba de Wilcoxon

N	$0.01 < \alpha \leq 0.05$	$\alpha \leq 0.01$
5	14	—
6	19	—
7	21	28
8	30	34
9	37	42
10	44	50
12	61	68
14	79	89
16	100	112
18	124	138
20	150	167

Estrategias de ventas de un producto							
Ciudad	A	B	$D_i = A - B$	$ D_i $	Jerarquización ascendente	R_i^-	R_i^+
1	12,36	11,32	1,04	1,04	16		16
2	15,65	10,21	5,44	5,44	20		20
3	9,83	9,79	0,04	0,04	1		1
4	18,37	19,21	-0,84	0,84	15	15	
5	11,72	11,43	0,29	0,29	6		6
6	13,04	12,96	0,08	0,08	2		2
7	10,15	9,81	0,34	0,34	8		8
8	15,41	14,65	0,76	0,76	14		14
9	11,07	11,79	-0,72	0,72	13	13	
10	12,31	11,05	1,26	1,26	18		18
11	10,15	10,03	0,12	0,12	3		3
12	13,15	13,02	0,13	0,13	4		4
13	14,65	12,05	2,6	2,6	19		19
14	11,47	12,62	-1,15	1,15	17	17	
15	13,15	12,96	0,19	0,19	5		5
16	15,67	14,97	0,7	0,7	12		12
17	13,12	13,65	-0,53	0,53	9	9	
18	10,17	9,87	0,3	0,3	7		7
19	14,78	14,17	0,61	0,61	11		11
20	9,63	10,17	-0,54	0,54	10	10	
						64	146

$$T_+ = \sum R_i^+ = 146 \quad T_- = \sum R_i^- = 64$$

El valor experimental es $\hat{T}_+ = \sum R_i^+ = 146$

Como el valor experimental $\hat{T}_+ = 146 < 150 = T_{0,05, 20}$ se acepta la hipótesis nula de igualdad de estrategias de venta.

169. Una empresa inicialmente establece una jerarquía entre diez agentes vendedores de acuerdo a los resultados obtenidos en un test (el más idóneo es 1, el peor es 10). Un año después se replantea una nueva jerarquía con el mismo criterio de acuerdo al volumen de ventas anuales. Los resultados obtenidos se plasman en la tabla adjunta.

¿Existe alguna diferencia estadísticamente significativa entre ambos controles a un nivel de significación del 5%?

Jerarquía empresarial		
Vendedor	Test	Ventas
Alonso	4	1
Espinosa	5	6
Jiménez	2	2
Fernández	10	8
Crespo	7	9
García	1	3
Mateos	8	4
Cortes	3	5
Rodenas	9	10
Cabildo	7	7

Solución:

Se establecen las hipótesis:

H_0 : Los controles de capacitación son independientes

H_1 : Los controles de capacitación son dependientes

Se trata de un contraste bilateral o de dos colas

Se utiliza la prueba de Spearman Rho para aceptar o no la diferencia estadísticamente significativa entre los dos conjunto de observaciones de tipo cardinal u ordinal.

El coeficiente de correlación de Spearman Rho, al igual que el de Pearson, muestra una asociación entre variables que no se comportan normalmente, entre variables ordinales.

Procede de la siguiente manera:

a) Se calculan las diferencias D_i entre las 10 observaciones

$$6 \cdot \sum_{i=1}^n D_i^2$$

b) El coeficiente de Spearman ρ viene dado por la expresión $\rho = 1 - \frac{6 \cdot \sum_{i=1}^n D_i^2}{n \cdot (n^2 - 1)}$ c) El

coeficiente se compara con un coeficiente crítico $\rho_{\alpha, n}$ dado en la tabla siguiente:

Valores críticos $\rho_{\alpha,n}$ para la prueba de Spearman Rho

N	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
5	0.70	0.80	0.90
10	0.44	0.55	0.73
15	0.35	0.44	0.60
20	0.30	0.38	0.52
25	0.26	0.34	0.47
30	0.24	0.31	0.43

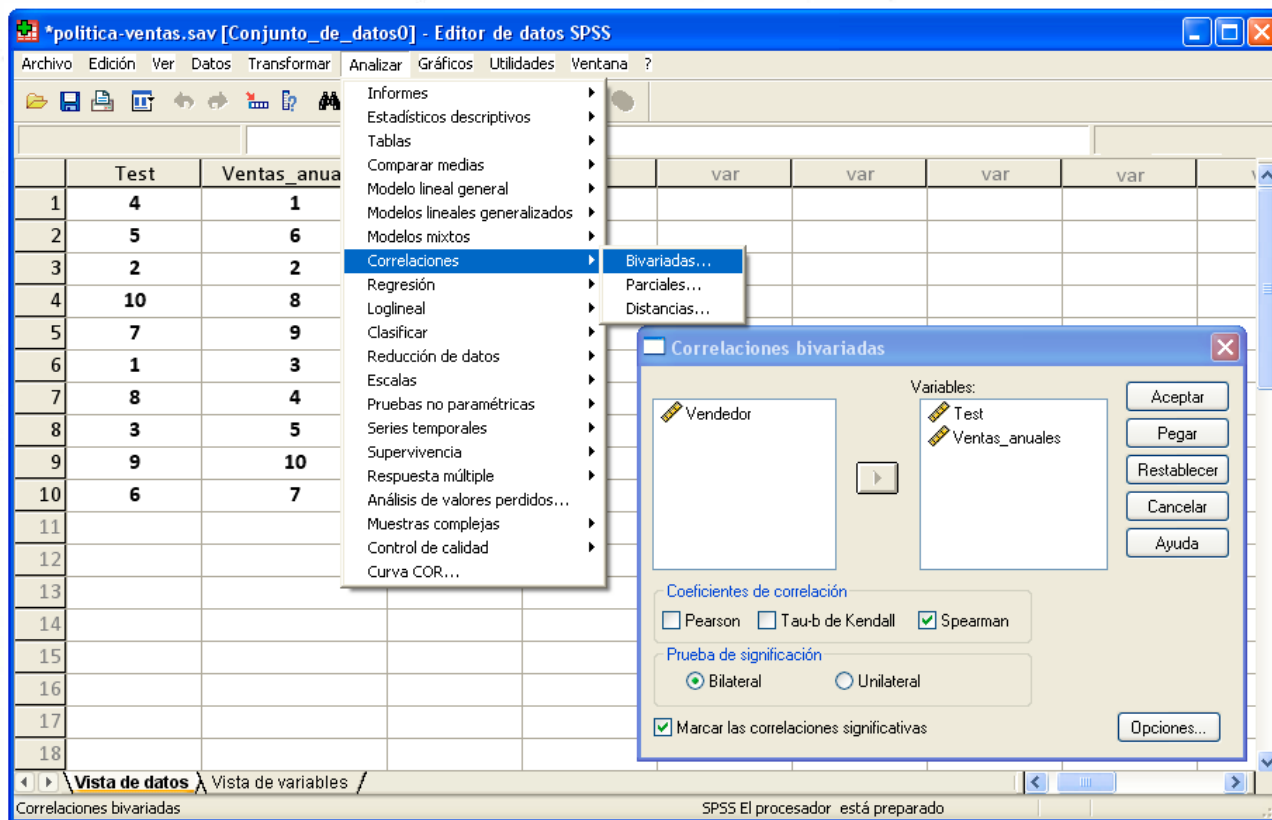
c) Si $\rho \leq \rho_{\alpha,n}$ se acepta la hipótesis nula de independencia significativa entre las observaciones.

En caso contrario, las observaciones son dependientes a un nivel de significación α

	Jerarquización Ordinal			
Vendedor	X	Y	$D_i = X - Y$	D_i^2
1	4	1	3	9
2	5	6	-1	1
3	2	2	0	0
4	10	8	2	4
5	7	9	-2	4
6	1	3	-2	4
7	8	4	4	16
8	3	5	-2	4
9	9	10	-1	1
10	6	7	-1	1
				$\sum_{i=1}^{10} D_i^2 = 44$

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^n D_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot \sum_{i=1}^{10} D_i^2}{10 \cdot (10^2 - 1)} = 0,733$$

Puesto que $\rho = 0,733 > 0,55 = \rho_{0,05,10}$ se rechaza la hipótesis nula admitiendo que las observaciones de control de capacitación son dependientes. En definitiva, la habilidad demostrada en el test está ligada a su desempeño en la vida real.



Correlaciones

			Test	Ventas_anuales
Rho de Spearman	Test	Coefficiente de correlación	1,000	,733*
		Sig. (bilateral)	.	,016
	Ventas_anuales	Coefficiente de correlación	,733*	1,000
		Sig. (bilateral)	,016	.
		N	10	10

*. La correlación es significativa al nivel 0,05 (bilateral).

El p_{valor} (Sig. bilateral) = 0,016 < 0,05 por lo que se rechaza la hipótesis nula, afirmando que los dos controles de capacitación están relacionados. Por otra parte, el coeficiente de correlación de Spearman es alto, con una correlación positiva, a medida que aumenta el nivel del test aumenta el volumen de ventas anual de los agentes.

170. De un cierto producto se tomaron 12 muestras lo más parecidas posibles y se dividieron en 5 familias con 4, 2, 2, 3, 1 muestras, que se almacenaron bajo diferentes condiciones. Se trata de ver con los datos de hidratación de la tabla adjunta si con un nivel de confianza del 95% hay diferencias significativas entre los métodos de almacenaje.

Método	1	2	3	4	5
	8,3	7,4	8,1	7,9	7,1
	7,6	7,1	6,4	9,5	
	8,4			10,0	
	8,3				

Solución:

Para asegurar que los datos de hidratación medios son iguales se aplica el modelo ANOVA de Fisher con un solo factor (método).

Bajo los supuestos de independencia de las variables, normalidad y homocedasticidad (la varianza del error condicional a las variables explicativas es constante a lo largo de las observaciones). Denotando por $N(\mu_1, \sigma_1)$, $N(\mu_2, \sigma_2)$, $N(\mu_3, \sigma_3)$, $N(\mu_4, \sigma_4)$ y $N(\mu_5, \sigma_5)$ cada una de los métodos de almacenaje, se contrastan las hipótesis:

H_0 : Los métodos son igualmente efectivos $\equiv \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

H_1 : Los tratamientos no son igualmente efectivos $\equiv \mu_i \neq \mu_j$ para algún $i \neq j$

Se acepta la hipótesis nula H_0 cuando el estadístico observado (empírico) $F_{(k-1), (n-k)} = \frac{\hat{S}_e^2}{\hat{S}_r^2}$ es menor o igual que el estadístico teórico $F_{\alpha, (k-1), (n-k)}$

$$\text{donde, } \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2}_{\hat{S}_y^2 = \frac{SCT}{n-1}} = \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}_{\hat{S}_r^2 = \frac{SCR}{n-k}} + \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2}_{\hat{S}_e^2 = \frac{SCE}{k-1}}$$

Cálculos		Tamaño muestra n_i	$\sum_{j=1}^{n_i} y_{ij}$	$\sum_{j=1}^{n_i} y_{ij}^2$	$\frac{1}{n_i} \left(\sum_{j=1}^{n_i} y_{ij} \right)^2$	$\bar{y}_{i\cdot} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$
1	8,3 - 7,6 - 8,4 - 8,3	4	32,6	266,1	265,69	8,2
2	7,4 - 7,1	2	14,5	105,17	105,13	7,3
3	8,1 - 6,4	2	14,5	106,57	105,13	7,3
4	7,9 - 9,5 - 10	3	27,4	252,66	250,25	9,1
5	7,1	1	7,1	50,41	50,41	7,1
		$\sum_{i=1}^5 n_i = 12$	$\sum_{i=1}^5 \sum_{j=1}^{n_i} y_{ij} = 96,1$	$\sum_{i=1}^5 \sum_{j=1}^{n_i} y_{ij}^2 = 780,91$	$\sum_{i=1}^5 \frac{\left(\sum_{j=1}^{n_i} y_{ij} \right)^2}{n_i} = 776,61$	$\bar{y}_{..} = 7,8$

$$\frac{\left[\sum_{i=1}^5 \sum_{j=1}^{n_i} y_{ij} \right]^2}{12} = \frac{(96,1)^2}{12} = 769,6$$

$$SCT = \sum_{i=1}^5 \sum_{j=1}^{n_i} y_{ij}^2 - \frac{\left(\sum_{i=1}^5 \sum_{j=1}^{n_i} y_{ij} \right)^2}{12} = 780,91 - 769,6 = 11,31 \quad \text{Suma cuadrados total}$$

$$SCE = \sum_{i=1}^5 \left[\frac{1}{n_i} \left(\sum_{j=1}^{n_i} y_{ij} \right)^2 \right] - \frac{\left(\sum_{i=1}^5 \sum_{j=1}^{n_i} y_{ij} \right)^2}{12} = 776,61 - 769,6 = 7,01$$

Suma de cuadrados explicados (entre los métodos) debida a que hay distintos niveles del factor.

$$\hat{s}_e^2 = \frac{SCE}{k-1} = \frac{7,01}{4} = 1,75 \quad \text{Variabilidad explicada en la muestra}$$

$$SCT = SCR + SCE \rightarrow SCR = 11,31 - 7,01 = 4,30$$

SCR = 4,30 Suma de cuadrados residuos (dentro grupos) variabilidad interna dentro de cada nivel

$$R^2 = \frac{SCE}{SCT} = \frac{7,01}{11,31} = 0,62 \quad \text{Coeficiente de determinación, porcentaje de variabilidad explicada}$$

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

Fuente de variación	Suma cuadrados desviaciones	Grados libertad	Variabilidad muestral	Test F
Explicada entre grupos	SCE = 7,01	$k - 1 = 4$	$\hat{S}_e^2 = \frac{SCE}{k - 1} = \frac{7,01}{4} = 1,75$	$F = \frac{\hat{S}_e^2}{\hat{S}_r^2} = 2,86$ $R^2 = \frac{SCE}{SCT} = 0,62$
Residual dentro grupos	SCR = 4,30	$n - k = 7$	$\hat{S}_r^2 = \frac{SCR}{n - k} = \frac{4,30}{7} = 0,61$	
Total	SCT = 11,31	$n - 1 = 11$	$\hat{S}_y^2 = \frac{SCT}{n - 1} = \frac{11,31}{11} = 1,03$	

Hipótesis nula H_0 : No existe diferencia significativa en los métodos de almacenaje.

Siendo $F = 2,86 < 4,1203 = F_{0,05; 4, 7}$ se acepta la hipótesis nula, afirmando que no hay diferencia significativa entre los métodos de almacenaje con una fiabilidad del 95% (nivel de confianza).

171. Se someten 24 muestras de agua a 4 tratamientos de descontaminación diferentes y asignados al azar. Para cada muestra se mide un indicador de la calidad del agua (cuanto más alto está el indicador, mayor es la calidad del agua). Con un nivel de confianza del 95% ¿son los tratamientos igualmente efectivos?.

Tratamiento	T1	T2	T3	T4
	62	63	68	56
	60	67	66	62
	63	71	71	60
	59	64	67	61
		65	68	63
		66	68	64
				63
				59

Solución:

En un principio parece que el tratamiento más efectivo es el 3. Para asegurar que los indicadores medios son iguales se aplica el modelo ANOVA de Fisher con un solo factor (tratamientos).

Bajo los supuestos de independencia de las variables, normalidad y homocedasticidad (la varianza del error condicional a las variables explicativas es constante a lo largo de las observaciones). Denotando por $N(\mu_1, \sigma_1)$, $N(\mu_2, \sigma_2)$, $N(\mu_3, \sigma_3)$, $N(\mu_4, \sigma_4)$ cada uno de los tratamientos, se contrastan las hipótesis:

H_0 : Los tratamientos son igualmente efectivos $\equiv \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1 : Los tratamientos no son igualmente efectivos $\equiv \mu_i \neq \mu_j$ para algún $i \neq j$

Se acepta la hipótesis nula H_0 cuando el estadístico observado (empírico) $F_{(k-1), (n-k)} = \frac{\hat{S}_e^2}{\hat{S}_r^2}$ es

menor o igual que el estadístico teórico $F_{\alpha, (k-1), (n-k)}$

$$\text{donde, } \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2}_{\hat{S}_y^2 = \frac{SCT}{n-1}} = \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2}_{\hat{S}_r^2 = \frac{SCR}{n-k}} + \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2}_{\hat{S}_e^2 = \frac{SCE}{k-1}}$$

$$SCT = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij}^2 - \frac{\left(\sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} \right)^2}{n}$$

$$SCE = \sum_{i=1}^k \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2 = \sum_{i=1}^k \frac{\left(\sum_{j=1}^{n_i} y_{ij} \right)^2}{n_i} - \frac{\left(\sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} \right)^2}{n}$$

Cálculos: k = 4		Tamaño muestra n_i	$\sum_{j=1}^{n_i} y_{ij}$	$\sum_{j=1}^{n_i} y_{ij}^2$	$\frac{1}{n_i} \left(\sum_{j=1}^{n_i} y_{ij} \right)^2$	$\bar{y}_{i.} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$
T1	62 - 60 - 63 - 59	4	244	14894	14884	61
T2	63 - 67 - 71 - 64 - 65 - 66	6	396	26176	26136	66
T3	68 - 66 - 71 - 67 - 68 - 68	6	408	27758	27744	68
T4	56 - 62 - 60 - 61 - 63 - 64 - 63 - 59	8	488	29816	29768	61
		$\sum_{i=1}^4 n_i = 24$	$\sum_{i=1}^4 \sum_{j=1}^{n_i} y_{ij} = 1536$	$\sum_{i=1}^4 \sum_{j=1}^{n_i} y_{ij}^2 = 98644$	$\sum_{i=1}^4 \frac{\left(\sum_{j=1}^{n_i} y_{ij} \right)^2}{n_i} = 98532$	$\bar{y}_{..} = 64$

$$\frac{\left[\sum_{i=1}^4 \sum_{j=1}^{n_i} y_{ij} \right]^2}{24} = \frac{1536^2}{24} = 98304$$

$$SCT = \sum_{i=1}^4 \sum_{j=1}^{n_i} y_{ij}^2 - \frac{\left(\sum_{i=1}^4 \sum_{j=1}^{n_i} y_{ij} \right)^2}{24} = 98644 - 98304 = 340 \quad \text{Suma cuadrados total}$$

$$SCE = \sum_{i=1}^4 \left[\frac{1}{n_i} \left(\sum_{j=1}^{n_i} y_{ij} \right)^2 \right] - \frac{\left(\sum_{i=1}^4 \sum_{j=1}^{n_i} y_{ij} \right)^2}{24} = 98532 - 98304 = 228$$

Suma de cuadrados explicados (entre tratamientos) debida a que hay distintos niveles del factor.

$$\hat{s}_e^2 = \frac{SCE}{k-1} = \frac{228}{3} = 76 \quad \text{Varianza explicada en la muestra}$$

$$SCT = SCR + SCE \rightarrow SCR = 340 - 228 = 112$$

SCR $\equiv$ Suma de cuadrados residuos (dentro grupos) variabilidad interna dentro de cada nivel

$$R^2 = \frac{SCE}{SCT} = \frac{228}{340} = 0,6706 \quad \text{Coeficiente de determinación, porcentaje de variabilidad explicada}$$

Fuente de variación	Suma cuadrados	Grados libertad	Variabilidad muestral	Test F
Explicada entre grupos	SCE = 228	k - 1 = 3	$\hat{s}_e^2 = \frac{SCE}{k-1} = \frac{228}{3} = 76$	$F = \frac{\hat{s}_e^2}{\hat{s}_r^2} = \frac{76}{5,6} = 13,571$ $R^2 = \frac{SCE}{SCT} = 0,6706$
Residual dentro grupos	SCR = 112	n - k = 20	$\hat{s}_r^2 = \frac{SCR}{n-k} = \frac{112}{20} = 5,6$	
Total	SCT = 340	n - 1 = 23	$\hat{s}_y^2 = \frac{SCT}{n-1} = \frac{340}{23} = 14,782$	

Siendo $F_{(4-1), (20-4)} = 13,571 > 3,0984 = F_{0,05; 3, 20}$ se rechaza la hipótesis nula, afirmando que no todos los tratamientos son iguales de efectivos.

Cuando se rechaza la hipótesis nula se busca qué grupos son distintos entre sí realizando comparaciones múltiples (Pruebas Post Hoc) dos a dos. Una de las opciones más utilizadas es el Test de Bonferroni.

Se construye un intervalo de confianza para la diferencia de medias poblacionales entre todas las posibles parejas de niveles en cada factor.

$$IC(\mu_i - \mu_j) = \left[(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) \pm t_{\alpha/2c, (n-k)} \hat{s}_r \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \right] \quad \text{Error típico} = \hat{s}_r \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

El test de Bonferroni fija un nivel de significación total $(\alpha / 2c)$, donde $c \equiv$ número de

contrastes, es decir, el número de tests $c = \binom{k}{2}$ que hay que realizar para comparar todos los pares de medias.

Señalar que puede ocurrir que se rechace la hipótesis nula H_0 en ANOVA y no se encuentren diferencias entre ningún par de medias aplicando el test de Bonferroni.

El test de Bonferroni es muy conservador, especialmente cuando 'c' es grande. Pueden aplicarse otros contrastes múltiples: Test de Tukey (cuando el diseño es equilibrado), test de Scheffé (cuando los tamaños muestrales son diferentes, coincide siempre con ANOVA), test de Dunnett (cuando hay un grupo control), test de Duncan, ..., etc

$$c = \binom{4}{2} = 6 \quad \frac{\alpha}{2c} = \frac{0,05}{12} = 0,005 \quad t_{0,005, (24-4)} = t_{0,005, 20} = 2,845 \quad \hat{s}_r = \sqrt{5,6} = 2,367$$

$$\text{IC}(\mu_1 - \mu_2) = \left[(61 - 66) \pm 2,845 \cdot 2,367 \sqrt{\frac{1}{4} + \frac{1}{6}} \right] = [-9,47, -0,53]$$

El intervalo no cubre el 0 y se rechaza la hipótesis de igualdad de medias, siendo mejor la calidad del agua en el tratamiento 2.

$$\text{IC}(\mu_1 - \mu_3) = \left[(61 - 68) \pm 2,845 \cdot 2,367 \sqrt{\frac{1}{4} + \frac{1}{6}} \right] = [-11,47, -2,53]$$

El intervalo no cubre el 0 y se rechaza la hipótesis de igualdad de medias, siendo mejor la calidad del agua en el tratamiento 3.

$$\text{IC}(\mu_1 - \mu_4) = \left[(61 - 61) \pm 2,845 \cdot 2,367 \sqrt{\frac{1}{4} + \frac{1}{8}} \right] = [-4,24, 4,24]$$

El intervalo cubre el 0 y se acepta la hipótesis de igualdad de medias, siendo igual la calidad del agua en los dos tratamientos.

$$\text{IC}(\mu_2 - \mu_3) = \left[(66 - 68) \pm 2,845 \cdot 2,367 \sqrt{\frac{1}{6} + \frac{1}{6}} \right] = [-6, 2]$$

El intervalo cubre el 0 y se acepta la hipótesis de igualdad de medias, siendo igual la calidad del agua en tratamientos 2 y 3.

En definitiva los tratamientos 2 y 3 ofrecen la misma y mejor calidad en el agua con un nivel de confianza del 95%.

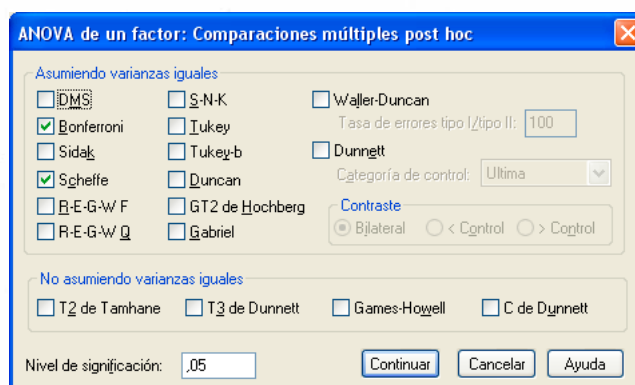
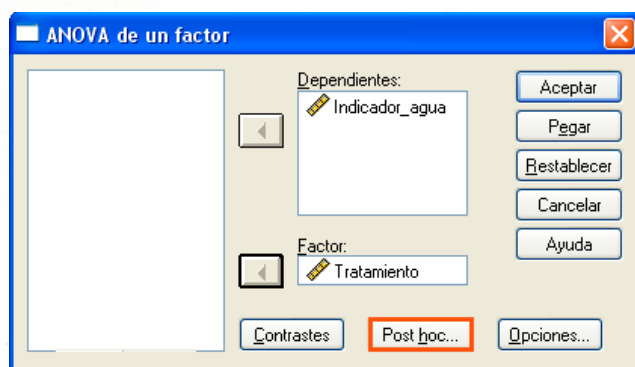
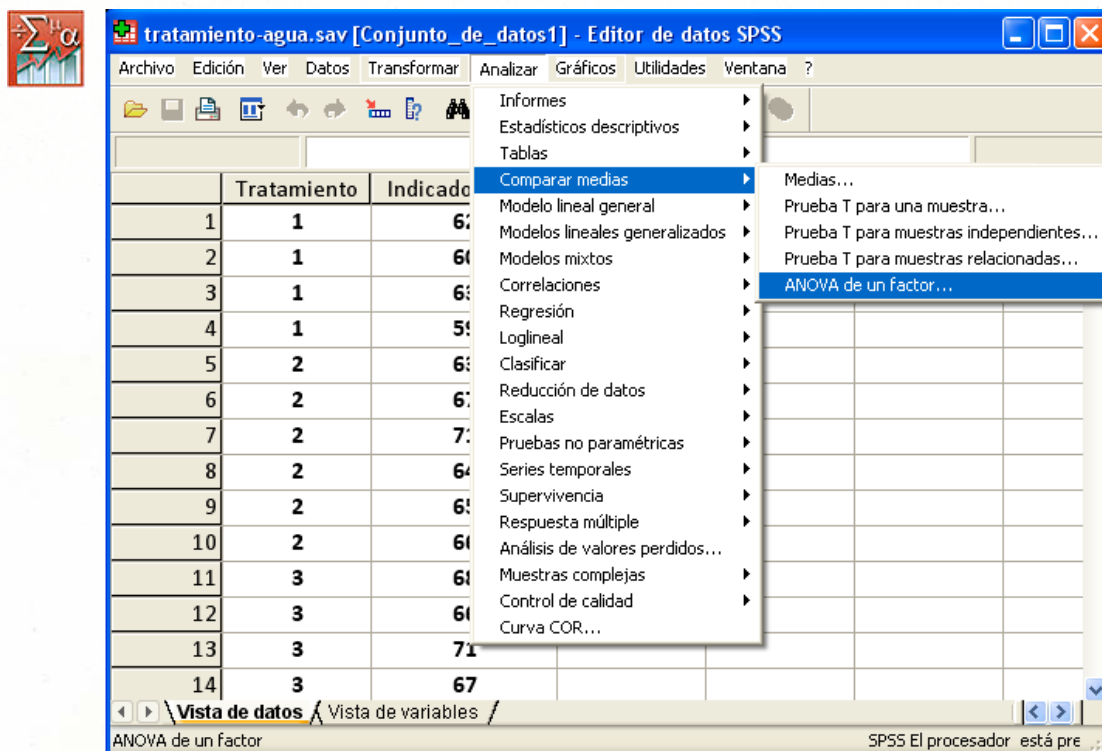
El método de comparaciones múltiples de Sheffé es muy conservador, basado en la distribución F, controla la tasa de error para el conjunto total de comparaciones que es posible diseñar con J medias. Se utiliza solo para efectuar solo comparaciones por pares, tiende a considerar significativas menos diferencias de las que debería.

$$\text{IC}(\mu_i - \mu_j) = \left[(\bar{y}_{i\cdot} - \bar{y}_{j\cdot}) \pm \sqrt{(k-1) \cdot F_{\alpha; (k-1), (n-k)}} \hat{s}_r \sqrt{\frac{1}{n_i} + \frac{1}{n_j}} \right] \quad \text{Error típico} = \hat{s}_r \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$$

$$\text{IC}(\mu_1 - \mu_2) = \left[(61 - 66) \pm \sqrt{3 \cdot F_{0,05; 3, 20}} \cdot 2,367 \sqrt{\frac{1}{4} + \frac{1}{6}} \right] = [-9,66, -0,34]$$

$$IC(\mu_1 - \mu_3) = \left[(61 - 68) \pm \sqrt{3 \cdot F_{0,05; 3, 20}} \cdot 2,367 \sqrt{\frac{1}{4} + \frac{1}{6}} \right] = [-11,66, -2,34]$$

$$IC(\mu_1 - \mu_4) = \left[(61 - 61) \pm \sqrt{3 \cdot F_{0,05; 3, 20}} \cdot 2,367 \sqrt{\frac{1}{4} + \frac{1}{8}} \right] = [-4,42, 4,42]$$



ANOVA

Indicador_agua	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	228,000	3	76,000	13,571	,000
Intra-grupos	112,000	20	5,600		
Total	340,000	23			

Siendo $p_valor \equiv Signatura = 0,00 < 0,05 \rightarrow$ Se rechaza H_0

En consecuencia, no todos los tratamientos son iguales de efectivos.

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

Variable dependiente: Indicador_agua
Bonferroni

Comparaciones múltiples

(I) Tratamiento	(J) Tratamiento	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
1	2	-5,000*	1,528	,023	-9,47	-,53
	3	-7,000*	1,528	,001	-11,47	-2,53
	4	,000	1,449	1,000	-4,24	4,24
2	1	5,000*	1,528	,023	,53	9,47
	3	-2,000	1,366	,953	-6,00	2,00
	4	5,000*	1,278	,005	1,26	8,74
3	1	7,000*	1,528	,001	2,53	11,47
	2	2,000	1,366	,953	-2,00	6,00
	4	7,000*	1,278	,000	3,26	10,74
4	1	,000	1,449	1,000	-4,24	4,24
	2	-5,000*	1,278	,005	-8,74	-1,26
	3	-7,000*	1,278	,000	-10,74	-3,26

*. La diferencia de medias es significativa al nivel .05.

Variable dependiente: Indicador_agua
Scheffé

Comparaciones múltiples

(I) Tratamiento	(J) Tratamiento	Diferencia de medias (I-J)	Error típico	Sig.	Intervalo de confianza al 95%	
					Límite inferior	Límite superior
1	2	-5,000*	1,528	,032	-9,66	-,34
	3	-7,000*	1,528	,002	-11,66	-2,34
	4	,000	1,449	1,000	-4,42	4,42
2	1	5,000*	1,528	,032	,34	9,66
	3	-2,000	1,366	,555	-6,17	2,17
	4	5,000*	1,278	,009	1,10	8,90
3	1	7,000*	1,528	,002	2,34	11,66
	2	2,000	1,366	,555	-2,17	6,17
	4	7,000*	1,278	,000	3,10	10,90
4	1	,000	1,449	1,000	-4,42	4,42
	2	-5,000*	1,278	,009	-8,90	-1,10
	3	-7,000*	1,278	,000	-10,90	-3,10

*. La diferencia de medias es significativa al nivel .05.

172. En un restaurante desean contratar a un camarero profesional cualificado. En una preselección se eligieron a cuatro personas de prueba. Para evaluar su rendimiento y el grado de satisfacción de los clientes se utilizó las cantidades entregadas como propinas. En la tabla adjunta se refleja el porcentaje de propina entregado sobre el importe total de la factura de los distintos clientes.

Camareros	1	2	3	4
	8	10	8,5	13
	7,5	9,5	5	10,5
	10,9	11,4	15	9,5
	10	10,3	11,1	12
	12	12	10,5	16
	6,5		7	8
	10		11	

Se supone que los porcentajes de propinas sobre el importe total de la factura se pueden considerar normalmente distribuidos. Con un nivel de significación del 10% se pide:

- Se puede aceptar la hipótesis de homocedasticidad de las distribuciones correspondientes al porcentaje de propina.
- Se pueden considerar los cuatro candidatos igualmente cualificados para el trabajo.

Solución:

- Para decidir si los candidatos se encuentran igualmente cualificados se aplica el modelo ANOVA de Fisher con un solo factor (propinas).

Las variables aleatorias $Y_i \equiv$ "Porcentaje de propina sobre el importe total de una factura recibida por un candidato" siguen distribuciones normales $Y_i \sim N(\mu_i, \sigma_i)$ y pueden considerarse independientes.

Antes de aplicar el modelo ANOVA hay que verificar que pueden considerarse homocedásticas (la varianza del error condicional a las variables explicativas es constante a lo largo de las observaciones), realizando para ello el test de Bartlett de homogeneidad de varianzas:

$$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 \quad H_1: \sigma_i^2 \neq \sigma_j^2 \text{ para algún } i \neq j$$

Se acepta H_0 si $B \leq \chi_{\alpha, (k-1)}^2$

$$B = \frac{(n-k) \cdot \ln \hat{s}_r^2 - \sum_{i=1}^k (n_i - 1) \cdot \ln \hat{s}_i^2}{1 + \frac{1}{3(k-1)} \cdot \left(\sum_{i=1}^k \frac{1}{n_i - 1} - \frac{1}{n-k} \right)} \approx \chi_{k-1}^2$$

$k \equiv$ grupos

$n_i \equiv$ tamaño de cada grupo

$\hat{s}_i^2 \equiv$ cuasivarianza de cada grupo

$\hat{s}_r^2 \equiv$ media ponderada de las cuasivarianzas

Para calcular el valor experimental del estadístico B hay que calcular las cuasivarianzas muestrales y el valor de la variabilidad residual dentro de los grupos.

1	2	3	4	
8	10	8,5	13	
7,5	9,5	5	10,5	
10,9	11,4	15	9,5	
10	10,3	11,1	12	
12	12	10,5	16	
6,5		7	8	
10		11		
$\sum_{i=1}^7 y_1 = 64,9$	$\sum_{i=1}^5 y_2 = 53,2$	$\sum_{i=1}^7 y_3 = 68,1$	$\sum_{i=1}^6 y_4 = 69$	$\sum_{i=1}^4 \sum_{j=1}^{n_i} y_{ij} = 255,2$
$\sum_{i=1}^7 y_1^2 = 625,31$	$\sum_{i=1}^5 y_2^2 = 570,3$	$\sum_{i=1}^7 y_3^2 = 725,71$	$\sum_{i=1}^6 y_4^2 = 833,5$	$\sum_{i=1}^4 \sum_{j=1}^{n_i} y_{ij}^2 = 2754,82$
$\hat{s}_1^2 = 3,9324$	$\hat{s}_2^2 = 1,063$	$\hat{s}_3^2 = 10,5324$	$\hat{s}_4^2 = 8$	
$\ln \hat{s}_1^2 = 1,3692$	$\ln \hat{s}_2^2 = 0,0611$	$\ln \hat{s}_3^2 = 2,3545$	$\ln \hat{s}_4^2 = 2,079$	

$$\hat{s}_i^2 = \frac{1}{n_i - 1} \sum_{i=1}^{n_i} (y_i - \bar{y}_i)^2 = \frac{1}{n_i - 1} \left[\sum_{i=1}^{n_i} y_i^2 - \frac{1}{n_i} \left(\sum_{j=1}^{n_i} y_i \right)^2 \right]$$

$$\hat{s}_1^2 = \frac{1}{6} \left[625,31 - \frac{64,9^2}{7} \right] = 3,9324 \quad \hat{s}_2^2 = \frac{1}{4} \left[570,3 - \frac{53,2^2}{5} \right] = 1,063$$

$$\hat{s}_3^2 = \frac{1}{6} \left[725,71 - \frac{68,1^2}{7} \right] = 10,5324 \quad \hat{s}_4^2 = \frac{1}{5} \left[833,5 - \frac{69^2}{6} \right] = 8$$

La variabilidad residual dentro de los grupos $\hat{s}_r^2$ es la media ponderada de la variabilidad de los grupos:

$$\hat{s}_r^2 = \frac{(n_1 - 1) \cdot \hat{s}_1^2 + (n_2 - 1) \cdot \hat{s}_2^2 + (n_3 - 1) \cdot \hat{s}_3^2 + (n_4 - 1) \cdot \hat{s}_4^2}{n_1 + n_2 + n_3 - 4}$$

$$\hat{s}_r^2 = \frac{6 \cdot 3,9324 + 4 \cdot 1,064 + 6 \cdot 10,5324 + 5 \cdot 8}{25 - 4} = 6,24023 \rightarrow \ln \hat{s}_r^2 = 1,831$$

$$B = \frac{(25-4) \cdot \ln 6,24 - \sum_{i=1}^4 (n_i - 1) \cdot \ln \hat{s}_i^2}{1 + \frac{1}{3(4-1)} \cdot \left(\sum_{i=1}^4 \frac{1}{n_i - 1} - \frac{1}{25-4} \right)} = \frac{21 \cdot 1,831 - [6 \cdot 1,3692 + 4 \cdot 0,0611 + 6 \cdot 2,3545 + 5 \cdot 2,079]}{1 + \frac{1}{3 \cdot 3} \cdot \left(\frac{1}{6} + \frac{1}{4} + \frac{1}{6} + \frac{1}{5} - \frac{1}{25-4} \right)} =$$

$$= \frac{5,4694}{1,0817} = 5,0563$$

Como $B = 5,0563 < 6,251 = \chi_{0,1,3}^3 = \chi_{\alpha, k-1}^2$ se acepta la hipótesis nula de homogeneidad de varianzas.

b) Se establecen las hipótesis:

H_0 : Los tratamientos son igualmente efectivos $\equiv \mu_1 = \mu_2 = \mu_3 = \mu_4$

H_1 : Los tratamientos no son igualmente efectivos $\equiv \mu_i \neq \mu_j$ para algún $i \neq j$

Se acepta la hipótesis nula H_0 cuando el estadístico observado (empírico) $F_{(k-1), (n-k)} = \frac{\hat{s}_e^2}{\hat{s}_r^2}$ es menor o igual que el estadístico teórico $F_{\alpha, (k-1), (n-k)}$

Se calculan las sumas de cuadrados y la variabilidad muestral

$$\underbrace{\text{SCT}}_{\hat{s}_y^2 = \frac{\text{SCT}}{n-1}} = \underbrace{\text{SCR}}_{\hat{s}_r^2 = \frac{\text{SCR}}{n-k}} + \underbrace{\text{SCE}}_{\hat{s}_e^2 = \frac{\text{SCE}}{k-1}}$$

$$\checkmark \quad \hat{s}_r^2 = \frac{\text{SCR}}{n-k} = \frac{\text{SCR}}{21} = 6,2402 \quad \text{variabilidad residual} \rightarrow \text{SCR} = 131,045$$

$$\checkmark \quad \alpha_1 = \bar{y} = \frac{1}{n} \sum_{i=1}^4 \sum_{j=1}^{n_i} y_{ij} = \frac{255,2}{25} = 10,208 \quad \alpha_2 = \frac{1}{n} \sum_{i=1}^4 \sum_{j=1}^{n_i} y_{ij}^2 = \frac{2754,82}{25} = 110,1928$$

$$\hat{\sigma}_y = \alpha_2 - \alpha_1^2 = 110,1928 - 10,208^2 = 5,9895 \rightarrow \hat{s}_y = \frac{25 \cdot 5,9895}{24} = 6,2391 \quad \text{variabilidad total}$$

$$\hat{s}_y = \frac{\text{SCT}}{n-1} \rightarrow \text{SCT} = 25 \cdot 5,9895 = 149,7375$$

$$\checkmark \quad \text{SCT} = \text{SCR} + \text{SCE} \rightarrow \text{SCE} = \text{SCT} - \text{SCR} = 149,7375 - 131,045 = 18,6925$$

$$\hat{s}_e = \frac{\text{SCE}}{k-1} = \frac{18,6925}{3} = 6,2308 \quad \text{variabilidad explicada}$$

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

Fuente de variación	Suma cuadrados	Grados libertad	Variabilidad muestral	Test F
Explicada entre grupos	SCE = 18,69	$k - 1 = 3$	$\hat{s}_e^2 = \frac{SCE}{k - 1} = \frac{18,69}{3} = 6,23$	$F = \frac{\hat{s}_e^2}{\hat{s}_r^2} = \frac{6,23}{6,24} = 0,999$ $R^2 = \frac{SCE}{SCT} = 0,125$
Residual dentro grupos	SCR = 131,04	$n - k = 21$	$\hat{s}_r^2 = \frac{SCR}{n - k} = \frac{131,04}{21} = 6,24$	
Total	SCT = 149,74	$n - 1 = 24$	$\hat{s}_y^2 = \frac{SCT}{n - 1} = \frac{149,74}{24} = 6,24$	

Siendo $F_{(4-1), (25-4)} = 0,999 < 2,3649 = F_{0,1; 3, 21}$ se acepta la hipótesis nula, afirmando que todos los candidatos se encuentran igualmente cualificados.

173. Un inversor tiene acciones en diferentes sectores del mercado continuo de valores. Después de quedarse sin asesor, para tener idea sobre la cotización de sus acciones seleccionó nueve sociedades en las que participaba y revisó sus cotizaciones de cierre el pasado viernes:

Sociedades	Cotización compra	Cotización actual
Acerinox	10,07	9,92
Bankinter	7,70	7,85
Bankia	3,98	4,37
Caixabank	2,56	1,44
Mapfre	3,05	3,05
BBVA	4,41	3,32
IAG	6,73	6,46
Telefónica	7,85	8,09
Mercadona	2,95	2,95

A un nivel de significación del 5%, ¿puede afirmarse que no hay diferencia significativa entre el valor de las acciones en las sociedades en las que participa este inversor?

Nota: Utilizar la prueba de rangos con signos de Wilcoxon

Solución:

La *prueba de rangos con signos de Wilcoxon* para dos muestras pareadas y la variable de respuesta ordinal o cuantitativa es la prueba homóloga no paramétrica a la prueba paramétrica *t* de Student para muestras pareadas.

Se utiliza como alternativa de la *prueba t de Student* cuando no se puede suponer la normalidad de las muestras.

Se establecen las hipótesis:

H_0 : No hay diferencias en valor de las acciones en las que participa el inversor

H_1 : El valor de las acciones actuales es inferior al valor de compra

Se introducen los datos en SPSS en Vista de variables:

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

*acciones-bolsa.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida
1	Sociedad	Numérico	8	0		{1, Acerinox}..	Ninguno	8	Centrado	Escala
2	Cotización_c	Numérico	8	2		Ninguno	Ninguno	11	Centrado	Escala
3	Cotización_ac	Numérico	8	2		Ninguno	Ninguno	12	Centrado	Escala
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										

Etiquetas de valor

Etiquetas de valor

Valor:

Etiqueta:

- 1 = "Acerinox"
- 2 = "Bankinter"
- 3 = "Bankia"
- 4 = "Caixabank"
- 5 = "Mapfre"
- 6 = "BBVA"
- 7 = "IAG"
- 8 = "Telefónica"
- 9 = "Mercadona"

Vista de datos Vista de variables /

SPSS El procesador está preparado

*acciones-bolsa.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

	Sociedad	Cotización_co
1	Acerinox	10,07
2	Bankinter	7,70
3	Bankia	3,98
4	Caixabank	2,56
5	Mapfre	3,05
6	BBVA	4,41
7	IAG	6,73
8	Telefónica	7,85
9	Mercadona	2,95
10		
11		
12		
13		
14		

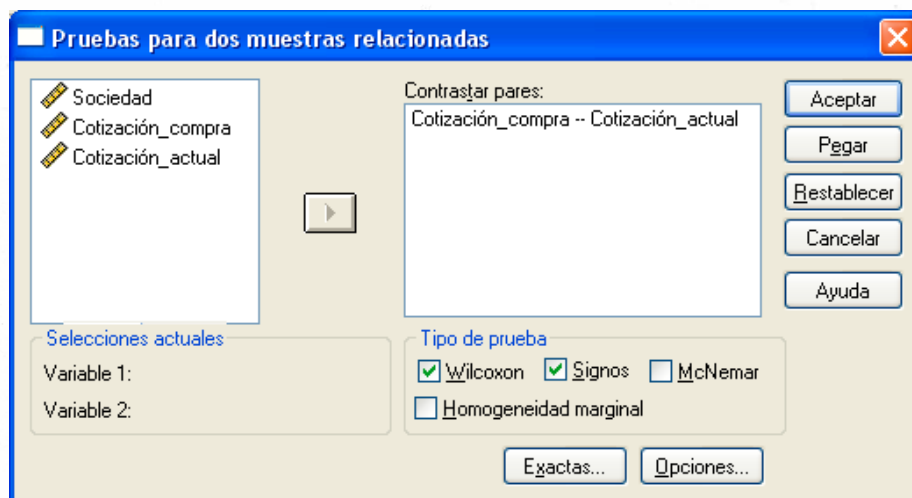
Informes
 Estadísticos descriptivos
 Tablas
 Comparar medias
 Modelo lineal general
 Modelos lineales generalizados
 Modelos mixtos
 Correlaciones
 Regresión
 Loglineal
 Clasificar
 Reducción de datos
 Escalas
Pruebas no paramétricas
 Series temporales
 Supervivencia
 Respuesta múltiple
 Análisis de valores perdidos...
 Muestras complejas
 Control de calidad
 Curva COR...

Chi-cuadrado...
 Binomial...
 Rachas...
 K-S de 1 muestra...
 2 muestras independientes...
 K muestras independientes...
2 muestras relacionadas...
 K muestras relacionadas...

Vista de datos Vista de variables /

2 muestras relacionadas

SPSS El procesador está preparado



Prueba de los rangos con signo de Wilcoxon

Rangos

		N	Rango promedio	Suma de rangos
Cotización_actual - Cotización_compra	Rangos negativos	4 ^a	4,63	18,50
	Rangos positivos	3 ^b	3,17	9,50
	Empates	2 ^c		
	Total	9		

a. Cotización_actual < Cotización_compra

b. Cotización_actual > Cotización_compra

c. Cotización_actual = Cotización_compra

Estadísticos de contraste^b

	Cotización_actual - Cotización_compra
Z	-,762 ^a
Sig. asintót. (bilateral)	,446

a. Basado en los rangos positivos.

b. Prueba de los rangos con signo de Wilcoxon

Como el p_valor (Sig. asintótica bilateral) = 0,446 > 0,05 se acepta la hipótesis nula y se concluye que no hay diferencias en el valor de las acciones de las sociedades en que participa el inversor.

$$z = \frac{T - \frac{n \cdot (n+1)}{4}}{\sqrt{\frac{n \cdot (n+1) \cdot (2n+1)}{24} - \sum_{i=1}^k \frac{t_i^3 - t_i}{48}}}$$

$$T_+ = \sum R_i^+ \quad T_- = \sum R_i^- \quad T = \min(T_+, T_-)$$

k ≡ número de rangos distintos en donde hay empates

t_i ≡ número puntuaciones empatadas en el rango i-ésimo



La prueba no paramétrica de Wilcoxon sirve para rechazar o no la diferencia estadísticamente significativa entre dos conjuntos de observaciones que tienen una distribución cualesquiera desconocida. La prueba consiste en lo siguiente:

- Se calculan las diferencias de dos observaciones
- Se jerarquiza en orden ascendente el valor absoluto de las diferencias
- Se suman las jerarquizaciones de las diferencias positivas, denominando a la suma con el símbolo T_+ = Suma de rangos de las $D_i > 0$
- Se calculan los valores críticos $\bar{T}$ de la tabla para el test de rangos-signos de Wilcoxon
- Si $T_+ < T_{\alpha, n}$ no existe diferencia significativa entre las observaciones a un nivel de significación α

La prueba de Wilcoxon supone que la distribución de las diferencias es continua, simétrica, independiente y que las diferencias tienen la misma mediana.

Sociedad	Cotización compra	Cotización actual	Diferencia D_i	Diferencia absoluta $ D_i $	Jerarquía ascendente	Rango Diferencias R_i		R_i^-	R_i^+
1	10,07	9,92	0,15	0,15	4	2	1,5	1,5	
2	7,7	7,85	-0,15	0,15	3	1	1,5		1,5
3	3,98	4,37	-0,39	0,39	7	5	5		5
4	2,56	1,44	1,12	1,12	9	7	7	7	
5	3,05	3,05	0	0	1				
6	4,41	3,32	1,09	1,09	8	6	6	6	
7	6,73	6,46	0,27	0,27	6	4	4	4	
8	7,85	8,09	-0,24	0,24	5	3	3		3
9	2,95	2,95	0	0	1				
								18,5	9,5

Para obtener los rangos, los valores absolutos de las diferencias $|D_i|$ se jerarquizan en orden ascendente (de menor a mayor) y se asignan los rangos o números de orden desde 1 hasta n (en este caso, $n = 1, \dots, 9$). Si existen valores de $|D_i|$ repetidos, el rango correspondiente será el promedio de los que se les asignarían si fueran diferentes.

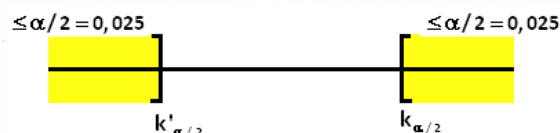
Cuando hay diferencias nulas no se tienen en cuenta, el tamaño muestral se reduce en el número de observaciones menos el número de diferencias nulas.

✓ En este caso, hay dos diferencias nulas y el tamaño muestral es $n = 9 - 7 = 2$, con lo que los rangos o números de orden van de 1 a 7. Por otra parte, la sociedad 1 y 2 presentan el mismo valor $|D_i|$ por lo que el rango de cada una de ellas será el promedio que se les asignarían si fueran diferentes $(1 + 2) / 2 = 1,5$

$$T_+ = \sum R_i^+ = 9,5 \quad T_- = \sum R_i^- = 18,5$$

El valor experimental es $\hat{T}_+ = \sum R_i^+ = 1,5 + 5 + 3 = 9,5$

Al tratarse de un contraste bilateral, la región crítica tendrá dos colas:



$$P(T_+ \leq k'_{\alpha/2, n}) \leq \alpha/2 \equiv P(T_+ \leq k'_{0,025, 7}) \leq 0,025 \rightarrow k'_{0,025, 7} = 2$$

$$P(T_+ \geq k_{\alpha/2, n}) \leq \alpha/2 \equiv P(T_+ \geq k_{0,025, 7}) \leq 0,05 = P(T_+ \leq k_{0,025, 7}) \leq 1 - 0,025 = 0,975 \rightarrow k_{0,025, 7} = 26$$

Así pues, como $k'_{0,025, 7} = 2 < \hat{T}_+ = 18,5 < 26 = k_{0,025, 7}$ se acepta la hipótesis nula, afirmando que no hay diferencias en el valor de las acciones de las sociedades en que participa el inversor.

Valores críticos para el test de rangos-signos de Wilcoxon

$$T_+ = \sum_{i=1}^n z_i \cdot \text{rango}(|D_i|)$$

α^* n	$P(T_+ \leq k_\alpha) \leq \alpha$					$P(T_+ \geq k_\alpha) \leq 1 - \alpha$				
	0,005	0,01	0,025	0,05	0,10	0,90	0,95	0,975	0,99	0,995
4	—	—	—	—	0	10	—	—	—	—
5	—	—	—	0	2	13	15	—	—	—
6	—	—	0	2	3	18	19	21	—	—
7	—	0	2	3	5	23	25	26	28	—
8	0	1	3	5	8	28	31	33	35	36
9	1	3	5	8	10	35	37	40	42	44
10	3	5	8	10	14	41	45	47	50	52
11	5	7	10	13	17	49	53	56	59	61
12	7	9	13	17	21	57	61	65	69	71
13	9	12	17	21	26	65	70	74	79	82
14	12	15	21	25	31	74	80	84	90	93
15	15	19	25	30	36	84	90	95	101	105

α^* no tiene por qué coincidir con el nivel de significación

174. Se emplean dos estrategias de ventas (A y B) de un mismo producto en veinte ciudades. El volumen de ventas anuales del mismo producto en cada ciudad (en miles de euros) por tipo de estrategia, está reflejado en la tabla adjunta. ¿Existe alguna diferencia significativa entre las estrategias de ventas a un nivel de significación del 5%?

Ventas (miles euros)			Ventas (miles euros)		
Ciudad	A	B	Ciudad	A	B
1	12,36	11,32	11	10,15	10,03
2	15,65	10,21	12	13,15	13,02
3	9,83	9,79	13	14,65	12,05
4	18,37	19,21	14	11,47	12,62
5	11,72	11,43	15	13,15	12,96
6	13,04	12,96	16	15,67	14,97
7	10,15	9,81	17	13,12	13,65
8	15,41	14,65	18	10,17	9,87
9	11,07	11,79	19	14,78	14,17
10	12,31	11,05	20	9,63	10,17

Solución:

Se trata de un contraste bilateral o de dos colas donde se establecen las hipótesis:

H_0 : No existe diferencia en la estrategia de ventas

H_1 : La estrategia de ventas son distintas

Se aplica la prueba de Wilcoxon suponiendo que la distribución de las diferencias en las estrategias de ventas es continua, simétrica, independiente y tienen la misma mediana.

Valores críticos $\bar{T}$ para la prueba de Wilcoxon

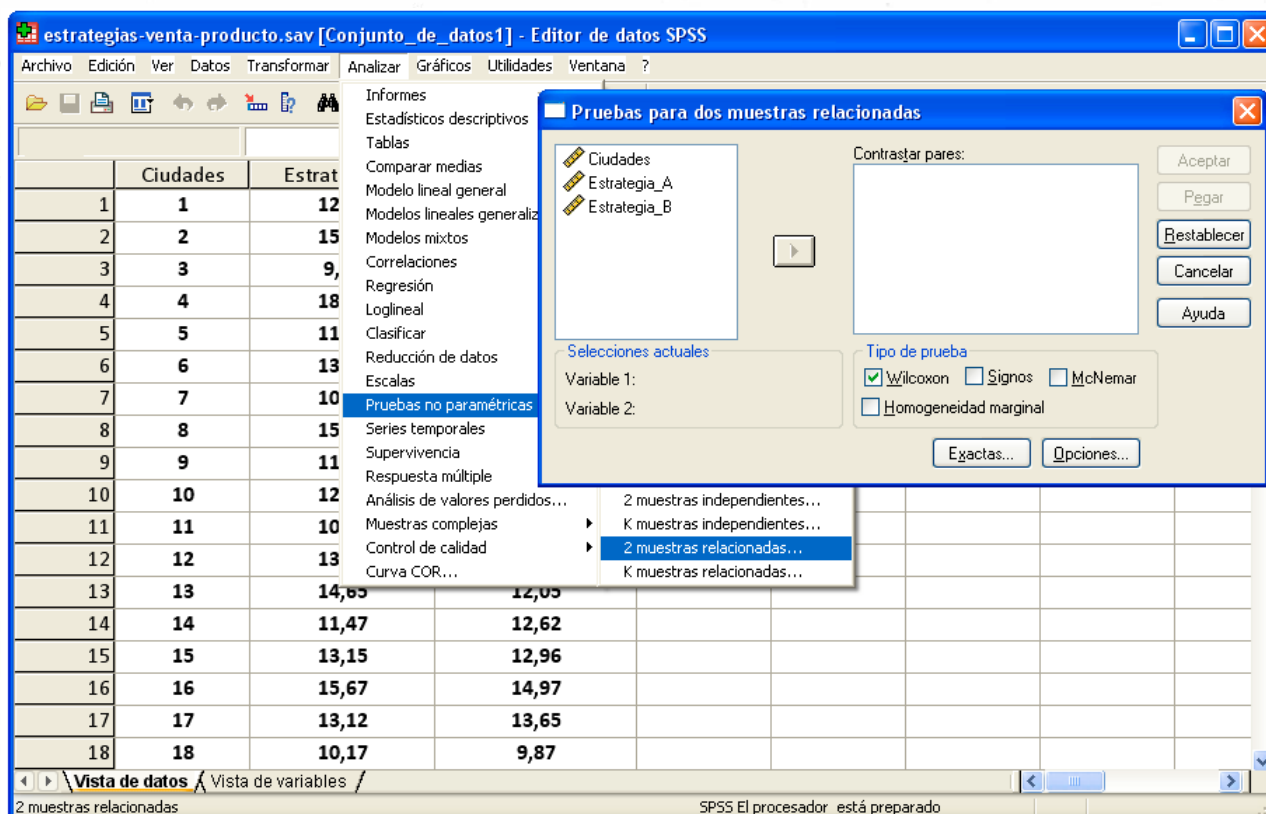
N	$0.01 < \alpha \leq 0.05$	$\alpha \leq 0.01$
5	14	—
6	19	—
7	21	28
8	30	34
9	37	42
10	44	50
12	61	68
14	79	89
16	100	112
18	124	138
20	150	167

Estrategias de ventas de un producto							
Ciudad	A	B	$D_i = A - B$	$ D_i $	Jerarquización ascendente	R_i^-	R_i^+
1	12,36	11,32	1,04	1,04	16		16
2	15,65	10,21	5,44	5,44	20		20
3	9,83	9,79	0,04	0,04	1		1
4	18,37	19,21	-0,84	0,84	15	15	
5	11,72	11,43	0,29	0,29	6		6
6	13,04	12,96	0,08	0,08	2		2
7	10,15	9,81	0,34	0,34	8		8
8	15,41	14,65	0,76	0,76	14		14
9	11,07	11,79	-0,72	0,72	13	13	
10	12,31	11,05	1,26	1,26	18		18
11	10,15	10,03	0,12	0,12	3		3
12	13,15	13,02	0,13	0,13	4		4
13	14,65	12,05	2,6	2,6	19		19
14	11,47	12,62	-1,15	1,15	17	17	
15	13,15	12,96	0,19	0,19	5		5
16	15,67	14,97	0,7	0,7	12		12
17	13,12	13,65	-0,53	0,53	9	9	
18	10,17	9,87	0,3	0,3	7		7
19	14,78	14,17	0,61	0,61	11		11
20	9,63	10,17	-0,54	0,54	10	10	
						64	146

$$T_+ = \sum R_i^+ = 146 \quad T_- = \sum R_i^- = 64$$

El valor experimental es $\hat{T}_+ = \sum R_i^+ = 146$

Como el valor experimental $\hat{T}_+ = 146 < 150 = T_{0,05, 20}$ se acepta la hipótesis nula de igualdad de estrategias de venta.



Prueba de los rangos con signo de Wilcoxon

Rangos

	N	Rango promedio	Suma de rangos
Estrategia_B - Estrategia_A Rangos negativos	15 ^a	9,73	146,00
Rangos positivos	5 ^b	12,80	64,00
Empates	0 ^c		
Total	20		

a. Estrategia_B < Estrategia_A

b. Estrategia_B > Estrategia_A

c. Estrategia_B = Estrategia_A

Estadísticos de contraste^b

	Estrategia_B - Estrategia_A
Z	-1,531 ^a
Sig. asintót. (bilateral)	,126

a. Basado en los rangos positivos.

b. Prueba de los rangos con signo de Wilcoxon

El p_valor (Sig. asintótica bilateral) = 0,126 > 0,05 con lo que se acepta la hipótesis nula, concluyendo que no existe diferencia en las estrategias de venta del producto con una significación de 0,05

175. Una empresa inicialmente establece una jerarquía entre diez agentes vendedores de acuerdo a los resultados obtenidos en un test (el más idóneo es 1, el peor es 10). Un año después se replantea una nueva jerarquía con el mismo criterio de acuerdo al volumen de ventas anuales. Los resultados obtenidos se plasman en la tabla adjunta. ¿Existe alguna diferencia estadísticamente significativa entre ambos controles a un nivel de significación del 5%?

Jerarquía empresarial		
Vendedor	Test	Ventas
Alonso	4	1
Espinosa	5	6
Jiménez	2	2
Fernández	10	8
Crespo	7	9
García	1	3
Mateos	8	4
Cortes	3	5
Rodenas	9	10
Cabildo	7	7

Solución:

Se establecen las hipótesis:

H_0 : Los controles de capacitación son independientes

H_1 : Los controles de capacitación son dependientes

Se trata de un contraste bilateral o de dos colas

Se utiliza la prueba de Spearman Rho para aceptar o no la diferencia estadísticamente significativa entre los dos conjunto de observaciones de tipo cardinal u ordinal.

El coeficiente de correlación de Spearman Rho, al igual que el de Pearson, muestra una asociación entre variables que no se comportan normalmente, entre variables ordinales.

Procede de la siguiente manera:

a) Se calculan las diferencias D_i entre las 10 observaciones

b) El coeficiente de Spearman ρ viene dado por la expresión $\rho = 1 - \frac{6 \cdot \sum_{i=1}^n D_i^2}{n \cdot (n^2 - 1)}$ c) El

coeficiente se compara con un coeficiente crítico $\rho_{\alpha, n}$ dado en la tabla siguiente:

Valores críticos $\rho_{\alpha,n}$ para la prueba de Spearman Rho

N	$\alpha = 0.10$	$\alpha = 0.05$	$\alpha = 0.01$
5	0.70	0.80	0.90
10	0.44	0.55	0.73
15	0.35	0.44	0.60
20	0.30	0.38	0.52
25	0.26	0.34	0.47
30	0.24	0.31	0.43

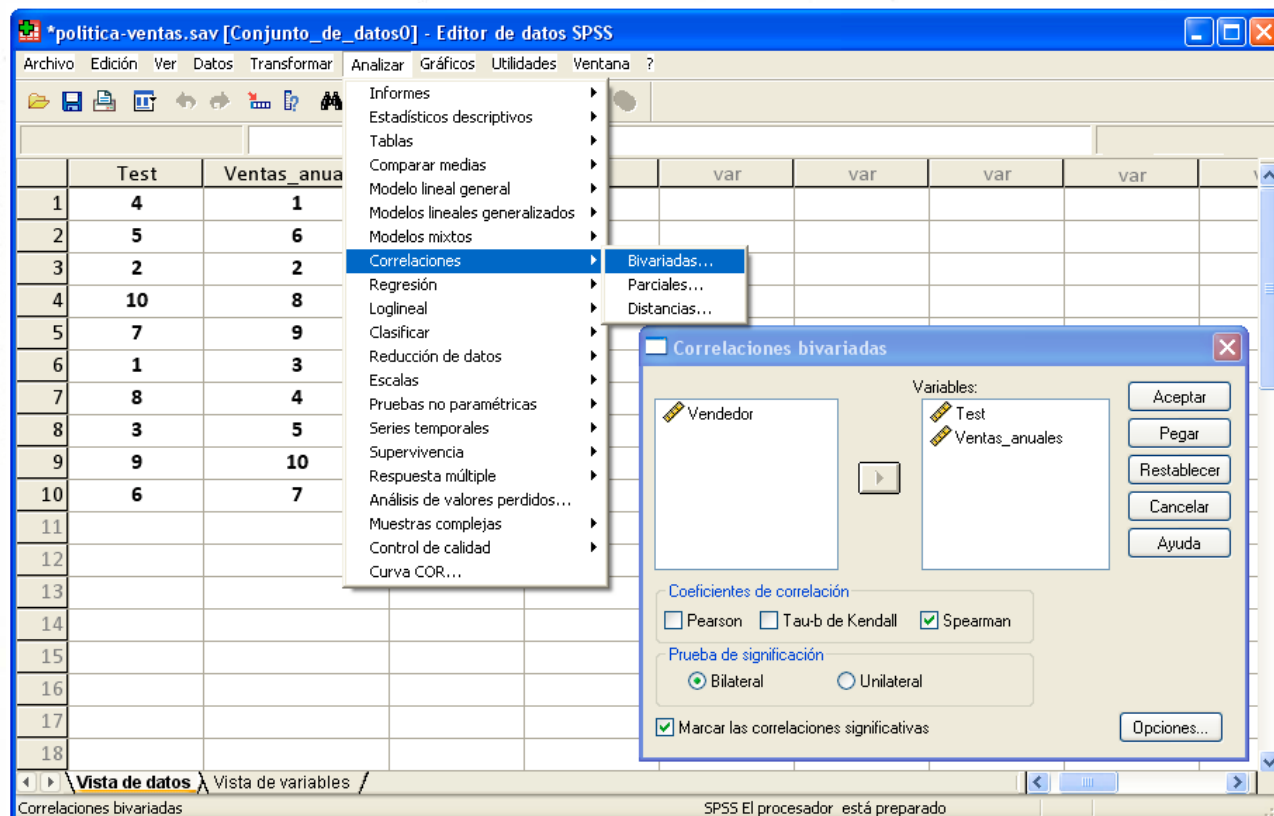
c) Si $\rho \leq \rho_{\alpha,n}$ se acepta la hipótesis nula de independencia significativa entre las observaciones.

En caso contrario, las observaciones son dependientes a un nivel de significación α

	Jerarquización Ordinal			
Vendedor	X	Y	$D_i = X - Y$	D_i^2
1	4	1	3	9
2	5	6	-1	1
3	2	2	0	0
4	10	8	2	4
5	7	9	-2	4
6	1	3	-2	4
7	8	4	4	16
8	3	5	-2	4
9	9	10	-1	1
10	6	7	-1	1
				$\sum_{i=1}^{10} D_i^2 = 44$

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^n D_i^2}{n \cdot (n^2 - 1)} = 1 - \frac{6 \cdot \sum_{i=1}^{10} D_i^2}{10 \cdot (10^2 - 1)} = 0,733$$

Puesto que $\rho = 0,733 > 0,55 = \rho_{0,05,10}$ se rechaza la hipótesis nula admitiendo que las observaciones de control de capacitación son dependientes. En definitiva, la habilidad demostrada en el test está ligada a su desempeño en la vida real.



Correlaciones

	Test	Ventas_anuales
Rho de Spearman	Coefficiente de correlación	1,000
	Sig. (bilateral)	,733*
	N	10
Ventas_anuales	Coefficiente de correlación	,733*
	Sig. (bilateral)	,016
	N	10

*. La correlación es significativa al nivel 0,05 (bilateral).

El p_{valor} (Sig. bilateral) = 0,016 < 0,05 por lo que se rechaza la hipótesis nula, afirmando que los dos controles de capacitación están relacionados. Por otra parte, el coeficiente de correlación de Spearman es alto, con una correlación positiva, a medida que aumenta el nivel del test aumenta el volumen de ventas anual de los agentes.

176. El director de una agencia de viajes, para diseñar un conjunto de estrategias publicitarias, está interesado en tener información sobre las edades de los compradores de un determinado paquete de viajes. Por este motivo, decide registrar la edad de las personas que compran este tipo de oferta de viajes, obteniendo la secuencia:

34 26 27 31 28 32 23 26 29 36
40 25 42 23 20 29 31 34 28 24
23 35 29 34 32 31 25 24 30 31
37 29 28 27

Con una significación del 5% se pide:

- Se admite la aleatoriedad en la serie de edades
- Observando los datos, ¿se puede afirmar que las personas que compran este paquete de viajes tienen al menos 30 años?

Solución:

- Pra verificar si la sucesión de edades es aleatoria, se realiza el contraste bilateral:

H_0 : La muestra es aleatoria H_1 : La muestra no es aleatoria

Para ello se acude a la prueba de rachas de Wald-Wolfowitz que permite verificar la hipótesis nula de que la muestra es aleatoria, es decir, si las sucesivas observaciones son independientes.

El estadístico de prueba del test de rachas de Wald-Wolfowitz es:

$R = \text{"Número de rachas en la muestra"}$

El contraste se basa en el número de rachas que presenta la muestra. Una racha es una secuencia de valores muestrales con una característica en común precedida y seguida por valores que no presentan esa característica.

En esta línea, se considera una racha la secuencia de k valores consecutivos superiores o iguales a un valor de corte (media, mediana, moda, ...) siempre que se encuentran precedidos y seguidos por valores inferiores a este valor de corte.

El número total de rachas en una muestra proporciona un indicio de si hay o no aleatoriedad en la muestra. Un número reducido de rachas (el caso extremo es 2) es indicio de que las observaciones no se han extraído de forma aleatoria, los elementos de la primera racha proceden de una población con una determinada característica (valores mayores o menores al punto de corte) mientras que los de la segunda proceden de otra población.

De forma idéntica un número excesivo de rachas puede ser también indicio de no aleatoriedad de la muestra.

Si la muestra es suficientemente grande y la hipótesis de aleatoriedad es cierta, la distribución muestral del número de rachas R puede aproximarse mediante una distribución normal de parámetros:

$$E(R) = \frac{2n_1 n_2}{n} + 1 \quad \text{Var}(R) = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}}$$

Se acepta la hipótesis nula cuando $z_R = \frac{|R + c - E(R)|}{\sqrt{\text{Var}(R)}} \leq z_{\alpha/2}$

$R < E(R) \rightarrow c = 0,5$
 $R > E(R) \rightarrow c = -0,5$

✓ Como las observaciones son cuantitativas, se construye una sucesión dicotómica asignando a cada observación el signo de su desviación respecto a la mediana muestral, es decir:

$$D_i = X_i - m_e$$

Como los datos $n = 34$ (par), la mediana una vez ordenados los datos en orden ascendente, será el valor medio de las observaciones centrales (lugares 17 y 18), con lo cual,

$$m_e = \frac{29 + 29}{2} = 29$$

El signo de $D_i = X_i - 29$ a partir de la serie original será:

+	-	-	+	-	+	-	-	0	+
+	-	+	-	-	0	+	+	-	-
-	+	0	+	+	+	-	-	+	+
+	0	-	-						

Como hay cuatro observaciones iguales a la mediana se ignoran y se reduce el tamaño de la muestra a $n = 34 - 4 = 30$ observaciones, donde:

$$n_1 = \text{"Números con signo positivo"} = 15 > 10$$

$$n_2 = \text{"Números con signo negativo"} = 15 > 10$$

Al ser $n_1, n_2 > 10$ se puede utilizar la aproximación normal:

$$E(R) = \frac{2n_1 n_2}{n} + 1 = \frac{2 \cdot 15 \cdot 15}{30} + 1 = 16$$

$$\text{Var}(R) = \sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}} = \sqrt{\frac{2 \cdot 15 \cdot 15 (2 \cdot 15 \cdot 15 - 30)}{30^2 \cdot (30-1)}} = 2,691$$

$$z_R = \frac{|R + c - E(R)|}{\sqrt{\text{Var}(R)}} = \frac{|16 - 16|}{2,691} = 0$$

Como $-z_{0,025} = -1,96 < z_R = 0 < 1,96 = z_{0,025}$

Se acepta la hipótesis nula de aleatoriedad de la serie de edades con una significación del 5%

📖 Cuando $n_1 > 10$ y $n_2 > 10$ se puede aplicar la distribución asintótica normal de Wald y Wolfowitz, denotando por $\gamma = \frac{n_1}{n}$ la proporción de observaciones de una categoría en la muestra de tamaño n se verifica que: $R \xrightarrow{d} N[2\gamma(1-\gamma)n, 2\gamma(1-\gamma)\sqrt{n}]$

En esta línea, $\gamma = \frac{n_1}{n} = \frac{15}{30} = 0,5$, quedando la distribución normal:

$$R \sim N[2 \cdot 0,5 \cdot (1-0,5) \cdot 30, 2 \cdot 0,5 \cdot (1-0,5)\sqrt{30}] \equiv N(15, 2,74)$$

Los valores críticos k_1 y k_2 se determinan con el valor de significación $\alpha = 0,05$:

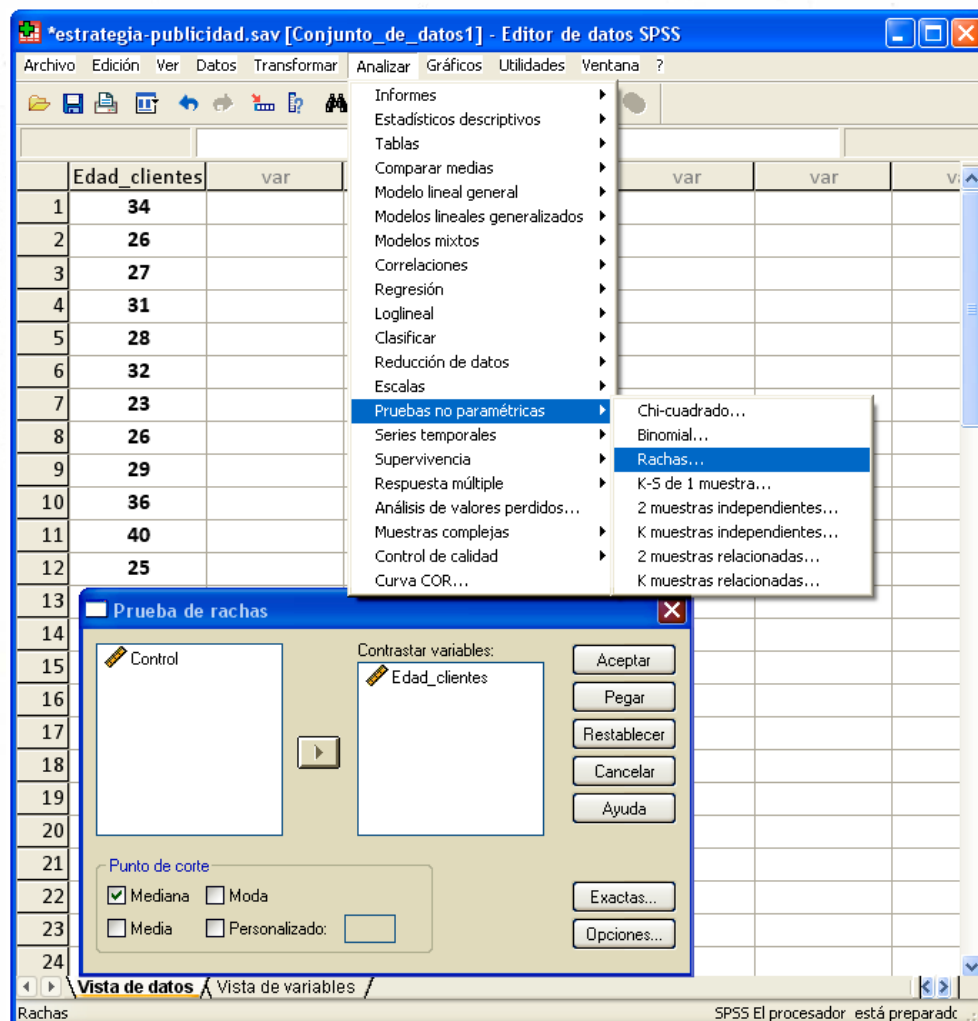
$$\begin{aligned} \alpha &= P[ET1] = P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P[R \mid N(15, 2,74)] = \\ &= P[(R \leq k_1) \cup (R \geq k_2) \mid N(15, 2,74)] = P\left[\frac{R-15}{2,74} \leq \frac{k_1-15}{2,74}\right] + P\left[\frac{R-15}{2,74} \geq \frac{k_2-15}{2,74}\right] = \\ &= P\left[z \leq \frac{k_1-15}{2,74}\right] + P\left[z \geq \frac{k_2-15}{2,74}\right] = 0,025 + 0,025 \end{aligned}$$

$$P\left[z \leq \frac{k_1-15}{2,74}\right] = P\left[z \geq \frac{-k_1+15}{2,74}\right] = 0,025 \rightarrow \frac{-k_1+15}{2,74} = 1,96 \rightarrow k_1 = 9,63$$

$$P\left[z \geq \frac{k_2-15}{2,74}\right] = 0,025 \rightarrow \frac{k_2-15}{2,74} = 1,96 \rightarrow k_2 = 20,37$$

El número de rachas $R = 16$ se encuentra en la región de aceptación: $9,63 < R = 16 < 20,37$ por lo que se acepta la hipótesis nula de aleatoriedad de las edades con una significación del 5%.

📖 Cuando hay más de 10 rachas y el tamaño de la muestra $n > 40$ se puede utilizar la aproximación asintótica de Levene: $R \xrightarrow{d} N\left[\frac{2n-1}{3}, \sqrt{\frac{16n-29}{90}}\right]$



Prueba de rachas

	Edad_clientes
Valor de prueba ^a	29
Casos < Valor de prueba	15
Casos >= Valor de prueba	19
Casos en total	34
Número de rachas	16
Z	-,447
Sig. asintót. (bilateral)	,655

a. Mediana

Como el p_{valor} (Sig. asintótica bilateral) = 0,655 > 0,05 = α se admite la hipótesis nula de aleatoriedad de la muestra.

b) El número de personas que compran el paquete de viajes con al menos 30 años de edad se puede analizar desde dos perspectivas diferentes: *contraste paramétrico* o *contraste no paramétrico*.

✓ En el *contraste paramétrico*, un supuesto común en donde aparecen una o más variables cuantitativas es que éstas se distribuyen normalmente en la población de referencia. El cumplimiento de este supuesto puede ser comprobado mediante la *prueba de Kolmogorov-Smirnov* o la prueba de *Shapiro-Wilk*.

No obstante, diversos estudios han puesto de manifiesto que la importancia de este supuesto es relativa a medida que el tamaño muestral es grande, en especial a partir de muestras mayores de 30 casos.

✓ El *contraste no paramétrico* es una prueba de significación que no presupone el cumplimiento de normalidad, si bien debe verificar dos supuestos: (a) Que la muestra sea representativa de la población objeto de estudio. (b) Que las observaciones sean independientes.

El *contraste no paramétrico* tiene la ventaja de ser menos exigente que el *contraste paramétrico* a nivel de supuestos, aunque presenta un aspecto negativo al ofrecer menor potencia estadística ($1 - \beta$) que la prueba paramétrica, es decir, con la prueba no paramétrica es menor la probabilidad de cometer un Error Tipo I (Rechazar la hipótesis nula H_0 cuando es Cierta).

Entre los contrastes no paramétricos: La *prueba de signos* y el *test de rangos-signos de Wilcoxon* permiten el contraste de hipótesis acerca de la localización de la variable en estudio respecto a la mediana, no respecto a la media como la prueba paramétrica t de Student.

El *test de rangos-signos de Wilcoxon* requiere que la variable a analizar sea cuantitativa y que la distribución de la misma sea simétrica, supuestos que no lo son para la *prueba de signos*, que tan sólo asume que la variable a analizar sea al menos ordinal.

Debido al carácter habitualmente asimétrico positivo de la variable $X \equiv$ "Edad de las personas que compran el paquete de viaje", así como a la simplicidad de los supuestos, se analiza en primer lugar la *prueba de signos*, conocida también como *prueba binomial*.

⇒ *Contraste de signos (prueba binomial)*

Denotando a la mediana poblacional por M_e , se contrasta si $M_e = m = 30$

Suponiendo que la variable aleatoria en estudio X es continua alrededor de M_e se establecen las hipótesis: $H_0: M_e = 30$ $H_1: M_e \neq 30$ (contraste bilateral o de dos colas)

Denotando por:

S_+ = "Número de signos positivos en la muestra"

S_- = "Número de signos negativos en la muestra"

Los signos positivos y negativos se calculan haciendo las diferencias $D_i = X_i - M_e$

El estadístico de prueba $S = \min(S_+, S_-)$ bajo la hipótesis nula H_0 sigue una distribución

binomial $S \sim B(n, 1/2)$ $E[S] = n \cdot p = \frac{n}{2}$ $V[S] = n \cdot p \cdot q = \frac{n}{4}$

Cuando n es suficientemente grande $n > 10 \rightarrow S \sim N\left(\frac{n}{2}, \frac{n}{4}\right)$

Se acepta H_0 cuando $z_p = \frac{\left| (S - 0,5) - \frac{n}{2} \right|}{\sqrt{\frac{n}{4}}} = \frac{\left| (S - 0,5) - 0,5 \cdot n \right|}{0,5 \cdot \sqrt{n}} \leq z_{\alpha/2}$

Se calculan las diferencias $D_i = X_i - 30$

+	-	-	+	-	+	-	-	-	+
+	-	+	-	-	-	+	+	-	-
-	+	-	+	+	+	-	-	0	+
+	-	-	-						

Como hay una observación igual a 0 se reduce el tamaño muestral a $n = 34 - 1 = 33 > 10$ observaciones.

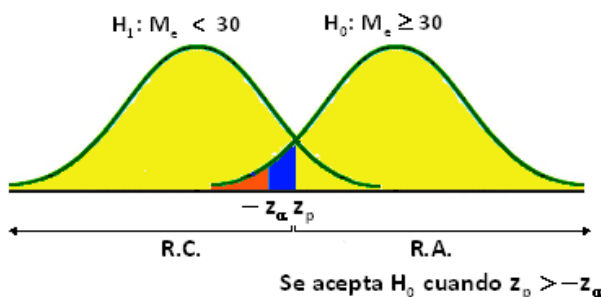
Al ser n suficientemente grande se puede utilizar la aproximación normal.

$S_+ = 14 > 10$ $S_- = 19 > 10$ $S = \min(14, 19) = 14$

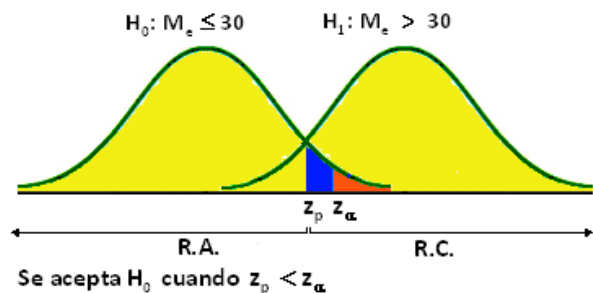
Siendo $z_p = \frac{\left| (14 - 0,5) - 0,5 \cdot 33 \right|}{0,5 \cdot \sqrt{33}} = 1,04 < 1,96 = z_{0,025}$

Se acepta la hipótesis nula $H_0: M_e = 30$, afirmando que la mitad de las personas que compran este tipo de viajes tienen 30 años.

Contraste unilateral a la izquierda (cola a la izquierda)



Contraste unilateral a la derecha (cola a la derecha)

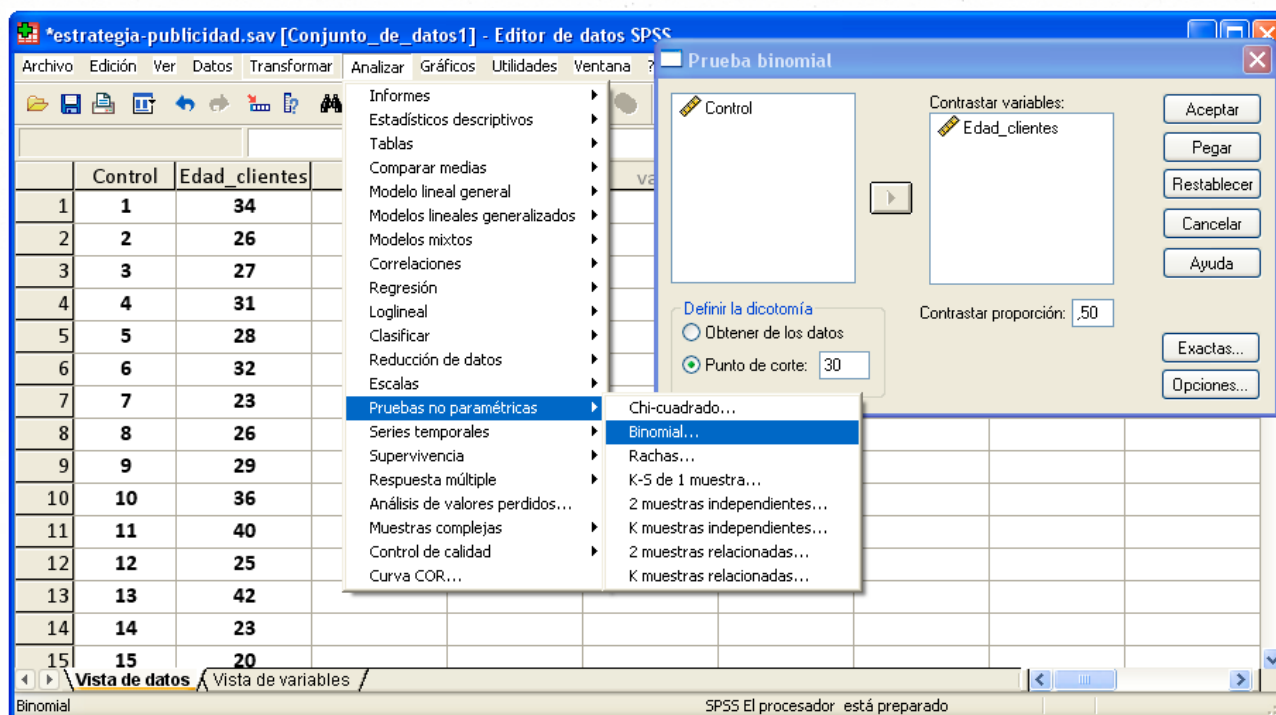


En el contraste unilateral a la izquierda (cola a la izquierda): $H_0: M_e \geq 30$ $H_1: M_e < 30$

Se acepta la hipótesis nula cuando $z_p = \frac{(S - 0,5) - 0,5 \cdot n}{0,5 \cdot \sqrt{n}} > -z_\alpha$

En consecuencia, siendo $z_p = \frac{(14 - 0,5) - 0,5 \cdot 33}{0,5 \cdot \sqrt{33}} = -1,04 > -1,645 = z_{0,05}$ se acepta la hipótesis

nula, en consecuencia la mitad de las personas que compran este tipo de viajes tienen al menos 30 años.



Prueba binomial

		Categoría	N	Proporción observada	Prop. de prueba	Sig. asintót. (bilateral)
Edad_clientes	Grupo 1	<= 30	20	,59	,50	,392 ^a
	Grupo 2	> 30	14	,41		
	Total		34	1,00		

a. Basado en la aproximación Z.

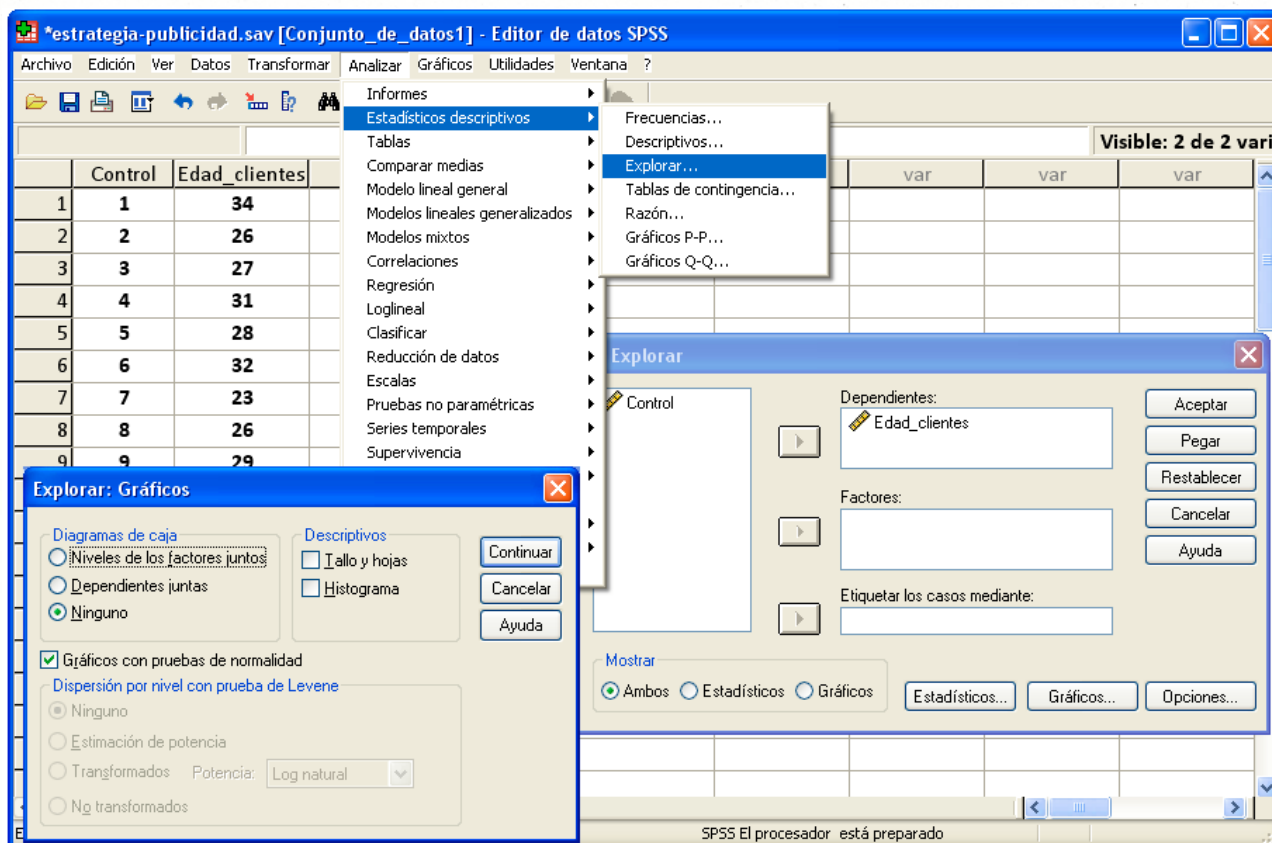
El p_{valor} (Sig. asintótica bilateral) = 0,392 > 0,05 = α con lo que se acepta la hipótesis nula $H_0: M_e = 30$ (contraste bilateral o de dos colas)

- Para realizar el contraste unilateral a la izquierda: $H_0: M_e \geq 30$ $H_1: M_e < 30$

$$p_{\text{valor}} (\text{Sig. asintótica unilateral izquierda}) = 1 - \frac{0,392}{2} = 0,804 > 0,05 = \alpha \rightarrow$$

concluyendo que las personas que compran este paquete de viajes tienen al menos 30 años.

Si se hubiera optado por un contraste paramétrico de la variable $X \equiv$ "Edad de las personas que compran el paquete de viaje" se tendría que haber realizado primero un contraste de normalidad de los datos obtenidos.



Descriptivos

			Estadístico	Error típ.
Edad_clientes	Media		29,50	,866
	Intervalo de confianza para la media al 95%	Límite inferior	27,74	
		Límite superior	31,26	
	Media recortada al 5%		29,31	
	Mediana		29,00	
	Varianza		25,470	
	Desv. típ.		5,047	
	Mínimo		20	
	Máximo		42	
	Rango		22	
	Amplitud intercuartil		7	
	Asimetría		,493	,403
	Curtosis		,086	,788

Pruebas de normalidad

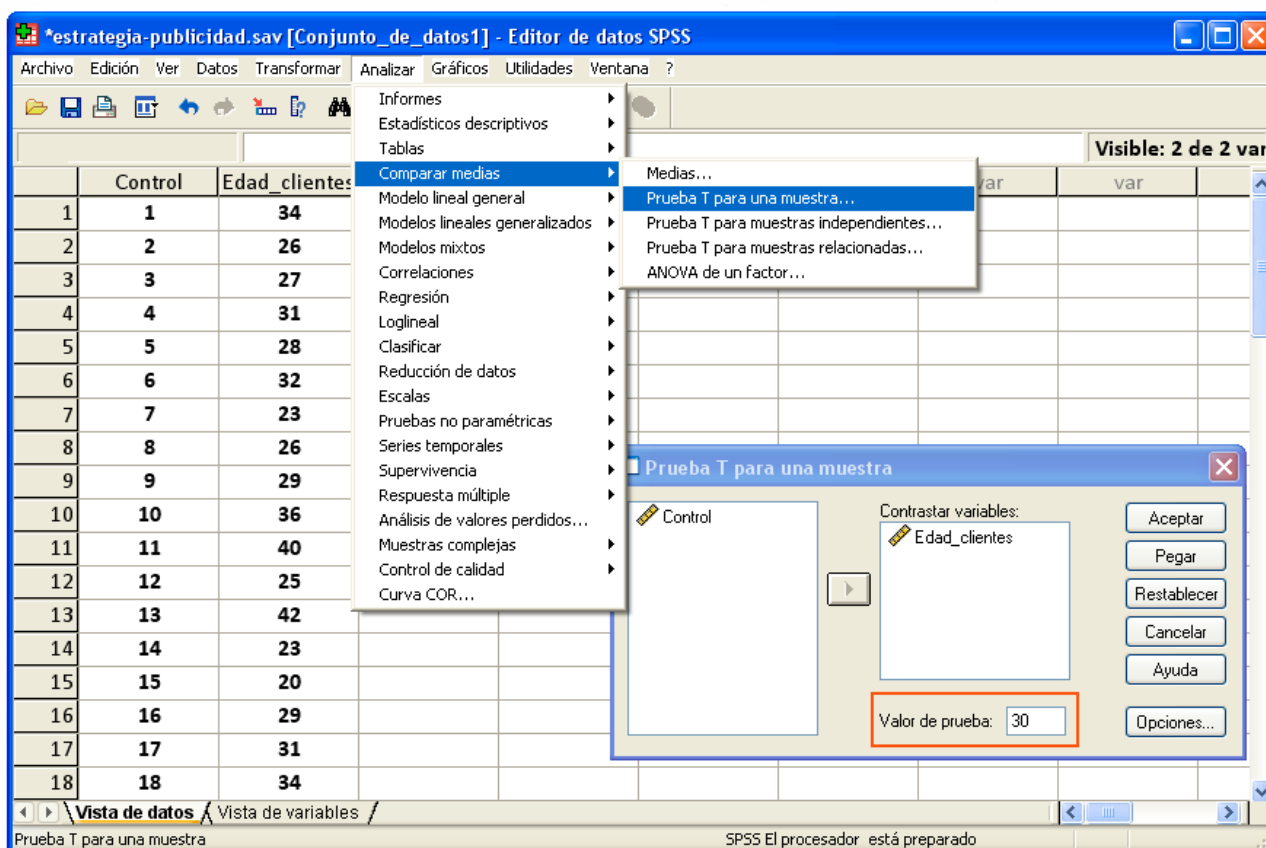
	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Edad_clientes	,098	34	,200*	,976	34	,638

*. Este es un límite inferior de la significación verdadera.

a. Corrección de la significación de Lilliefors

En ambos contrastes (Kolmogorov-Smirnov y Shapiro-Wilk) el p_valor (Signatura) es mayor que $\alpha = 0,05$ en consecuencia se admite la hipótesis de normalidad.

Realizando ahora el contraste paramétrico: $H_0: \mu \geq 30$ $H_1: \mu < 30$



The screenshot shows the SPSS 'Prueba T para una muestra' (One-Sample T-Test) dialog box. The 'Contrastar variables' (Contrast variables) list contains 'Edad_clientes'. The 'Valor de prueba' (Test value) is set to 30. The background shows a data table with columns 'Control' and 'Edad_clientes'.

Prueba para una muestra

	Valor de prueba = 30					
	t	gl	Sig. (bilateral)	Diferencia de medias	95% Intervalo de confianza para la diferencia	
					Inferior	Superior
Edad_clientes	-,578	33	,567	-,500	-2,26	1,26

El p_valor (Sig. bilateral) = 0,567 > 0,05 = α con lo que se admite la hipótesis nula del contraste bilateral $H_0: \mu = 30$, es decir, las personas que compran el paquete de viajes tienen 30 años.

El contraste que se desea realizar $H_0: \mu \geq 30$ $H_1: \mu < 30$ es un contraste unilateral a la izquierda (cola a la izquierda):

p_valor (Sig. unilateral izquierda) $= 1 - \frac{0,567}{2} = 0,7165 > 0,05 = \alpha \rightarrow$ Se acepta la hipótesis nula concluyendo que las personas que compran el paquete de viajes tienen al menos 30 años

⇒ Contraste de rangos-signos de Wilcoxon

Para aplicar el contraste es necesario suponer que la variable aleatoria X es continua y simétrica respecto a su mediana.

Se establece el contraste bilateral: $H_0: M_e = 30$ $H_1: M_e \neq 30$

Estadístico de prueba: $T_+ = \text{Suma de los rangos positivos de } D_i = X_i - M_e = \sum R_i^+$

Cuando el tamaño muestral $n > 15$ se puede utilizar la aproximación normal, y en lugar de

$T_+ = \sum R_i^+$, utilizar como estadístico de contraste $z_p = \frac{|T_+ - E[T_+]|}{\sqrt{\text{Var}[T_+]}} \xrightarrow[n \rightarrow \infty]{H_0} N(0,1)$

Se admite H_0 si $z_p = \frac{|T_+ - E[T_+]|}{\sqrt{\text{Var}[T_+]}} \leq z_{\alpha/2}$

El cálculo del valor experimental del estadístico, se obtiene:

$$E[T_+] = \frac{n \cdot (n+1)}{4} \quad \text{Var}[T_+] = \frac{n \cdot (n+1) \cdot (2n+1)}{24}$$

A continuación, para facilitar los cálculos necesarios para hallar el estadístico de contraste, se elabora la tabla siguiente:

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

1	2	3	4	5	6	7	8
Orden x_i	$D_i = x_i - 30$	x_i	$ D_i $	Posición $ D_i $	Rango R	R_i^-	R_i^+
20	-10	20	10	31	31,5	31,5	
23	-7	23	7	27	28,5	28,5	
23	-7	23	7	28	28,5	28,5	
23	-7	23	7	29	28,5	28,5	
24	-6	24	6	24	25	25	
24	-6	24	6	25	25	25	
25	-5	25	5	21	22	22	
25	-5	25	5	22	22	22	
26	-4	26	4	16	18	18	
26	-4	26	4	17	18	18	
27	-3	27	3	14	14,5	14,5	
27	-3	27	3	15	14,5	14,5	
28	-2	28	2	9	11	11	
28	-2	28	2	10	11	11	
28	-2	28	2	11	11	11	
29	-1	29	1	1	4,5	4,5	
29	-1	29	1	2	4,5	4,5	
29	-1	29	1	3	4,5	4,5	
29	-1	29	1	4	4,5	4,5	
30	0	31	1	5	4,5		4,5
31	1	31	1	6	4,5		4,5
31	1	31	1	7	4,5		4,5
31	1	31	1	8	4,5		4,5
31	1	32	2	12	11		11
32	2	32	2	13	11		11
32	2	34	4	18	18		18
34	4	34	4	19	18		18
34	4	34	4	20	18		18
34	4	35	5	23	22		22
35	5	36	6	26	25		25
36	6	37	7	30	28,5		28,5
37	7	40	10	32	31,5		31,5
40	10	42	12	33	33		33
42	12					327	234

- En la primera columna se ordenan los datos muestrales de menor a mayor
- La segunda columna son las diferencias entre los datos muestrales ordenados y la mediana

 Se reinicia el proceso

- La tercera columna son los datos muestrales ordenados, sin tener en cuenta los datos con diferencias nulas $|D_i| = 0$
- En la cuarta columna las diferencias absolutas $|D_i| = |x_i - 30| \neq 0$
- En la quinta columna la posición de $|D_i|$ en orden ascendente, empezando por la más pequeña.
- En la sexta columna los rangos de las diferencias $|D_i|$, cuando hay diferencias repetidas se asigna el rango medio de ellas.

Excel: JERARQUIA.MEDIA(Dato, Columna valor absoluto datos, Columna datos ordenados)

La diferencia $|D_i| = 10$ está repetida 2 veces, el rango se obtiene: $\frac{31+32}{2} = 31,5$

La diferencia $|D_i| = 1$ está repetida 8 veces, el rango se obtiene: $\frac{1+2+3+4+5+6+7+8}{8} = 4,5$

La diferencia $|D_i| = 3$ está repetida 2 veces, el rango se obtiene: $\frac{14+15}{2} = 14,5$

La diferencia $|D_i| = 4$ está repetida 5 veces, el rango se obtiene: $\frac{16+17+18+19+20}{5} = 18$

- En la séptima y octava columna aparecen, respectivamente, los rangos negativos y positivos.

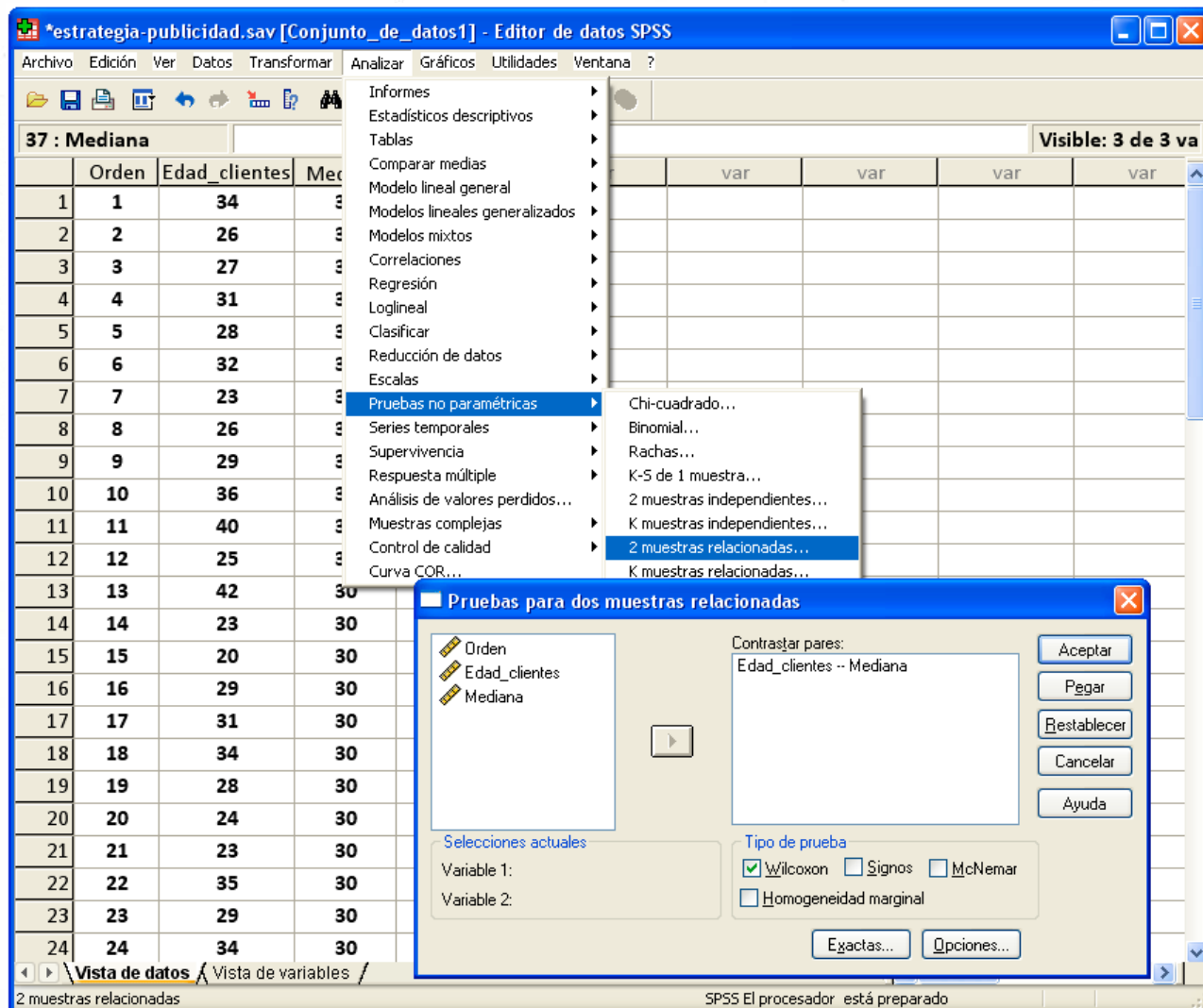
$$T_- = \sum R_i^- = 327 \quad T_+ = \sum R_i^+ = 234$$

Como $n = 33 > 15$, utilizando la aproximación normal, el estadístico de contraste:

$$E[T_+] = \frac{n \cdot (n+1)}{4} = \frac{33 \cdot 34}{4} = 280,5 \quad \text{Var}[T_+] = \frac{n \cdot (n+1) \cdot (2n+1)}{24} = \frac{33 \cdot 34 \cdot 67}{24} = 3132,25$$

$$z_p = \frac{|T_+ - E[T_+]|}{\sqrt{\text{Var}[T_+]}} = \frac{|234 - 280,5|}{\sqrt{3132,25}} = 0,831$$

Siendo $z_p = 0,831 < 1,96 = z_{0,025}$ se admite la hipótesis nula $H_0: M_e = 30$, coincidiendo con el diagnóstico del test de signos de la mediana.



*estrategia-publicidad.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

37 : Mediana

	Orden	Edad_clientes	Mediana
1	1	34	30
2	2	26	30
3	3	27	30
4	4	31	30
5	5	28	30
6	6	32	30
7	7	23	30
8	8	26	30
9	9	29	30
10	10	36	30
11	11	40	30
12	12	25	30
13	13	42	30
14	14	23	30
15	15	20	30
16	16	29	30
17	17	31	30
18	18	34	30
19	19	28	30
20	20	24	30
21	21	23	30
22	22	35	30
23	23	29	30
24	24	34	30

Pruebas para dos muestras relacionadas

Contrastar pares:
Edad_clientes -- Mediana

Tipo de prueba
☒ Wilcoxon ☐ Signos ☐ McNemar
☐ Homogeneidad marginal

Exactas... Opciones...

Prueba de los rangos con signo de Wilcoxon

Rangos

		N	Rango promedio	Suma de rangos
Mediana - Edad_clientes	Rangos negativos	14 ^a	16,71	234,00
	Rangos positivos	19 ^b	17,21	327,00
	Empates	1 ^c		
	Total	34		

a. Mediana < Edad_clientes

b. Mediana > Edad_clientes

c. Mediana = Edad_clientes

Estadísticos de contraste^b

	Mediana - Edad_clientes
Z	-,833 ^a
Sig. asintót. (bilateral)	,405

a. Basado en los rangos negativos.

b. Prueba de los rangos con signo de Wilcoxon

Se admite la hipótesis nula $M_e = 30$ al presentar un p_valor (Sig. asintótica bilateral) = $0,405 > 0,05 = \alpha$

177. Un cuestionario sobre innovación empresarial, entre otros aspectos, recoge datos sobre la edad (x_i) y los años de experiencia profesional (y_i), obteniendo la siguiente tabla:

x_i	35	37	42	31	33	47	52	50	60	39	43	34	34	41	49	40	34	57	53	32	45
y_i	5	7	10	6	8	12	17	16	23	7	13	4	9	11	12	14	9	21	28	7	10

- a) Contrastar que la edad de innovadores es una muestra aleatoria simple de una población continua y simétrica, con mediana igual o mayor que 40 para un nivel de significación del 5%.
- b) Contrastar que la experiencia profesional es una muestra aleatoria simple de una población continua y simétrica, con mediana 16 para un nivel de significación del 5%.

Solución:

- a) Como los datos proceden de una población continua y simétrica se puede emplear la prueba de signos de Wilcoxon, estableciendo el contraste unilateral a la izquierda:

$$H_0: M_e \geq 40 \quad H_1: M_e < 40$$

A simple vista se observa que la decimosexta observación es igual a la mediana, la diferencia es cero, con lo que el tamaño muestral se reduce una unidad, pasando a ser $n = 20$

El tamaño muestral $n > 15$ pudiendo utilizar la aproximación normal.

Se admite H_0 si $z_p = \frac{T_+ - E[T_+]}{\sqrt{\text{Var}[T_+]}} \geq -z_\alpha$

Para determinar el estadístico de contraste se recurre a la tabla siguiente:

- En la primera columna se ordenan los datos muestrales de menor a mayor
- La segunda columna son las diferencias entre los datos muestrales ordenados y la mediana
- En la tercera columna las diferencias absolutas $|D_i| = |x_i - 40|$
- En la cuarta columna la posición de $|D_i|$ en orden ascendente, empezando por la más pequeña.
- En la quinta columna los rangos de las diferencias $|D_i|$, cuando hay diferencias repetidas se asigna el rango medio de ellas.

Excel: JERARQUIA.MEDIA(Dato, Columna valor absoluto datos, Columna datos ordenados)

La diferencia $|D_i| = 1$ está repetida 2 veces, el rango se obtiene: $\frac{1+2}{2} = 1,5$

La diferencia $|D_i| = 3$ está repetida 2 veces, el rango se obtiene: $\frac{4+5}{2} = 4,5$

La diferencia $|D_i| = 5$ está repetida 2 veces, el rango se obtiene: $\frac{6+7}{2} = 6,5$

La diferencia $|D_i| = 6$ está repetida 3 veces, el rango se obtiene: $\frac{8+9+10}{3} = 9$

La diferencia $|D_i| = 7$ está repetida 2 veces, el rango se obtiene: $\frac{11+12}{2} = 11,5$

La diferencia $|D_i| = 9$ está repetida 2 veces, el rango se obtiene: $\frac{14+15}{2} = 14,5$

1	2	3	4	5	6	7
Orden x_i	$D_i = x_i - 40$	$ D_i $	Posición $ D_i $	Rango R	R_i^-	R_i^+
31	-9	9	14	14,5	14,5	
32	-8	8	13	13	13	
33	-7	7	11	11,5	11,5	
34	-6	6	8	9	9	
34	-6	6	9	9	9	
34	-6	6	10	9	9	
35	-5	5	6	6,5	6,5	
37	-3	3	4	4,5	4,5	
39	-1	1	1	1,5	1,5	
41	1	1	2	1,5		1,5
42	2	2	3	3		3
43	3	3	5	4,5		4,5
45	5	5	7	6,5		6,5
47	7	7	12	11,5		11,5
49	9	9	15	14,5		14,5
50	10	10	16	16		16
52	12	12	17	17		17
53	13	13	18	18		18
57	17	17	19	19		19
60	20	20	20	20		20
					78,5	131,5

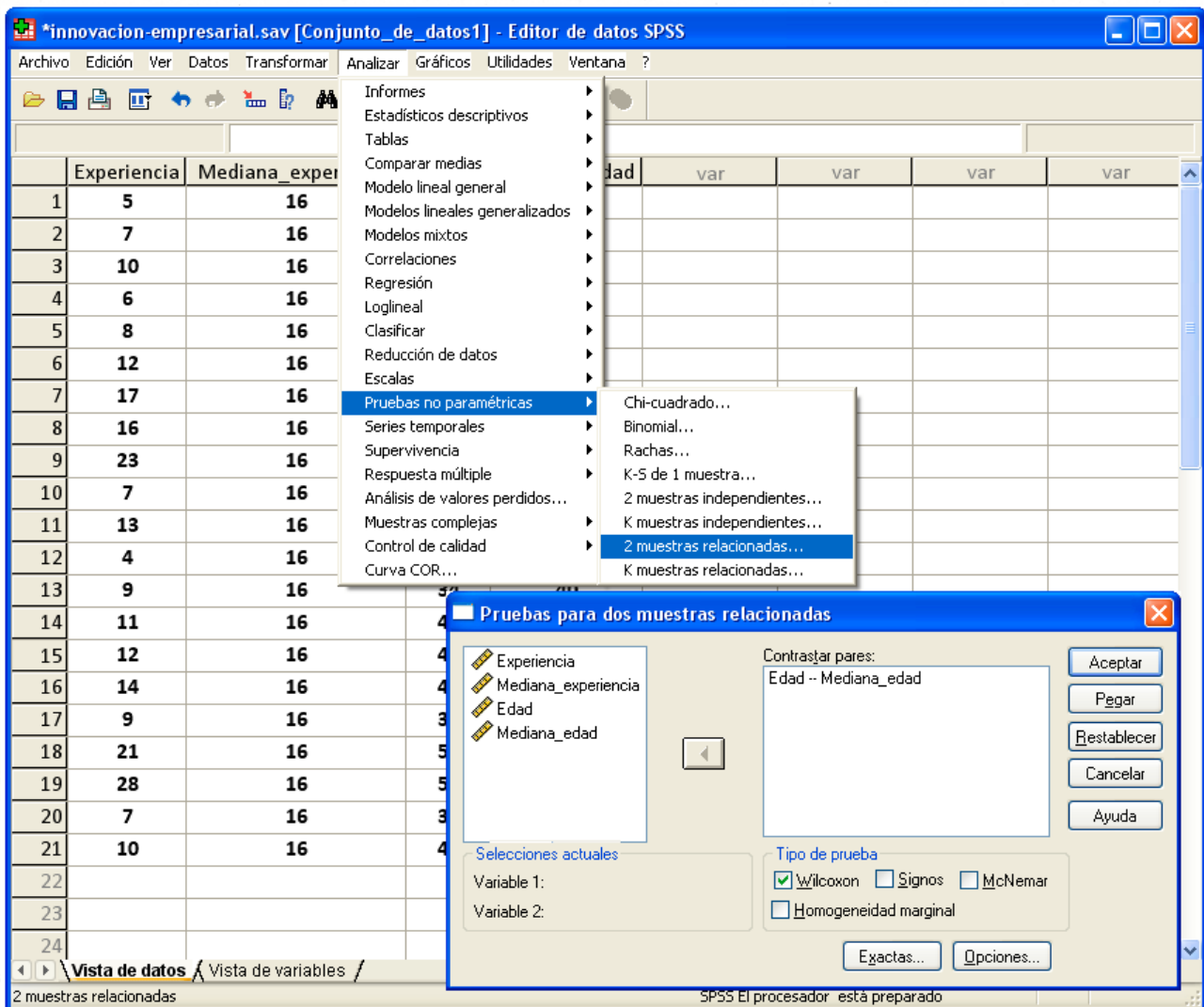
En la sexta y séptima columna aparecen, respectivamente, los rangos negativos y positivos.

$$T_- = \sum R_i^- = 78,5 \quad T_+ = \sum R_i^+ = 131,5$$

$$E[T_+] = \frac{n \cdot (n+1)}{4} = \frac{20 \cdot 21}{4} = 105 \quad \text{Var}[T_+] = \frac{n \cdot (n+1) \cdot (2n+1)}{24} = \frac{20 \cdot 21 \cdot 41}{24} = 717,5$$

$$z_p = \frac{T_+ - E[T_+]}{\sqrt{\text{Var}[T_+]}} = \frac{131,5 - 105}{\sqrt{717,5}} = 0,99$$

$z_p = 0,99 \geq -1,645 = -z_{0,05} \rightarrow$ aceptando la hipótesis nula



*innovacion-empresarial.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

Informes
Estadísticos descriptivos
Tablas
Comparar medias
Modelo lineal general
Modelos lineales generalizados
Modelos mixtos
Correlaciones
Regresión
Loglineal
Clasificar
Reducción de datos
Escalas
Pruebas no paramétricas
Series temporales
Supervivencia
Respuesta múltiple
Análisis de valores perdidos...
Muestras complejas
Control de calidad
Curva COR...

Chi-cuadrado...
Binomial...
Rachas...
K-S de 1 muestra...
2 muestras independientes...
K muestras independientes...
2 muestras relacionadas...
K muestras relacionadas...

Pruebas para dos muestras relacionadas

Selecciones actuales:
Experiencia
Mediana_experiencia
Edad
Mediana_edad

Contrastar pares:
Edad - Mediana_edad

Tipo de prueba:
☒ Wilcoxon ☐ Signos ☐ McNemar
☐ Homogeneidad marginal

Exactas... Opciones...

2 muestras relacionadas

SPSS El procesador está preparado

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

Prueba de los rangos con signo de Wilcoxon

Rangos

	N	Rango promedio	Suma de rangos
Mediana_edad - Edad Rangos negativos	11 ^a	11,95	131,50
Rangos positivos	9 ^b	8,72	78,50
Empates	1 ^c		
Total	21		

a. Mediana_edad < Edad

b. Mediana_edad > Edad

c. Mediana_edad = Edad

Estadísticos de contraste^b

	Mediana_edad - Edad
Z	-,990 ^a
Sig. asintót. (bilateral)	,322

a. Basado en los rangos positivos.

b. Prueba de los rangos con signo de Wilcoxon

SPSS siempre un contraste bilateral o de dos colas: $H_0: M_e = 40$ $H_1: M_e \neq 40$ que se acepta al presentar un p_valor (Sig. asintótica bilateral) = 0,322 > 0,05

En el caso del contraste unilateral a la izquierda: $H_0: M_e \geq 40$ $H_1: M_e < 40$ corresponde una Sig.asintótica izquierda = $1 - p_valor / 2 = 1 - 0,161 = 0,839 > 0,05$, con lo que se acepta la hipótesis nula.

b) Como los datos proceden de una población continua y simétrica se puede emplear la prueba de signos de Wilcoxon, estableciendo el contraste bilateral o de dos colas:

$$H_0: M_e = 16 \quad H_1: M_e \neq 16$$

La octava posición es igual a la mediana propuesta, la diferencia es cero, con lo que se elimina del estudio, y el tamaño muestral es $n = 20 > 15$, utilizando la aproximación asintótica bilateral a la distribución $N(0, 1)$

$$\text{Se admite } H_0 \text{ si } z_p = \frac{|T_+ - E[T_+]|}{\sqrt{\text{Var}[T_+]}} \leq z_{\alpha/2}$$

Para determinar el estadístico de contraste se recurre a la tabla siguiente:

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

1	2	3	4	5	6	7
Orden y_i	$D_i = y_i - 16$	$ D_i $	Posición $ D_i $	Rango R	R_i^-	R_i^+
4	-12	12	19	19,5	19,5	
5	-11	11	18	18	18	
6	-10	10	17	17	17	
7	-9	9	14	15	15	
7	-9	9	15	15	15	
7	-9	9	16	15	15	
8	-8	8	13	13	13	
9	-7	7	10	11	11	
9	-7	7	11	11	11	
10	-6	6	8	8,5	8,5	
10	-6	6	9	8,5	8,5	
11	-5	5	6	6,5	6,5	
12	-4	4	4	4,5	4,5	
12	-4	4	5	4,5	4,5	
13	-3	3	3	3	3	
14	-2	2	2	2	2	
17	1	1	1	1		1
21	5	5	7	6,5		6,5
23	7	7	12	11		11
28	12	12	20	19,5		19,5
					172	38

- En la primera columna se ordenan los datos muestrales de menor a mayor
- La segunda columna son las diferencias entre los datos muestrales ordenados y la mediana
- En la tercera columna las diferencias absolutas $|D_i| = |y_i - 16|$
- En la cuarta columna la posición de $|D_i|$ en orden ascendente, empezando por la más pequeña.

En la quinta columna los rangos de las diferencias $|D_i|$, cuando hay diferencias repetidas se asigna el rango medio de ellas.

Excel: JERARQUIA.MEDIA(Dato, Columna valor absoluto datos, Columna datos ordenados)

La diferencia $|D_i| = 4$ está repetida 2 veces, el rango se obtiene: $\frac{4+5}{2} = 4,5$

La diferencia $|D_i| = 5$ está repetida 2 veces, el rango se obtiene: $\frac{6+7}{2} = 6,5$

La diferencia $|D_i| = 6$ está repetida 2 veces, el rango se obtiene: $\frac{8+9}{2} = 8,5$

La diferencia $|D_i| = 7$ está repetida 3 veces, el rango se obtiene: $\frac{10+11+12}{3} = 11$

La diferencia $|D_i| = 9$ está repetida 3 veces, el rango se obtiene: $\frac{14+15+16}{3} = 15$

La diferencia $|D_i| = 12$ está repetida 2 veces, el rango se obtiene: $\frac{19+20}{2} = 19,5$

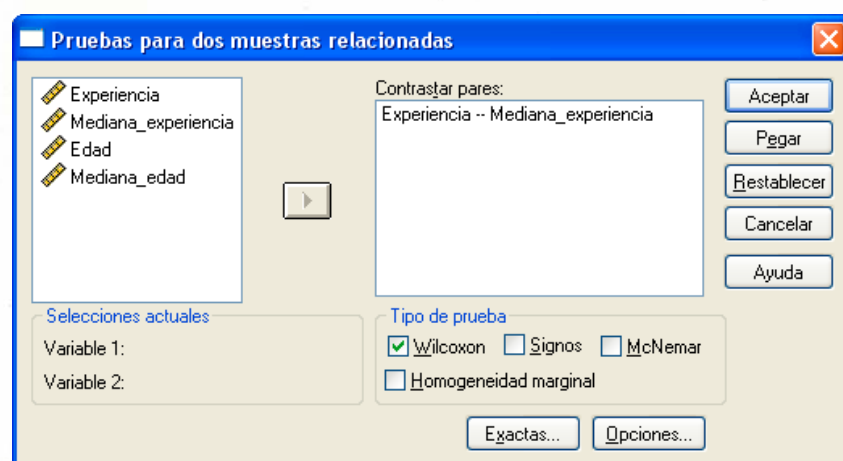
En la sexta y séptima columna aparecen, respectivamente, los rangos negativos y positivos.

$$T_- = \sum R_i^- = 172 \quad T_+ = \sum R_i^+ = 38$$

$$E[T_+] = \frac{n \cdot (n+1)}{4} = \frac{20 \cdot 21}{4} = 105 \quad \text{Var}[T_+] = \frac{n \cdot (n+1) \cdot (2n+1)}{24} = \frac{20 \cdot 21 \cdot 41}{24} = 717,5$$

$$z_p = \frac{|T_+ - E[T_+]|}{\sqrt{\text{Var}[T_+]}} = \frac{|38 - 105|}{\sqrt{717,5}} = 2,504$$

$z_p = 2,504 > 1,645 = z_{0,05} \rightarrow$ Se rechaza la hipótesis nula a un nivel de significación del 1%.



Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

Prueba de los rangos con signo de Wilcoxon**Rangos**

		N	Rango promedio	Suma de rangos
Mediana_experiencia	Rangos negativos	4 ^a	9,50	38,00
- Experiencia	Rangos positivos	16 ^b	10,75	172,00
	Empates	1 ^c		
	Total	21		

a. Mediana_experiencia < Experiencia

b. Mediana_experiencia > Experiencia

c. Mediana_experiencia = Experiencia

Estadísticos de contraste^b

	Mediana_experiencia - Experiencia
Z	-2,504 ^a
Sig. asintót. (bilateral)	,012

a. Basado en los rangos negativos.

b. Prueba de los rangos con signo de Wilcoxon

El p_valor (Sig. asintótica bilateral) = 0,012 < 0,05 por lo que se rechaza la hipótesis nula, es decir, la mediana no es igual a 16 años de experiencia profesional, para la muestra considerada a un nivel de significación del 5%.

178. En la tabla adjunta aparecen dos muestras aleatorias simples que reflejan los costes laborales unitarios de dos sectores de producción A y B::

SECTOR A	SECTOR B
1200	1263
1150	1370
1326	1614
1480	1416
1235	1139
1215	1041
1515	1493
1375	1817
1296	1388
1463	1110
1514	
1409	
1268	
1377	
1091	

Con un nivel de significación del 5%, utilizando la prueba de Ansari-Bradley, se pide:

- Contrastar que los costes unitarios de los dos sectores tienen la misma dispersión, suponiendo que las medianas de las dos poblaciones son iguales.
- Contrastar que los costes unitarios de los dos sectores tienen la misma dispersión, suponiendo que las medianas de las dos poblaciones son distintas.

Solución:

a) Siendo las dos poblaciones continuas se emplea la prueba de Ansari-Bradley

Se establece el contraste bilateral con las hipótesis: $H_0: \sigma_A = \sigma_B$ $H_1: \sigma_A \neq \sigma_B$

Como las poblaciones son continuas se puede emplear la prueba de Ansari-Bradley

Se construye la muestra conjunta de tamaño $n = n_A + n_B = 15 + 10 = 25 > 20$ (impar), por lo que se emplea la aproximación asintótica a una distribución normal $N(0, 1)$.

Se ordenan los valores de menor a mayor, dando al coeficiente instrumental c_i el valor 1 a los valores de A y 0 a los valores de B.

Se asigna a cada valor un coeficiente: $R_i = 1, 2, 3, \dots, \frac{n+1}{2}, \dots, 3, 2, 1$

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

SECTOR A	SECTOR B	SECTORES A - B	POSICION	c_i	R_i	$c_i \cdot R_i$
1200	1263	1200	1041	0	1	0
1150	1370	1150	1091	1	2	2
1326	1614	1326	1110	0	3	0
1480	1416	1480	1139	0	4	0
1235	1139	1235	1150	1	5	5
1215	1041	1215	1200	1	6	6
1515	1493	1515	1215	1	7	7
1375	1817	1375	1235	1	8	8
1296	1388	1296	1263	0	9	0
1463	1110	1463	1268	1	10	10
1514		1514	1296	1	11	11
1409		1409	1326	1	12	12
1268		1268	1370	0	13	0
1377		1377	1375	1	12	12
1091		1091	1377	1	11	11
		1263	1388	0	10	0
		1370	1409	1	9	9
		1614	1416	0	8	0
		1416	1463	1	7	7
		1139	1480	1	6	6
		1041	1493	0	5	0
		1493	1514	1	4	4
		1817	1515	1	3	3
		1388	1614	0	2	0
		1110	1817	0	1	0
						113

$$W = \sum_{i=1}^{25} c_i \cdot R_i = 113$$

$$E(W) = \frac{n_A \cdot (n+1)^2}{4 \cdot n} = \frac{15 \cdot 26^2}{4 \cdot 25} = 101,4$$

$$V(W) = \frac{n_A \cdot n_B \cdot (n+1) \cdot (n^2 + 3)}{48 \cdot n^2} = \frac{15 \cdot 10 \cdot 26 \cdot (25^2 + 3)}{48 \cdot 25^2} = 81,64$$

$$\text{Estadístico de contraste: } z_p = \frac{|W - E[W]|}{\sqrt{\text{Var}[W]}} = \frac{|113 - 101,4|}{\sqrt{81,64}} = 1,284$$

Como $z_p = 1,284 < 1,96 = z_{0,025} \rightarrow$ Se acepta la hipótesis nula de igualdad de dispersiones.

b) Se establece el contraste bilateral con las hipótesis: $H_0: \sigma_A = \sigma_B$ $H_1: \sigma_A \neq \sigma_B$

Como las poblaciones son continuas se puede emplear la prueba de Ansari-Bradley

Al no conocer las medianas de las poblaciones, para calcular el estadístico de contraste han de calcularse las medianas muestrales M_e^A y M_e^B . Para ello, se ordenan los valores muestrales de menor a mayor.

La Mediana del Sector A se obtiene al dividir $n_A / 2 = 15 / 2 = 7,5$, con lo que se encuentra en el octavo valor, es decir, $M_e^A = 1326$ La Mediana del Sector B se obtiene al dividir $n_B / 2 = 10 / 2 = 5$, con lo que se encuentra en el quinto valor, esto es, $M_e^B = 1370$

La Mediana del Sector B se obtiene al dividir $n_B / 2 = 10 / 2 = 5$, con lo que se encuentra en el quinto valor, esto es, $M_e^B = 1370$

Se hacen las diferencias entre los valores muestrales y las correspondientes medianas:

$x_i - M_e^A$ y $y_i - M_e^B$

SECTOR A	SECTOR B	$x_i - M_e^A$	$y_i - M_e^B$
1091	1041	-235	-329
1150	1110	-176	-260
1200	1139	-126	-231
1215	1263	-111	-107
1235	1370	-91	0
1268	1388	-58	18
1296	1416	-30	46
1326	1493	0	123
1375	1614	49	244
1377	1817	51	447
1409		83	
1463		137	
1480		154	
1514		188	
1515		189	

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

Se construye la muestra conjunta de las diferencias de tamaño $n = n_A + n_B$, ordenando las diferencias de menor a mayor y dando al coeficiente instrumental c_i el valor 1 a las diferencias de A y 0 a las diferencias de B.

Después se asigna a cada valor un coeficiente $R_i = 1, 2, 3, \dots, \frac{n+1}{2}, \dots, 3, 2, 1$ atribuyendo el rango o coeficiente medio cuando hay diferencias repetidas.

w_i	c_i	R_i	$c_i \cdot R_i$
-329	0	1	0
-260	0	2	0
-235	1	3	3
-231	0	4	0
-176	1	5	5
-126	1	6	6
-111	1	7	7
-107	0	8	0
-91	1	9	9
-58	1	10	10
-30	1	11	11
0	0	12,5	0
0	1	12,5	12,5
18	0	12	0
46	0	11	0
49	1	10	10
51	1	9	9
83	1	8	8
123	0	7	0
137	1	6	6
154	1	5	5
188	1	4	4
189	1	3	3
244	0	2	0
447	0	1	0
			108,5

El estadístico de contraste W (suma de los coeficientes de la muestra procedente de A) será:

$$W = \sum_{i=1}^{25} c_i \cdot R_i = 108,5$$

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

Señalar que estas diferencias no tienen una distribución exacta en el muestreo, tomando los resultados con cautela.

Como el tamaño de la diferencia de la muestra conjunta $n = n_A + n_B > 20$ se emplea la aproximación asintótica a una distribución normal $N(0, 1)$, considerando una varianza corregida por la presencia de valores repetidos:

$$E(W) = \frac{n_A \cdot (n+1)^2}{4 \cdot n} = \frac{15 \cdot 26^2}{4 \cdot 25} = 101,4$$

$$V_c(W) = \frac{n_A \cdot n_B \cdot \sum_{i=1}^g t_h \cdot R_h^2}{n \cdot (n-1)} - \frac{n_A \cdot n_B \cdot (n+1)^4}{16 \cdot n^2 \cdot (n-1)}$$

$g \equiv$ Número de rangos diferentes cuando aparecen observaciones repetidas

$t_h \equiv$ Tamaño h-ésimo grupo

$R_h \equiv$ Rango medio h-ésimo grupo

Para hallar $\sum_{i=1}^g t_h \cdot R_h^2$ se construye la tabla:

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

t_h	R_h	R_h^2	$t_h \cdot R_h^2$
1	1	1	1
1	2	4	4
1	3	9	9
1	4	16	16
1	5	25	25
1	6	36	36
1	7	49	49
1	8	64	64
1	9	81	81
1	10	100	100
1	11	121	121
2	12,5	156,25	312,5
1	12	144	144
1	11	121	121
1	10	100	100
1	9	81	81
1	8	64	64
1	7	49	49
1	6	36	36
1	5	25	25
1	4	16	16
1	3	9	9
1	2	4	4
1	1	1	1
			1468,5

$$\sum_{i=1}^{24} t_h \cdot R_h^2 = 1468,5$$

$$V_c(W) = \frac{n_A \cdot n_B \cdot \sum_{i=1}^g t_h \cdot R_h^2}{n \cdot (n-1)} - \frac{n_A \cdot n_B \cdot (n+1)^4}{16 \cdot n^2 \cdot (n-1)} = \frac{15 \cdot 10 \cdot 1468,5}{25 \cdot 24} - \frac{15 \cdot 10 \cdot 26^4}{16 \cdot 25^2 \cdot 24} = 81,515$$

$$\text{Estadístico de contraste: } z_p = \frac{|W - E[W]|}{\sqrt{V_c[W]}} = \frac{|108,5 - 101,4|}{\sqrt{81,515}} = 0,786$$

Como $z_p = 0,786 < 1,96 = z_{0,025} \rightarrow$ Se acepta la hipótesis nula de igualdad de dispersiones con un nivel de significación del 5% (probabilidad que se reparte a partes iguales en las dos colas de la distribución)

179. En la tabla adjunta se presentan los gastos diarios medios de los turistas que visitan dos ciudades españolas A y B.

x_{iA}	300	351	326	482	532	227	211	300	326	469	614	129	265	377
x_{iB}	128	300	189	97	223	341	493	63						

Sabiendo que la mediana de la población A es 400 y la de B es 250, con un nivel de significación del 5%, contrasta que los gastos medios de los turistas con destino A tienen una dispersión igual o menor que los del destino B. Utilizar la prueba de Ansari-Bradley.

Solución:

Se utiliza la prueba de Ansari-Bradley al ser dos poblaciones continuas.

Se establece el contraste unilateral a la derecha $H_0: \sigma_A \leq \sigma_B$ $H_1: \sigma_A > \sigma_B$

Se hacen las diferencias entre los valores muestrales y las correspondientes medianas:

poblacionales $x_{iA} - M_e^A = x_{iA} - 400$ y $x_{iB} - M_e^B = x_{iB} - 250$

CIUDAD A	CIUDAD B	$x_{iA} - M_e^A$	$x_{iB} - M_e^B$
300	128	-100	-122
351	300	-49	50
326	189	-74	-61
482	97	82	-153
532	223	132	-27
227	341	-173	91
211	493	-189	243
300	63	-100	-187
326		-74	
469		69	
614		214	
129		-271	
265		-135	
377		-23	

Se construye la muestra conjunta de las diferencias de tamaño $n = n_A + n_B$, ordenando las diferencias de menor a mayor y dando al coeficiente instrumental c_i el valor 1 a las diferencias de A y 0 a las diferencias de B.

Después se asigna a cada valor un coeficiente $R_i = 1, 2, 3, \dots, \frac{n}{2}, \dots, 3, 2, 1$ atribuyendo el rango o coeficiente medio cuando hay diferencias repetidas.

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

w_i	c_i	R_i	$c_i \cdot R_i$
-271	1	1	1
-189	1	2	2
-187	0	3	0
-173	1	4	4
-153	0	5	0
-135	1	6	6
-122	0	7	0
-100	1	8,5	8,5
-100	1	8,5	8,5
-74	1	10,5	10,5
-74	1	10,5	10,5
-61	0	11	0
-49	1	10	10
-27	0	9	0
-23	1	8	8
50	0	7	0
69	1	6	6
82	1	5	5
91	0	4	0
132	1	3	3
214	1	2	2
243	0	1	0
			85

El estadístico de contraste W (suma de los coeficientes de la muestra procedente de A) será:

$$W = \sum_{i=1}^{22} c_i \cdot R_i = 85$$

Como el tamaño de la diferencia de la muestra conjunta $n = n_A + n_B > 20$ se emplea la aproximación asintótica a una distribución normal $N(0, 1)$, considerando una varianza corregida por la presencia de valores repetidos:

$$E(W) = \frac{n_A \cdot (n+2)}{4} = \frac{14 \cdot 24}{4} = 84 \quad V_c(W) = \frac{n_A \cdot n_B \cdot \sum_{h=1}^g t_h \cdot R_h^2}{n \cdot (n-1)} - \frac{n_A \cdot n_B \cdot (n+2)^2}{16 \cdot (n-1)}$$

Para hallar $\sum_{h=1}^g t_h \cdot R_h^2$ se construye la tabla:

$g \equiv$ Número de rangos diferentes cuando aparecen observaciones repetidas

$t_h \equiv$ Tamaño h-ésimo grupo

$R_h \equiv$ Rango medio h-ésimo grupo

t_h	R_h	R_h^2	$t_h \cdot R_h^2$
1	1	1	1
1	2	4	4
1	3	9	9
1	4	16	16
1	5	25	25
1	6	36	36
1	7	49	49
2	8,5	72,25	144,5
2	10,5	110,25	220,5
1	11	121	121
1	10	100	100
1	9	81	81
1	8	64	64
1	7	49	49
1	6	36	36
1	5	25	25
1	4	16	16
1	3	9	9
1	2	4	4
1	1	1	1
			1011

$$\sum_{i=1}^{20} t_h \cdot R_h^2 = 1011$$

$$V_c(W) = \frac{n_A \cdot n_B \cdot \sum_{i=1}^g t_h \cdot R_h^2}{n \cdot (n-1)} - \frac{n_A \cdot n_B \cdot (n+2)^2}{16 \cdot (n-1)} = \frac{14 \cdot 8 \cdot 1011}{22 \cdot 21} - \frac{14 \cdot 8 \cdot 24^2}{16 \cdot 21} = 43,09$$

$$\text{Estadístico de contraste: } z_p = \frac{W - E[W]}{\sqrt{V_c[W]}} = \frac{85 - 84}{\sqrt{53,09}} = 0,137$$

Como $z_p = 0,137 < 1,645 = z_{0,05} \rightarrow$ Se acepta la hipótesis nula, concluyendo que la dispersión del gasto medio en la ciudad A es menor o igual que el gasto medio de la ciudad B, con un nivel de significación del 5%

180. En tres estaciones del metro de Madrid se analiza la puntualidad del servicio. Para ello, se toma una muestra en cada una de ellas, midiendo el tiempo que se retrasa cada tren que llega. Se han obtenido los resultados siguientes:

A	30	25	29	32	33	20	19	5	18	23		
B	26	21	22	37	38	13	10	9	3	14	16	34
C	31	35	42	15	17	11	12	24				

Contrastar si los retrasos en las tres estaciones son semejantes, con un nivel de significación del 5%. Utilizar la prueba de Kruskal-Wallis.

Solución:

La prueba de Kruskal-Wallis permite contrastar si varias muestras independientes, de iguales o diferentes tamaños, proceden de la misma población continua.

Sean F_1 , F_2 y F_3 las correspondientes funciones de distribución. Se trata de comprobar si estas funciones presentan diferencias significativas en cuanto a su retraso.

Aplicando el test de Kruskal-Wallis se contrasta:

$$H_0: F_1(x) = F_2(x) = F_3(x) \quad H_1: \text{Al menos dos son diferentes}$$

siendo F_i la función de distribución de la variable X_i , $i = 1, 2, 3$

$$\text{Estadístico de contraste: } H = \frac{12}{n \cdot (n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3 \cdot (n+1)$$

$$\text{Siendo, } n = \sum_{i=1}^k n_i \quad R_i = \sum_{j=1}^{n_i} r_{ij} \quad r_{ij} \equiv \text{Rango de la observación } j\text{-ésima de la muestra } i$$

$$\text{Se rechaza } H_0 \text{ cuando } P(H \geq h_\alpha | H_0) = \alpha$$

$h_\alpha \equiv$ valor crítico que se observa en las tabla de valores críticos de Kruskal-Wallis.

Cuando todos los tamaños muestrales son mayores que 5 se recurre a la aproximación asintótica, que emplea la distribución $\chi^2_{\alpha, k-1}$ con $k \equiv$ número de muestras

Para calcular el valor experimental H se ordenan las observaciones x_i de menor a mayor, asignando a cada una su rango R_i

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

estación	x_i	R_i
2	3	1
1	5	2
2	9	3
2	10	4
3	11	5
3	12	6
2	13	7
2	14	8
3	15	9
2	16	10
3	17	11
1	18	12
1	19	13
1	20	14
2	21	15
2	22	16
1	23	17
3	24	18
1	25	19
2	26	20
1	29	21
1	30	22
3	31	23
1	32	24
1	33	25
2	34	26
3	35	27
2	37	28
2	38	29
3	42	30

Se agrupan los rangos de cada estación y se hace la suma de cada una de las muestras:

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

estación	R_i^A	estación	R_i^B	estación	R_i^C
1	2	2	1	3	5
1	12	2	3	3	6
1	13	2	4	3	9
1	14	2	7	3	11
1	17	2	8	3	18
1	19	2	10	3	23
1	21	2	15	3	27
1	22	2	16	3	30
1	24	2	20		
1	25	2	26		
		2	28		
		2	29		
	169		167		129

$$R_i^A = 169$$

$$R_i^B = 167$$

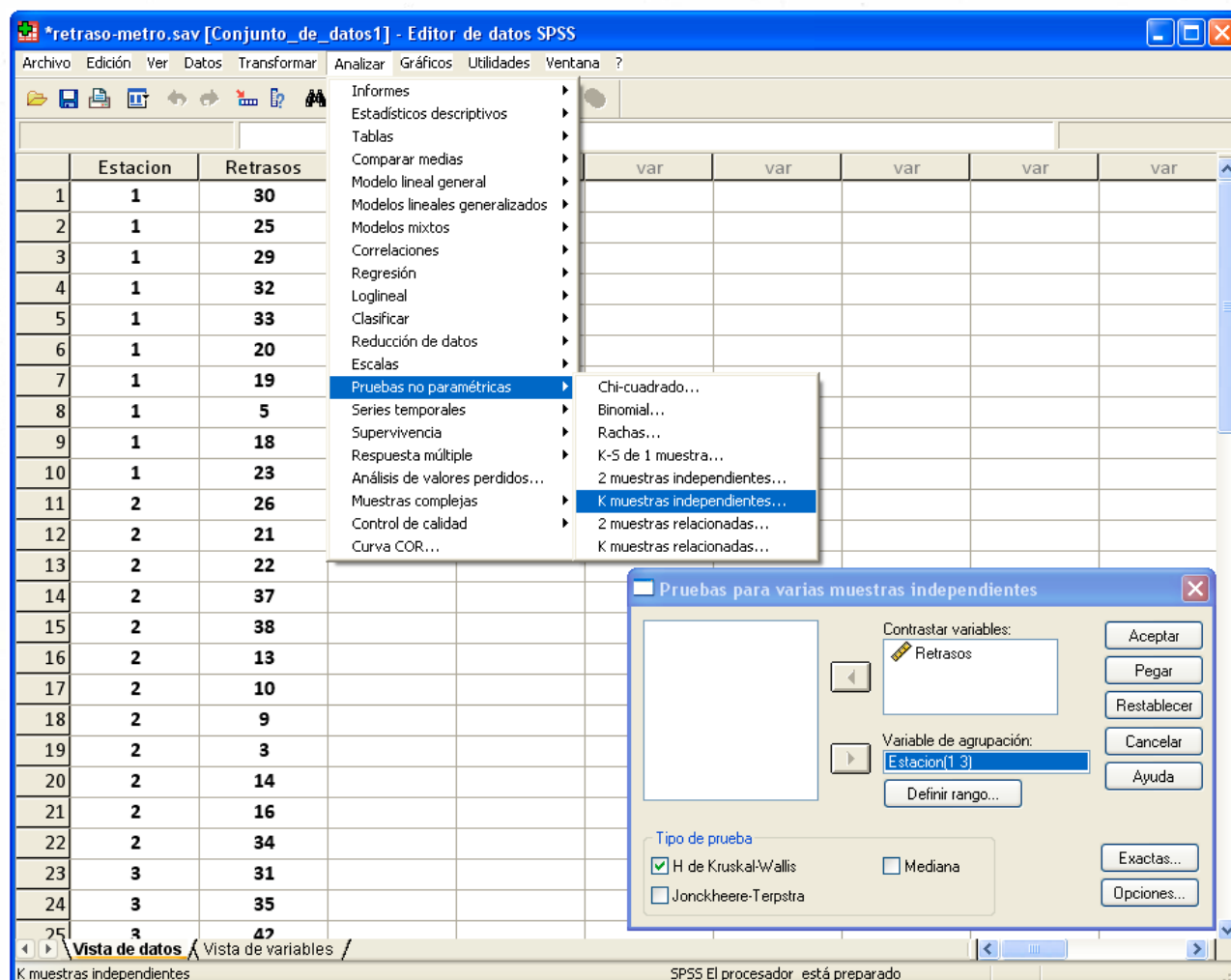
$$R_i^C = 129$$

Estadístico contraste:

$$H = \frac{12}{n \cdot (n+1)} \sum_{i=1}^3 \frac{R_i^2}{n_i} - 3 \cdot (n+1) = \frac{12}{30 \cdot 31} \left(\frac{169^2}{10} + \frac{167^2}{12} + \frac{129^2}{8} \right) - 3 \cdot 31 = 0,6814$$

Como se rechaza H_0 cuando $P(H \geq h_\alpha | H_0) = \alpha$ recurriendo a la aproximación asintótica a la $\chi_{k-1, \alpha}^2 = \chi_{2, 0,05}^2 = 5,991$

$H = 0,6814 < h_{0,05} = 5,991 \rightarrow$ aceptando la hipótesis nula, afirmando que los retrasos en las tres estaciones son semejantes o siguen una misma distribución, con un nivel de significación del 5%.



*retraso-metro.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

Informes
Estadísticos descriptivos
Tablas
Comparar medias
Modelo lineal general
Modelos lineales generalizados
Modelos mixtos
Correlaciones
Regresión
Loglineal
Clasificar
Reducción de datos
Escalas
Pruebas no paramétricas
Series temporales
Supervivencia
Respuesta múltiple
Análisis de valores perdidos...
Muestras complejas
Control de calidad
Curva COR...

Chi-cuadrado...
Binomial...
Rachas...
K-S de 1 muestra...
2 muestras independientes...
K muestras independientes...
2 muestras relacionadas...
K muestras relacionadas...

Pruebas para varias muestras independientes

Contrastar variables:
Retrasos

Variable de agrupación:
Estacion(1 3)

Definir rango...

Tipo de prueba
☒ H de Kruskal-Wallis
☐ Jonckheere-Terpstra
☐ Mediana

Aceptar
Pegar
Restablecer
Cancelar
Ayuda
Exactas...
Opciones...

Vista de datos / Vista de variables /
K muestras independientes

SPSS El procesador está preparado

Prueba de Kruskal-Wallis

Rangos

	Estacio	N	Rango promedio
Retrasos	1	10	16,90
	2	12	13,92
	3	8	16,13
	Total	30	

Estadísticos de contraste^{a,b}

	Retrasos
Chi-cuadrado	,681
gl	2
Sig. asintót.	,711

a. Prueba de Kruskal-Wallis

b. Variable de agrupación: Estacion

El p_valor (Sig. asintótica) = 0,711 > 0,05 = α → aceptando la hipótesis nula, es decir, los retrasos en las tres estaciones son semejantes o siguen una misma distribución, con un nivel de significación del 5%.

181. Un fabricante de juguetes desea conocer si existen diferencias en cuanto a la calidad de cuatro marcas de pilas extendidas en el mercado, con el fin de recomendarlas para su utilización en un nuevo juguete que se va a promocionar. Para comparar las cuatro marcas, toma muestras aleatorias de pilas de cada una de las marcas y controla el tiempo que permanecen funcionando el juguete en cuestión. Los resultados obtenidos fueron:

A	87	95	100	79	85	98	100	84	
B	97	62	59	37	90	42	44	100	94
C	100	45	50	95	70	78	80	47	69
D	75	50	54	65					

Utilizando un nivel de significación del 5%, ¿puede decirse que existen diferencias significativas en las calidades de estas marcas de pilas alcalinas?.

Utilizar la prueba de Kruskal-Wallis.

Solución:

Sean F_1, F_2, F_3 y F_4 las correspondientes funciones de distribución. Se trata de comprobar si estas funciones presentan diferencias significativas en cuanto a la duración de las pilas.

Aplicando el test de Kruskal-Wallis se contrasta:

$$H_0: F_1(x) = F_2(x) = F_3(x) \quad H_1: \text{Al menos dos son diferentes}$$

siendo F_i la función de distribución de la variable X_i , $i = 1, 2, 3, 4$

El estadístico de contraste corregido por observaciones repetidas:

$$H^* = \frac{1}{f_c} \cdot H = \frac{1}{f_c} \cdot \frac{12}{n \cdot (n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3 \cdot (n+1)$$

$$f_c = 1 - \frac{\sum_j \tau_j^3 - \tau_j}{n^2 \cdot (n-1)}$$

Siendo, $n = \sum_{i=1}^k n_i$ $R_i = \sum_{j=1}^{n_i} r_{ij}$ $r_{ij} \equiv$ Rango de la observación j -ésima de la muestra i

$\tau_j \equiv$ Número de observaciones que se repiten de cada grupo de repeticiones.

Se rechaza H_0 cuando $P(H^* \geq h_\alpha | H_0) = \alpha$

Para calcular el valor experimental se ordenan las observaciones de menor a mayor, asignándoles su correspondiente rango y sumando los rangos de las observaciones de cada muestra. En caso de empates, se procede de forma habitual:

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

Muestra pilas	x_i	Muestra pilas	x_i	R_i
1	87	2	37	1
1	95	2	42	2
1	100	2	44	3
1	79	3	45	4
1	85	3	47	5
1	98	4	50	6,5
1	100	3	50	6,5
1	84	4	54	8
2	97	2	59	9
2	62	2	62	10
2	59	4	65	11
2	37	3	69	12
2	90	3	70	13
2	42	4	75	14
2	44	3	78	15
2	100	1	79	16
2	94	3	80	17
3	100	1	84	18
3	45	1	85	19
3	50	1	87	20
3	95	2	90	21
3	70	2	94	22
3	78	3	95	23,5
3	80	1	95	23,5
3	47	2	97	25
3	69	1	98	26
4	75	1	100	28,5
4	50	2	100	28,5
4	54	3	100	28,5
4	65	1	100	28,5

Se observa que hay observaciones repetidas, como en el caso del valor 100 que aparece en las posiciones 27, 28, 29 y 30, siendo el rango medio:

$$\bar{R} = \frac{27 + 28 + 29 + 30}{4} = 28,5$$

Muestra A	R_i^A	Muestra B	R_i^B	Muestra C	R_i^C	Muestra D	R_i^D
1	16	2	1	3	4	4	6,5
1	18	2	2	3	5	4	8
1	19	2	3	3	6,5	4	11
1	20	2	9	3	12	4	14
1	23,5	2	10	3	13		
1	26	2	21	3	15		
1	28,5	2	22	3	17		
1	28,5	2	25	3	23,5		
		2	28,5	3	28,5		
	179,5		121,5		124,5		39,5

$$R_i^A = 179,5 \quad R_i^B = 121,5 \quad R_i^C = 124,5 \quad R_i^D = 39,5$$

Estadístico de contraste inicial:

$$H = \frac{12}{n \cdot (n+1)} \sum_{i=1}^4 \frac{R_i^2}{n_i} - 3 \cdot (n+1) = \frac{12}{30 \cdot 31} \left(\frac{179,5^2}{8} + \frac{121,5^2}{9} + \frac{124,5^2}{9} + \frac{39,5^2}{4} \right) - 3 \cdot 31 = 7,3883$$

Empleando el factor de corrección para observaciones repetidas:

$$f_c = 1 - \frac{\sum_j (\tau_j^3 - \tau_j)}{n^2 \cdot (n-1)}$$

$\tau_j \equiv$ Número de observaciones que se repiten de cada grupo de repeticiones: 2 (por el valor de 50), 2 (por el valor de 95) y 4 (por el valor de 100)

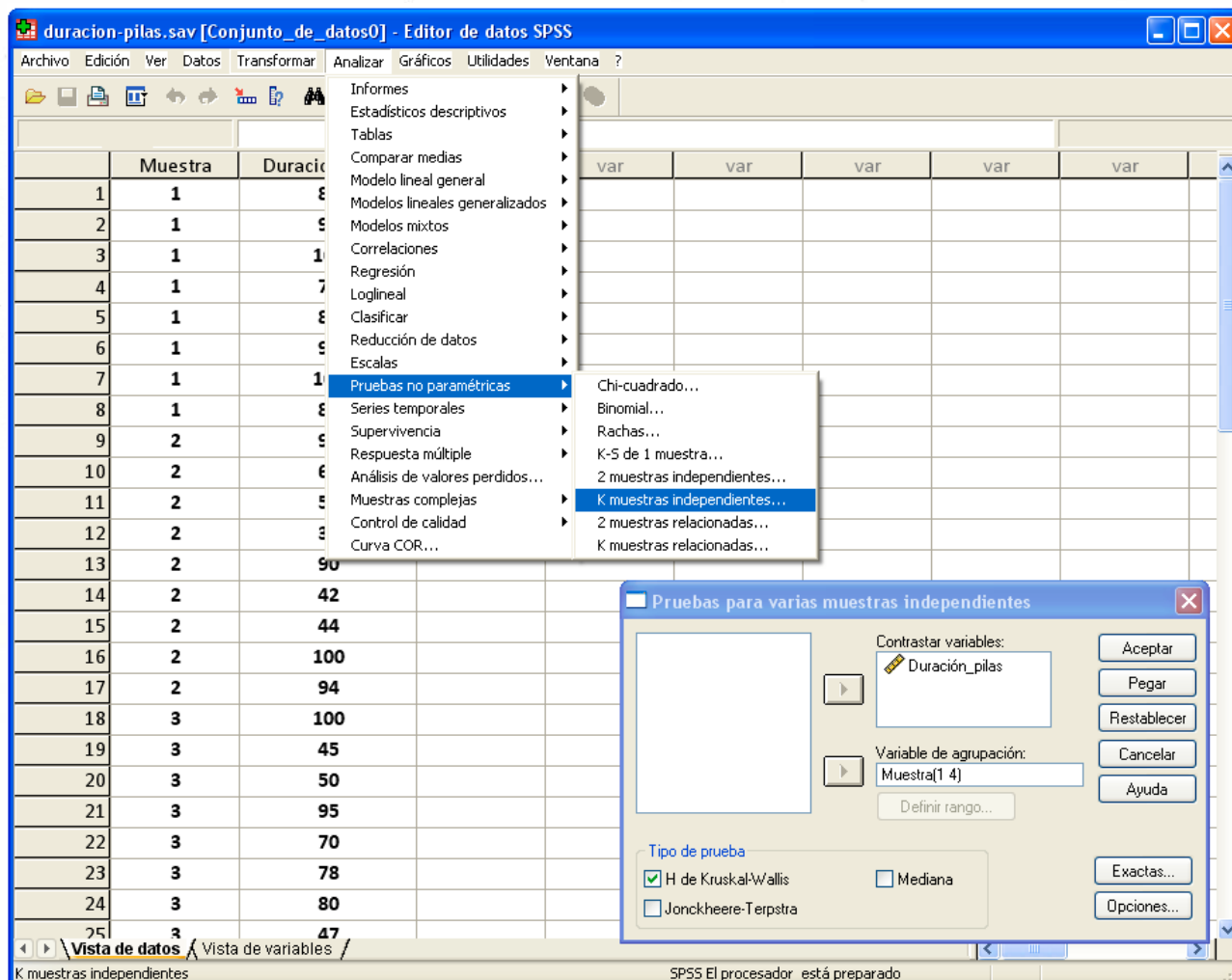
$$f_c = 1 - \frac{\sum_j (\tau_j^3 - \tau_j)}{n^2 \cdot (n-1)} = 1 - \frac{(2^3 - 2) + (2^3 - 2) + (4^3 - 4)}{30^2 \cdot 29} = 0,9972$$

El estadístico de contraste final: $H^* = \frac{1}{f_c} \cdot H = \frac{7,3883}{0,9972} = 7,408$

Se rechaza H_0 cuando $P(H^* \geq h_\alpha | H_0) = \alpha$

recurriendo a la aproximación asintótica a la $\chi_{k-1, \alpha}^2 = \chi_{3, 0,05}^2 = 7,815$

$H^* = 7,408 < h_{0,05} = 7,815 \rightarrow$ Se acepta la hipótesis nula, por tanto no existen diferencias significativas en la duración de las pilas alcalinas, con un nivel de significación del 5%.



The screenshot shows the SPSS 'Editor de datos' window for 'duracion-pilas.sav'. The 'Analizar' menu is open, showing 'Pruebas no paramétricas' selected. The 'Pruebas para varias muestras independientes' dialog box is open, showing 'Duración_pilas' as the variable to contrast and 'Muestra(1 4)' as the grouping variable. The 'H de Kruskal-Wallis' test is selected under 'Tipo de prueba'.

Prueba de Kruskal-Wallis

Rangos

	Muestra	N	Rango promedio
Duración_pilas	Muestra A	8	22,44
	Muestra B	9	13,50
	Muestra C	9	13,83
	Muestra D	4	9,88
	Total	30	

Estadísticos de contraste^{a,b}

	Duración_pilas
Chi-cuadrado	7,408
gl	3
Sig. asintót.	,060

a. Prueba de Kruskal-Wallis

b. Variable de agrupación: Muestra

Siendo el p_{valor} (Sig. asintótica) = $0,06 > 0,05 = \alpha \rightarrow$ Se acepta la hipótesis nula, concluyendo que no existen diferencias significativas en la duración de las pilas alcalinas, es decir, la duración de las pilas pueden tener una misma distribución, con un nivel de significación del 5%

182. En la siguiente tabla se adjuntan los datos sobre el ROE (ratio utilizado para realizar análisis mediante la medición de la rentabilidad obtenida por la entidad sobre sus recursos propio) y ROA (beneficio obtenido antes de intereses e impuestos - activos totales) de diez entidades financieras que cotizan en Bolsa.

Entidades	ROE	ROA
1 Endesa	13,68	1,80
2 Repsol	14,64	1,73
3 BBVA	20,28	1,66
4 Banesto	17,52	1,25
5 Carrefour	20,34	2,21
6 Ibercaja	15,12	2,28
7 Aceralia	18,6	1,75
8 Unicaja	18,72	2,83
9 BBK	19,32	2,62
10 Bankia	13,8	1,32

Contrastar que ambos conceptos no están relacionados empleando la τ de Kendall, con un nivel de significación del 5%.

Solución:

Se establece el contraste bilateral: $H_0: \tau_{xy} = 0$ $H_1: \tau_{xy} \neq 0$

$$\text{Estadístico de contraste: } \tau = \frac{|2 \cdot S|}{n \cdot (n-1)}$$

con $n \equiv$ tamaño de la muestra y $S \equiv$ Suma total de la función indicador: $S = \sum_{\substack{i=1 \\ i < j}}^n \psi_{ij}$

Se hallan los indicadores: $\psi_{ij}(R_{y_i}^*, R_{y_j}^*) = \begin{cases} 1 & \text{Si no hay inversión de rangos} \\ -1 & \text{Si hay inversión de rangos} \end{cases}$

utilizando la doble tabla de $R_{y_i}^*$

Se rechaza H_0 si $P[\tau \geq k | H_0 \text{ es Cierta}]$

El estadístico de contraste τ se contruye mediante una tabla que contenga los rangos R_{x_i} y R_{y_i} , considerando los valores ordenados de mayor a menor.

Posteriormente se reordenan siguiendo el orden natural de R_{x_i}

$$R_{x_i} \equiv \text{Jerarquia}(x_i; \$x\$1 : \$x\$10)$$

$$R_{y_i} \equiv \text{Jerarquia}(y_i; \$y\$1 : \$y\$10)$$

Entidades	x_i	y_i	R_{x_i}	R_{y_i}
Endesa	13,68	1,80	10	5
Repsol	14,64	1,73	8	7
BBVA	20,28	1,66	2	8
Banesto	17,52	1,25	6	10
Carrefour	20,34	2,21	1	4
Ibercaja	15,12	2,28	7	3
Aceralia	18,6	1,75	5	6
Unicaja	18,72	2,83	4	1
BBK	19,32	2,62	3	2
Bankia	13,8	1,32	9	9

Entidades	$R_{x_i}^*$	$R_{y_i}^*$
Carrefour	1	4
BBVA	2	8
BBK	3	2
Unicaja	4	1
Aceralia	5	6
Banesto	6	10
Ibercaja	7	3
Repsol	8	7
Bankia	9	9
Endesa	10	5

Se hallan los indicadores: $\psi_{ij}(R_{y_i}^*, R_{y_j}^*) = \begin{cases} 1 & \text{Si no hay inversión de rangos} \\ -1 & \text{Si hay inversión de rangos} \end{cases}$

utilizando la doble tabla de $R_{y_i}^*$

		$R_{y_2}^*$	$R_{y_3}^*$	$R_{y_4}^*$	$R_{y_5}^*$	$R_{y_6}^*$	$R_{y_7}^*$	$R_{y_8}^*$	$R_{y_9}^*$	$R_{y_{10}}^*$
		8	2	1	6	10	3	7	9	5
$R_{y_1}^*$	4	1	-1	-1	1	1	-1	1	1	1
$R_{y_2}^*$	8		-1	-1	-1	1	-1	-1	1	-1
$R_{y_3}^*$	2			-1	1	1	1	1	1	1
$R_{y_4}^*$	1				1	1	1	1	1	1
$R_{y_5}^*$	6					1	-1	1	1	-1
$R_{y_6}^*$	10						-1	-1	-1	-1
$R_{y_7}^*$	3							1	1	1
$R_{y_8}^*$	7								1	-1
$R_{y_9}^*$	9									-1
		1	-2	-3	2	5	-2	3	6	-1

$$S = \sum_{\substack{i=1 \\ i < j}}^9 \psi_{ij} = 1 - 2 - 3 + 2 + 5 - 2 + 3 + 6 - 1 = 9$$

$$\tau = \frac{|2 \cdot S|}{n \cdot (n-1)} = \frac{2 \cdot 9}{10 \cdot 9} = 0,200$$

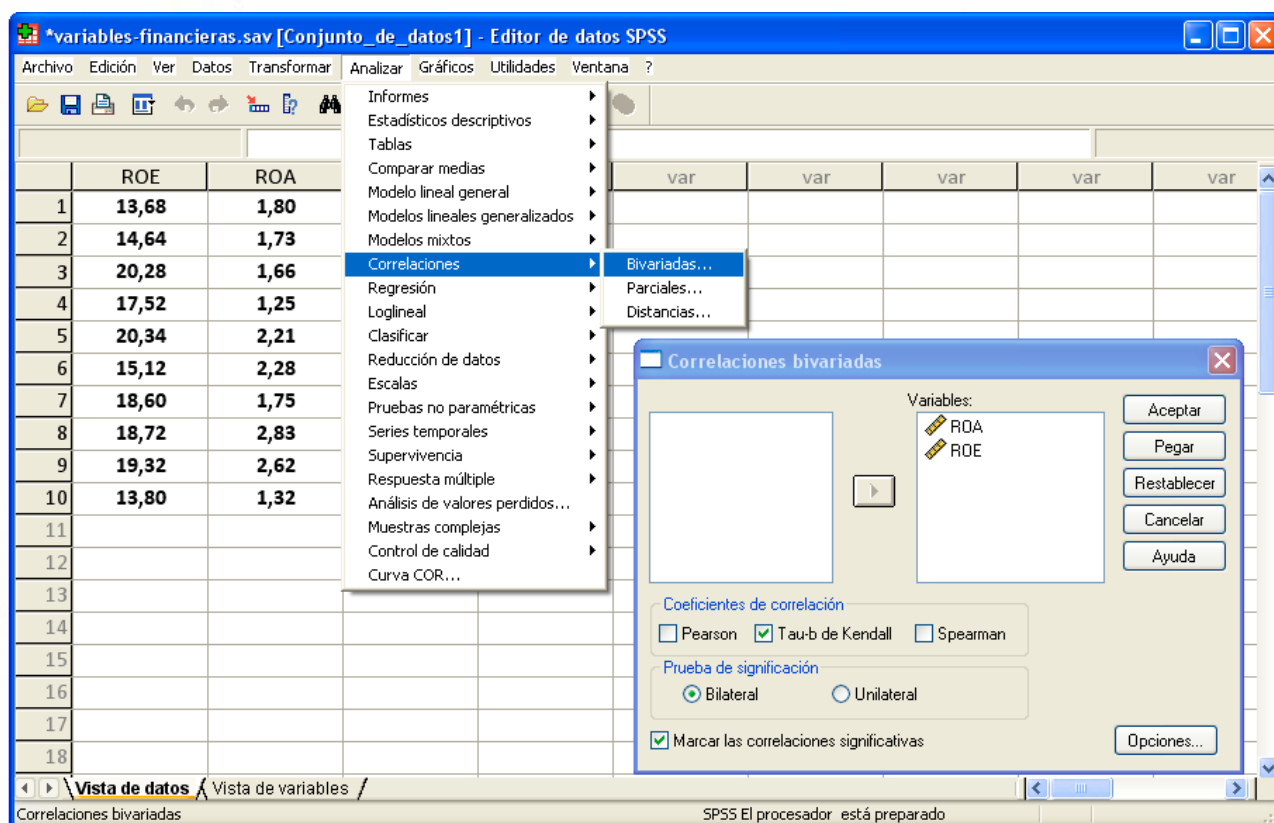
Teniendo en cuenta la aproximación asintótica

$$\tau \approx N\left(0, \sqrt{\frac{2 \cdot (2n+5)}{9n \cdot (n-1)}}\right) = N\left(0, \sqrt{\frac{2 \cdot 25}{9 \cdot 10 \cdot 9}}\right) = N(0, 0,248)$$

$$\text{Se rechaza } H_0 \text{ si } P[\tau \geq k \mid H_0 \text{ es Cierta}] = P[\tau \geq k \mid N(0, 0,248)] = P\left[z \geq \frac{k}{0,248}\right]$$

$$\frac{k}{0,248} = 1,96 = z_{0,025} \Rightarrow k = 0,486$$

Como se verifica que $\tau = 0,200 < 0,486 = z_{0,025} \rightarrow$ Se acepta la hipótesis nula de que las variables se encuentran incorreladas, es decir, las variables ROE y ROA no se encuentran asociadas.



Correlaciones

			ROA	ROE
Tau_b de Kendall	ROA	Coefficiente de correlación	1,000	,200
		Sig. (bilateral)	.	,421
		N	10	10
	ROE	Coefficiente de correlación	,200	1,000
		Sig. (bilateral)	,421	.
		N	10	10

El p_{valor} (Signatura bilateral) = $0,421 > 0,05 = \alpha \rightarrow$ Se acepta la hipótesis nula de incorrelación de las variables ROA Y ROE. En consecuencia, no se encuentran asociadas.

183. La tabla adjunta presenta las cotizaciones máximas de las acciones de una compañía durante los últimos diez días del mes de octubre:

14,47 12,63 15,12 17,75 18,05 17,1 13,42 20,4 19,3 17,4

a) Con un nivel de significación del 10%, ¿se puede decir que la muestra es aleatoria con respecto a la media de cotizaciones alcanzada en mes anterior que fue de 17,5 euros?

b) Contrastar si las acciones se pueden considerar normalmente distribuidas, con un nivel de significación del 5%. Utilizar el contraste de Lilliefors y Shapiro-Wilk

Solución:

a) Se utiliza el contraste de rachas de Wald-Wolfowitz

Se establece el contraste: H_0 : La muestra es aleatoria H_1 : La muestra no es aleatoria

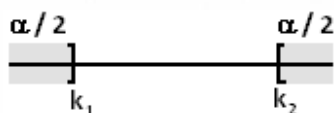
Estadístico contraste: R = Número de rachas

Para obtener una sucesión dicotómica se compara cada observación con 17,5, sustituyendo el valor por el signo S cuando el valor observado supere a esta cantidad o por el signo I cuando la observación sea inferior a 17,5. En caso de haber alguna observación igual a esta cantidad se ignora y el tamaño de la muestra se reduce convenientemente.

14,47 12,63 15,12 17,75 18,05 17,1 13,42 20,4 19,3 17,4
I I I S S I I S S I

La sucesión dicotómica queda: III | SS | II | SS | I $\rightarrow \hat{R} = 5$

n_1 = Número de signos S = 4 n_2 = Número de signos I = 6

Región crítica del contraste:  Siendo k_1 y k_2 el menor y el mayor entero, respectivamente, que verifican:

$$P(R \leq k_1) \leq \frac{\alpha}{2} = 0,05 \quad P(R \geq k_2) \leq \frac{\alpha}{2} = 0,05$$

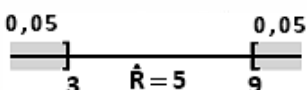
Utilizando la tabla para el test de rachas de aleatoriedad:

Distribución de probabilidades para el test de rachas de aleatoriedad
distribución del número total de rachas R ; $P(R \leq r)$ en una muestra
 $n = n_1 + n_2$, para el test de rachas de aleatoriedad de Wald-Wolfowitz:

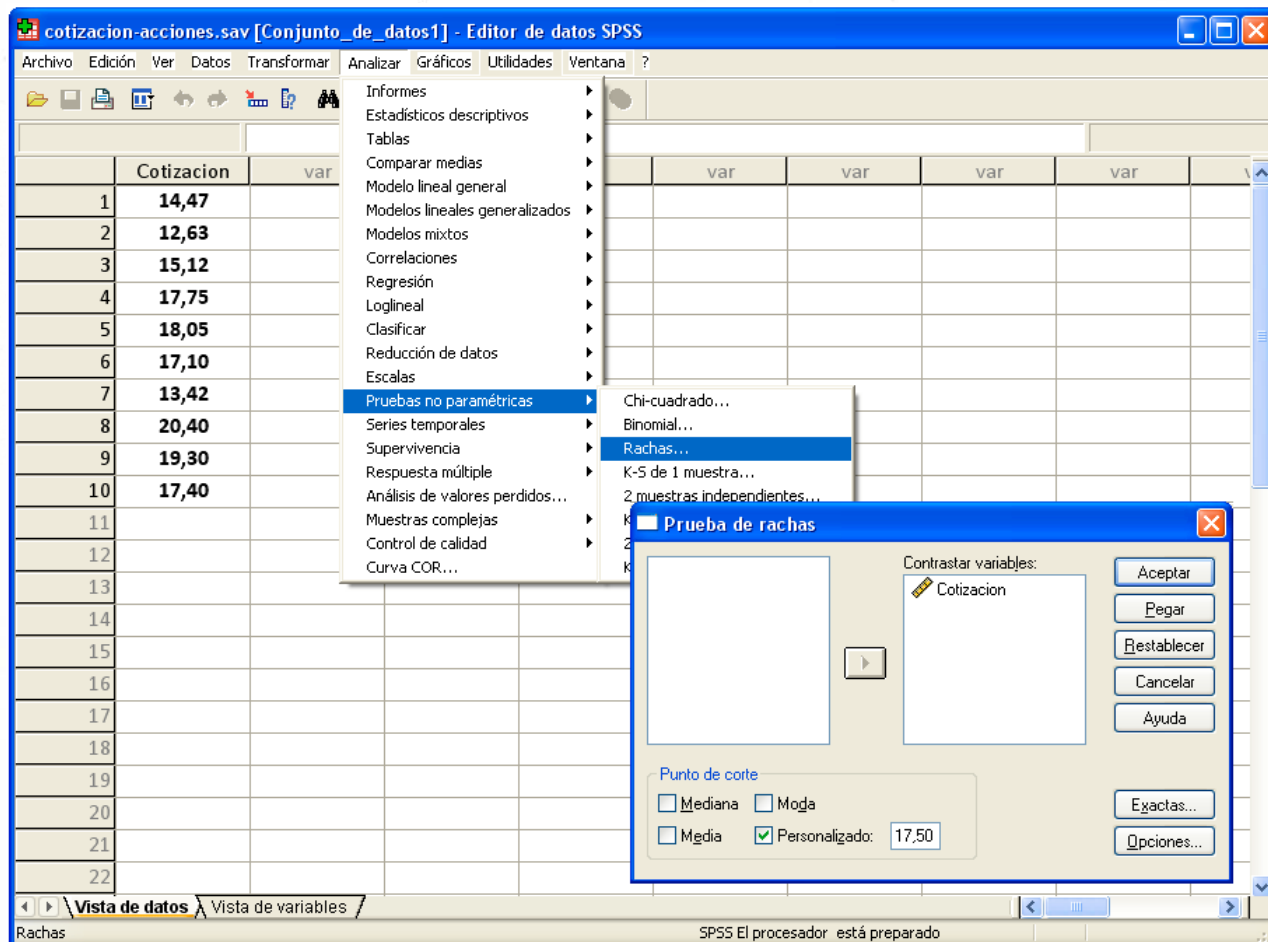
n_1	n_2	r										
		2	3	4	5	6	7	8	9	10	11	12
2	3	0,2000	0,5000	0,9000	1,0000							
2	4	0,1333	0,4000	0,8000	1,0000							
2	5	0,0952	0,3333	0,7143	1,0000							
2	6	0,0714	0,2857	0,6429	1,0000							
2	7	0,0556	0,2500	0,5833	1,0000							
2	8	0,0444	0,2222	0,5333	1,0000							
2	9	0,0364	0,2000	0,4909	1,0000							
2	10	0,0303	0,1818	0,4545	1,0000							
3	3	0,1000	0,3000	0,7000	0,9000	1,0000						
3	4	0,0571	0,2000	0,5429	0,8000	0,9714	1,0000					
3	5	0,0357	0,1429	0,4286	0,7143	0,9286	1,0000					
3	6	0,0238	0,1071	0,3452	0,6429	0,8810	1,0000					
3	7	0,0167	0,0833	0,2833	0,5833	0,8333	1,0000					
3	8	0,0121	0,0667	0,2364	0,5333	0,7879	1,0000					
3	9	0,0091	0,0545	0,2000	0,4909	0,7454	1,0000					
3	10	0,0070	0,0454	0,1713	0,4545	0,7063	1,0000					
4	4	0,0286	0,1143	0,3714	0,6286	0,8857	0,9714	1,0000				
4	5	0,0159	0,0714	0,2619	0,5000	0,7857	0,9286	0,9921	1,0000			
4	6	0,0095	0,0476	0,1905	0,4048	0,6905	0,8810	0,9762	1,0000			
4	7	0,0061	0,0333	0,1424	0,3333	0,6061	0,8333	0,9545	1,0000			
4	8	0,0040	0,0242	0,1091	0,2788	0,5333	0,7879	0,9293	1,0000			
4	9	0,0028	0,0182	0,0853	0,2364	0,4713	0,7454	0,9021	1,0000			
4	10	0,0020	0,0140	0,0679	0,2028	0,4186	0,7063	0,8741	1,0000			

$$P(R \leq 3) = 0,0476 \leq 0,05 \quad P(R \leq 4) = 0,1905 > 0,05 \quad \rightarrow \quad k_1 = 3$$

$$\begin{cases} P(R \leq 7) = 0,8810 \rightarrow P(R \geq 8) = 1 - 0,8810 = 0,1190 > 0,05 \\ P(R \leq 8) = 0,9762 \rightarrow P(R \geq 9) = 1 - 0,9762 = 0,0238 \leq 0,05 \end{cases} \rightarrow k_2 = 9$$

Región crítica del contraste: 

Con una significación del 5% se acepta la aleatoriedad de la muestra de cotizaciones respecto a la media histórica alcanzada el mes anterior.



Prueba de rachas

	Cotizacion
Valor de prueba ^a	17,500
Casos en total	10
Número de rachas	5
Z	-,211
Sig. asintót. (bilateral)	,833

a. Especificado por el usuario.

El p_{valor} (Sig. sintót. Bilateral) = 0,833 > 0,10 = α
con lo que se acepta la hipótesis nula de aleatoriedad
de la muestra.

b) Contraste de normalidad de Lilliefors

▪ Contraste bilateral:

H_0 : La muestra procede de una distribución normal

H_1 : La muestra no procede de una distribución normal

Estadístico contraste: $D_n = \max |F_n(z) - F_0(z)| = \max(a_i, b_i)$

Se acepta H_0 si $D_n = \max |F_n(z) - F_0(z)| = \max(a_i, b_i) < D_{\alpha, n}$ (valor crítico tabla de Lilliefors)

Para determinar el estadístico de contraste se construye una tabla donde se ordenan los datos, tipificando cada valor, se calcula $F_0(z_i)$, $F_n(z_i)$ y se crean las columnas $a_i = |F_0(z_i) - F_n(z_i)|$ y

$$b_i = |F_0(z_i) - F_n(z_{i-1})|$$

Valores tipificados $z_i = \frac{x_i - \bar{x}}{s_x} = \frac{x_i - 16,564}{2,5524}$

$$a_1 = \bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = \frac{165,64}{10} = 16,564 \quad a_2 = \frac{1}{10} \sum_{i=1}^{10} x_i^2 = \frac{2802,2936}{10} = 280,22936$$

$$\sigma_x^2 = a_2 - a_1^2 = 280,22936 - 16,564^2 = 5,8633$$

$$s_x^2 = \frac{10 \cdot 5,8633}{9} = 6,5147 \quad s_x = \sqrt{6,5147} = 2,5524$$

$F_0(z_i) \equiv$ Función de distribución de una $N(0, 1)$: DISTR.NORM.ESTAND(z_i)

$F_n(z_i) \equiv$ Función de distribución empírica de la muestra tipificada: $F_n(z_i) = \frac{n^\circ \text{ observaciones}}{n} = \frac{N(z)}{n}$

x_i	z_i	$F_n(z_i)$	$F_0(z_i)$	$a_i = F_0(z_i) - F_n(z_i) $	$b_i = F_0(z_i) - F_n(z_{i-1}) $
12,63	-1,5413	0,100	0,0616	0,0384	0,0616
13,42	-1,2318	0,200	0,1090	0,0910	0,0090
14,47	-0,8204	0,300	0,2060	0,0940	0,0060
15,12	-0,5657	0,400	0,2858	0,1142	0,0142
17,1	0,2100	0,500	0,5832	0,0832	0,1832
17,4	0,3275	0,600	0,6284	0,0284	0,1284
17,75	0,4647	0,700	0,6789	0,0211	0,0789
18,05	0,5822	0,800	0,7198	0,0802	0,0198
19,3	1,0719	0,900	0,8581	0,0419	0,0581
20,4	1,5029	1,000	0,9336	0,0664	0,0336

$$D_n = \max(a_i, b_i) = 0,1832 < 0,239 = D_{0,10, 10} \text{ (valor crítico tabla de Lilliefors)}$$

Por lo que se acepta la hipótesis nula de normalidad para la muestra, con un nivel de significación del 10%

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

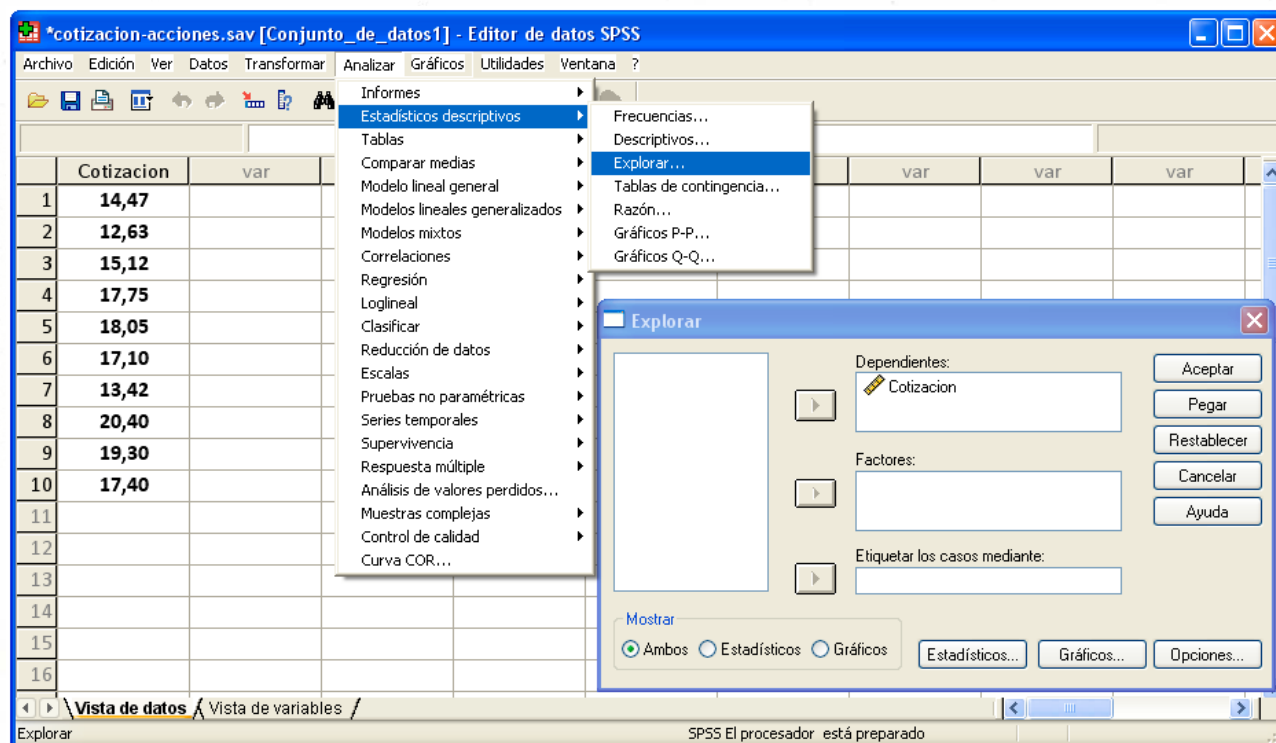
Corrección de Lilliefors para Normalidad

<i>n</i>	Nivel de significación α				
	0.20	0.15	0.10	0.05	0.01
4	0.300	0.319	0.352	0.381	0.417
5	0.285	0.299	0.315	0.337	0.405
6	0.265	0.277	0.294	0.319	0.364
7	0.247	0.258	0.276	0.300	0.348
8	0.233	0.244	0.261	0.285	0.331
9	0.223	0.233	0.249	0.271	0.311
10	0.215	0.224	0.239	0.258	0.294
11	0.206	0.217	0.230	0.249	0.284
12	0.199	0.212	0.223	0.242	0.275
13	0.190	0.202	0.214	0.234	0.268
14	0.183	0.194	0.207	0.227	0.261
15	0.177	0.187	0.201	0.220	0.257
16	0.173	0.182	0.195	0.213	0.250
17	0.169	0.177	0.189	0.206	0.245
18	0.166	0.173	0.184	0.200	0.239
19	0.163	0.169	0.179	0.195	0.235
20	0.160	0.166	0.174	0.190	0.231
21	0.157	0.162	0.170	0.186	0.226
22	0.154	0.158	0.167	0.182	0.220
23	0.150	0.154	0.164	0.179	0.213
24	0.146	0.151	0.161	0.176	0.207
25	0.142	0.147	0.158	0.173	0.200
26	0.139	0.145	0.155	0.171	0.197
27	0.137	0.143	0.153	0.168	0.194
28	0.135	0.140	0.150	0.166	0.192
29	0.133	0.138	0.147	0.163	0.190
30	0.131	0.136	0.144	0.161	0.187
<i>n</i> > 30	$0.736/\sqrt{n}$	$0.768/\sqrt{n}$	$0.805/\sqrt{n}$	$0.886/\sqrt{n}$	$1.031/\sqrt{n}$

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....



Descriptivos

			Estadístico	Error típ.
Cotizacion	Media		16,5640	,80714
	Intervalo de confianza para la media al 90%	Límite inferior	15,0844	
		Límite superior	18,0436	
	Media recortada al 5%		16,5694	
	Mediana		17,2500	
	Varianza		6,5147	
	Desv. típ.		2,5524	
	Mínimo		12,63	
	Máximo		20,40	
	Rango		7,77	
	Amplitud intercuartil		4,16	
	Asimetría		-,180	,687
	Curtosis		-1,044	1,334

Pruebas de normalidad

	Kolmogorov-Smirnov ^a			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
Cotizacion	,1832	10	,200*	,9579	10	,762

*. Este es un límite inferior de la significación verdadera.

a. Corrección de la significación de Lilliefors

El $p_valor (Sig) = 0,2 > 0,1 = \alpha \rightarrow$ Se acepta la hipótesis nula de normalidad de la muestra.

✓ **Contraste de normalidad de Shapiro-Wilk**

▪ **Contraste bilateral:**

Dada una muestra aleatoria simple de tamaño n ($x_1, x_2, \dots, x_n$)

Se establecen las hipótesis:

H_0 : La muestra procede de una distribución normal

H_1 : La muestra no procede de una distribución normal

$$\text{Estadístico contraste: } W = \frac{\left(\sum_{i=1}^k a_i \cdot (x_{n-i+1} - x_i) \right)^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad k = \begin{cases} n/2 & n \text{ par} \\ (n-1)/2 & n \text{ impar} \end{cases}$$

$x_i \equiv$ Estadístico ordenado de orden i

$a_i \equiv$ Coeficientes para cada $(x_{n-i+1} - x_i)$, con $(n, i = 1, 2, \dots, k)$ valores de la tabla de Shapiro-Wilk

Se rechaza la hipótesis de normalidad H_0 cuando $W < W_{\alpha, n}$ (valor crítico de Shapiro-Wilk)

Para calcular el estadístico W se organizan las operaciones en la siguiente tabla:

x_i	x_{n-i+1}	x_i	$x_{n-i+1} - x_i$	a_i	$a_i \cdot (x_{n-i+1} - x_i)$
20,4	20,4	12,63	7,77	0,5739	4,4592
19,3	19,3	13,42	5,88	0,3291	1,9351
18,05	18,05	14,47	3,58	0,2141	0,7665
17,75	17,75	15,12	2,63	0,1224	0,3219
17,4	17,4	17,1	0,30	0,0399	0,0120
17,1					7,4947
15,12					
14,47					
13,42					
12,63					

$$\left(\sum_{i=1}^5 a_i \cdot (x_{n-i+1} - x_i) \right)^2 = 7,4947^2 = 56,1705$$

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

Coeficientes a_i del test W de Shapiro-Wilks de normalidad

$n \backslash i$	2	3	4	5	6	7	8	9	10
1	0,7071	0,7071	0,6872	0,6646	0,6431	0,6233	0,6052	0,5888	0,5739
2		0,0000	0,1667	0,2413	0,2806	0,3031	0,3164	0,3244	0,3291
3				0,0000	0,0875	0,1401	0,1743	0,1976	0,2141
4						0,0000	0,0561	0,0947	0,1224
5								0,0000	0,0399

Se había calculado $s_x^2 = \frac{1}{9} \sum_{i=1}^{10} (x_i - \bar{x})^2 = 6,5147 \rightarrow \sum_{i=1}^{10} (x_i - \bar{x})^2 = 58,6323$

$$\text{Así pues: } W = \frac{\left(\sum_{i=1}^5 a_i \cdot (x_{n-i+1} - x_i) \right)^2}{\sum_{i=1}^{10} (x_i - \bar{x})^2} = \frac{56,1705}{58,6323} = 0,9579$$

Siendo $W = 0,9579 > 0,869 = W_{0,10,10}$ se acepta la hipótesis nula de normalidad, con un 10% de significación.

Valores críticos del test W de Shapiro-Wilks de normalidad

$$P[W < W_\alpha] = \alpha$$

n	0,01	0,02	0,05	Nivel de significación α			0,95	0,98	0,99
				0,10	0,50	0,90			
3	0,753	0,756	0,767	0,789	0,959	0,998	0,999	1,000	1,000
4	0,687	0,707	0,748	0,792	0,935	0,987	0,992	0,996	0,997
5	0,686	0,715	0,762	0,806	0,927	0,979	0,986	0,991	0,993
6	0,713	0,743	0,788	0,826	0,927	0,974	0,981	0,986	0,989
7	0,730	0,760	0,803	0,838	0,928	0,972	0,979	0,985	0,988
8	0,749	0,778	0,818	0,851	0,932	0,972	0,978	0,984	0,987
9	0,764	0,791	0,829	0,859	0,935	0,972	0,978	0,984	0,986
10	0,781	0,806	0,842	0,869	0,938	0,972	0,978	0,983	0,986

184. En la tabla aparecen los resultados de una encuesta que se realizó a cien personas sobre las veces que viajaban en avión durante un año.

x_i	1	2	3	4	5	6	7	8	9	10	11	12	14
n_i	2	5	10	14	16	12	11	8	5	6	8	2	1

Con una significación del 5%, determinar si la muestra se ajusta a una distribución de Weibull de parámetros $b = 6$ y $c = 2$, empleando la prueba de Kolmogorov-Smirnov.

Solución:

La función de Weibull se emplea con frecuencia en análisis de supervivencia (función de supervivencia, vida permanente de construcciones, vida de electrodomésticos, etc.) La función de densidad y función de distribución, respectivamente, son:

$$f(x; b, c) = \frac{c \cdot x^{c-1}}{b} \cdot e^{-(x/b)^c} \quad x \geq 0, b > 0, c > 0 \quad c \equiv \text{parámetro de forma} \quad y \quad b \equiv \text{parámetro de escala}$$

$$F(x; b, c) = 1 - e^{-(x/b)^c}$$

Con el test de Kolmogorov-Smirnov se establece el contraste bilateral:

H_0 : La muestra procede de una distribución de Weibull

H_1 : La muestra no procede de una distribución de Weibull

Estadístico de contraste: $D_n = \max |F_n(x_i) - F_0(x_i)|$

Se acepta H_0 si $D_n = \max |F_n(x_i) - F_0(x_i)| < D_{\alpha, n}$ (valor crítico tabla Kolmogorov-Smirnov)

$F_0(x_i) \equiv$ Función de distribución: $\text{DIST.WEIBULL}(x_i; 2; 6; 1)$

$F_n(x_i) \equiv$ Función de distribución empírica de la muestra: $F_n(x_i) = \frac{n^\circ \text{ observaciones}}{n} = \frac{N(x_i)}{n}$

Se construye una tabla con $F_0(x_i)$, $F_n(x_i)$ y $|F_n(x_i) - F_0(x_i)|$ para determinar el estadístico de contraste.

Las probabilidades acumuladas a cada valor x_i , es decir la función de distribución $F_n(x_i)$ se determinan de la forma siguiente:

$$F_n(x_1) = P(X \leq 1) = 1 - e^{-(1/6)^2} = 0,0274$$

$$F_n(x_2) = P(X \leq 2) = 1 - e^{-(2/6)^2} = 0,1052$$

$$\vdots \quad \quad \quad \vdots \quad \quad \quad \vdots$$

$$F_n(x_{13}) = P(X \leq 14) = 1 - e^{-(14/6)^2} = 0,9957$$

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

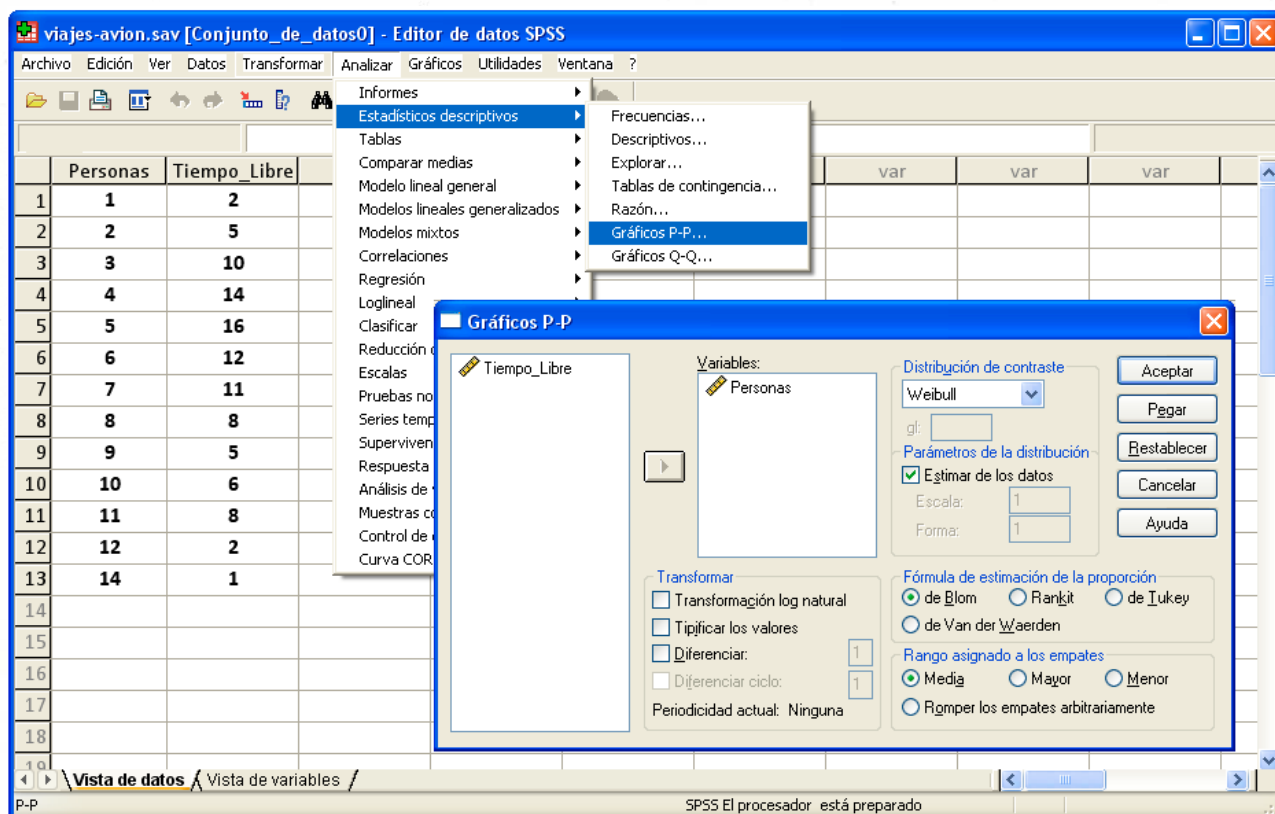
x_i	n_i	N_i	$F_n(x_i)$	$F_0(x_i)$	$ F_0(x_i) - F_n(x_i) $
1	2	2	0,020	0,0274	0,0074
2	5	7	0,070	0,1052	0,0352
3	10	17	0,170	0,2212	0,0512
4	14	31	0,310	0,3588	0,0488
5	16	47	0,470	0,5006	0,0306
6	12	59	0,590	0,6321	0,0421
7	11	70	0,700	0,7436	0,0436
8	8	78	0,780	0,8310	0,0510
9	5	83	0,830	0,8946	0,0646
10	6	89	0,890	0,9378	0,0478
11	8	97	0,970	0,9653	0,0047
12	2	99	0,990	0,9817	0,0083
14	1	100	1,000	0,9957	0,0043

$$D_n = \max |F_n(x_i) - F_0(x_i)| = 0,0646$$

Siendo $D_n = 0,0646 < 0,1358 = D_{0,05, 100} \rightarrow$ Se acepta la hipótesis nula H_0 , en consecuencia, la muestra se ajusta a una distribución de Weibull $W(6, 2)$, con una significación de 0,05

Valores críticos del test de Kolmogorov-Smirnov para una muestra

Tamaño de muestra	Test bilateral, $\alpha = 0,20$	0,10	0,050	0,02	0,010
Aproximación para $n > 40$:	$\frac{1,0730}{\sqrt{n}}$	$\frac{1,2239}{\sqrt{n}}$	$\frac{1,3581}{\sqrt{n}}$	$\frac{1,5174}{\sqrt{n}}$	$\frac{1,6276}{\sqrt{n}}$



Parámetros de distribución estimados

		Personas
Distribución de Weibull	Escala	7,021
	Forma	2,323

Los casos se ponderan por Tiempo_Libre.

185. Una compañía dedicada al envasado, fabricación y venta de productos lácteos pretende analizar el consumo anual de leche en una ciudad. Para realizar el estudio, decide llevar a cabo una encuesta a cien personas, clasificando por edades los litros de leche medios consumidos, obteniendo los siguientes datos:

$L_{i-1} - L_i$	0 - 10	10 - 20	20 - 30	30 - 40	40 - 50	50 - 60
n_i	26	21	16	15	12	10

Con un nivel de significación del 5%, empleando la prueba de Kolmogorov-Smirnov, determinar si la muestra se ajusta a una distribución uniforme $U(0,60)$

Solución:

Se establece el contraste bilateral:

H_0 : La muestra procede de una distribución de uniforme $U(0,60)$

H_1 : La muestra no procede de una distribución de uniforme $U(0,60)$

Estadístico de contraste: $D_n = \max |F_n(x_i) - F_0(x_i)|$

Se acepta H_0 si $D_n = \max |F_n(x_i) - F_0(x_i)| < D_{\alpha, n}$ (valor crítico tabla Kolmogorov-Smirnov)

$$F_0(x_i) \equiv \text{Función de distribución } U(0,60) \equiv \frac{x-0}{60-0} = \frac{x}{60}$$

$$F_n(x_i) \equiv \text{Función de distribución empírica de la muestra: } F_n(x_i) = \frac{n^\circ \text{ observaciones}}{n} = \frac{N(x_i)}{n}$$

Se construye una tabla con $F_0(x_i)$, $F_n(x_i)$ y $|F_n(x_i) - F_0(x_i)|$ para determinar el estadístico de contraste.

Las probabilidades acumuladas a cada valor x_i , es decir la función de distribución $F_n(x_i)$ se determinan de la forma siguiente:

$$F_n(x_1) = P(X \leq 10) = 10 / 60 = 0,1667$$

$$F_n(x_2) = P(X \leq 20) = 20 / 60 = 0,3333$$

$$\vdots \quad \quad \quad \vdots$$

$$F_n(x_6) = P(X \leq 60) = 60 / 60 = 1$$

$L_{i-1} - L_i$	n_i	N_i	$F_n(x_i)$	$F_0(x_i)$	$ F_0(x_i) - F_n(x_i) $
0 - 10	26	26	0,260	0,1667	0,0933
10 - 20	21	47	0,470	0,3333	0,1367
20 - 30	16	63	0,630	0,5000	0,1300
30 - 40	15	78	0,780	0,6667	0,1133
40 - 50	12	90	0,900	0,8333	0,0667
50 - 60	10	100	1,000	1,0000	0,0000

$$D_n = \max |F_n(x_i) - F_0(x_i)| = 0,1367$$

Siendo $D_n = 0,1367 > 0,1358 = D_{0,05, 100} \rightarrow$ Se rechaza la hipótesis nula H_0 , en consecuencia, la muestra no se ajusta a una distribución $U(0,60)$, con una significación de 0,05

Valores críticos del test de Kolmogorov-Smirnov para una muestra

Tamaño muestra	Test bilateral				
	$\alpha = 0,20$	0,10	0,050	0,02	0,010
Aproximación $n > 40$	$\frac{1,0730}{\sqrt{n}}$	$\frac{1,2239}{\sqrt{n}}$	$\frac{1,3581}{\sqrt{n}}$	$\frac{1,5174}{\sqrt{n}}$	$\frac{1,6276}{\sqrt{n}}$

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

*leche-consumida.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

Calcular variable

	Control	Intervalos	Frecuencia	Prob_U
1	1	1	26	,01667
2	2	2	21	,01667
3	3	3	16	,01667
4	4	4	15	,01667
5	5	5	12	,01667
6	6	6	10	,01667
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				

Variable de destino: Prob_U = Expresión numérica: PDF.UNIFORM(Control,0,60)

Tipo y etiqueta...

Control
Intervalos
Frecuencia
Prob_U
Fun_Distr_U

Grupo de funciones:
Aritméticas
Búsqueda
Cadena
Conversión
Creación de duración del tier
Creación de fechas
Cálculo de fechas
Estadísticas
Extensión de duración del tier

Funciones y variables especiales:
\$Casenum
\$Date
\$Date11
\$Jdate
\$Sysmis
\$Time
Abs
Any
Arasin
Artan
Cdf.Bernoulli
Cdf.Beta

CDF.UNIFORM(cant,mín,máx). Numérico.
Devuelve la probabilidad acumulada de que un valor de la distribución uniforme, con el mínimo y el máximo especificados, sea menor que la cantidad cant.

Si... (condición de selección de casos opcional)

Aceptar Pegar Restablecer Cancelar Ayuda

Vista de datos / Vista de variables /

SPSS El procesador está preparado

*leche-consumida.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

Calcular variable

	Control	Intervalos	Frecuencia	Prob_U	Fun_Distr
1	1	1	26	,0167	,01667
2	2	2	21	,0167	,03333
3	3	3	16	,0167	,05000
4	4	4	15	,0167	,06667
5	5	5	12	,0167	,08333
6	6	6	10	,0167	,10000
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					

Variable de destino: Fun_Distr_U = Expresión numérica: CDF.UNIFORM(Control,0,60)

Tipo y etiqueta...

Control
Intervalos
Frecuencia
Prob_U
Fun_Distr_U

Grupo de funciones:
Aritméticas
Búsqueda
Cadena
Conversión
Creación de duración del tier
Creación de fechas
Cálculo de fechas
Estadísticas
Extensión de duración del tier

Funciones y variables especiales:
\$Casenum
\$Date
\$Date11
\$Jdate
\$Sysmis
\$Time
Abs
Any
Arasin
Artan
Cdf.Bernoulli
Cdf.Beta

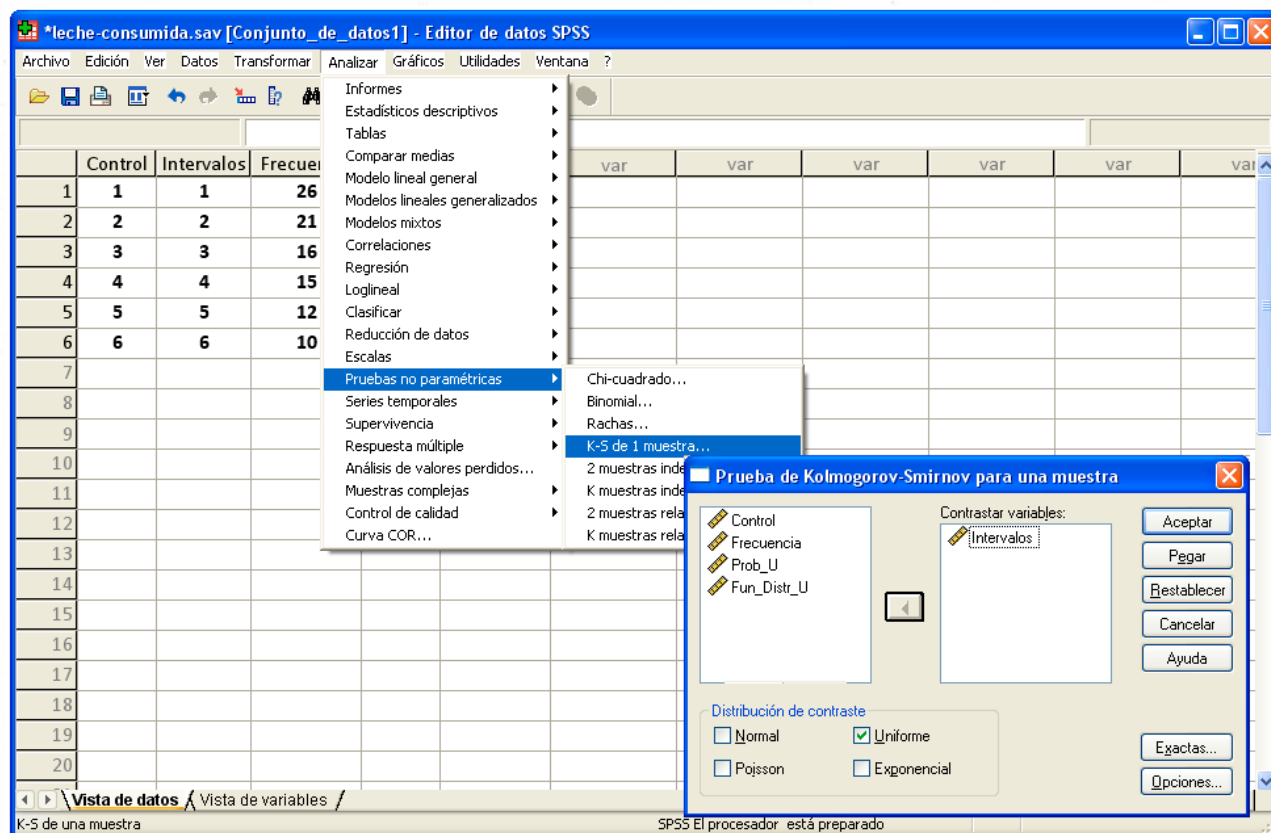
CDF.UNIFORM(cant,mín,máx). Numérico.
Devuelve la probabilidad acumulada de que un valor de la distribución uniforme, con el mínimo y el máximo especificados, sea menor que la cantidad cant.

Si... (condición de selección de casos opcional)

Aceptar Pegar Restablecer Cancelar Ayuda

Vista de datos / Vista de variables /

SPSS El procesador está preparado



Prueba de Kolmogorov-Smirnov para una muestra

		Intervalos
N		100
Parámetros uniformes^{a,b}	Mínimo	1
	Máximo	6
Diferencias más extremas	Absoluta	,270
	Positiva	,270
	Negativa	-,100
Z de Kolmogorov-Smirnov		2,700
Sig. asintót. (bilateral)		,000

a. La distribución de contraste es la Uniforme.

b. Se han calculado a partir de los datos.

El p_valor (Sig. asintótica bilateral) = $0 < 0,05 = \alpha \rightarrow$ Se rechaza la hipótesis nula H_0 , concluyendo que la muestra no se ajusta a una distribución $U(0,60)$, con una significación de 0,05

186. En la tabla aparecen las puntuaciones obtenidas en un test de estrés que se ha pasado en dos ciudades.

A	18	13	22	9	12	21	14	17	15	10	16	19	11
B	12	16	11	15	9	8	13	6	9	13	7	22	

Con un nivel de significación del 1%, a partir de los datos, puede afirmarse que el estrés es mayor en la ciudad A que en la ciudad B. Utilizar el test de la U de Wilcoxon-Mann-Whitney, el contraste de Kolmogorov-Smirnov y la prueba de la Mediana.

Solución:

Sean las variables aleatorias:

$X \equiv$ Puntuaciones de estrés en el test de la ciudad A

$Y \equiv$ Puntuaciones de estrés en el test de la ciudad B

Se realiza el contraste unilateral a la izquierda: $H_0: F(z) \geq G(z)$ $H_1: F(z) < G(z)$

Siendo F y G las respectivas funciones de distribución de X e Y

✓ *Utilizando el test de la U de Wilcoxon-Mann-Whitney*

Estadístico de contraste: $U = U_{x_A} = n_A \cdot n_B + \frac{n_A \cdot (n_A + 1)}{2} - W_{x_1}$

siendo $W_{x_1} \equiv \sum_{i=1}^n R_i$ suma de rangos de la muestra X_1

Como $n_A = 13 > 10$ y $n_B = 12 > 10$ la distribución del estadístico de contraste U se puede aproximar por una normal

$$U = U_{x_1} \sim N\left(\frac{n_A \cdot n_B}{2}, \sqrt{\frac{n_A \cdot n_B \cdot (n_A + n_B + 1)}{12}}\right)$$

$$E(U) = \frac{n_A \cdot n_B}{2} = \frac{13 \cdot 12}{2} = 78 \quad \sigma_U = \sqrt{\frac{n_A \cdot n_B \cdot (n_A + n_B + 1)}{12}} = \sqrt{\frac{13 \cdot 12 \cdot (13 + 12 + 1)}{12}} = 18,3848$$

$$\text{Se acepta } H_0 \text{ si } z_p = \frac{U - \frac{n_A \cdot n_B}{2}}{\sqrt{\frac{n_A \cdot n_B \cdot (n_A + n_B + 1)}{12}}} \leq z_\alpha$$

Para calcular el valor experimental de U se ordenan las observaciones muestrales de menor a mayor, asignando un rango de 1 hasta 25. Si hubiera observaciones repetidas, se asigna el rango medio de los que les corresponderían si fueran diferentes.

Para facilitar el cálculo $W_{x_1} \equiv \sum_{i=1}^{25} c_i \cdot R_i$ se crea el coeficiente instrumental $c_i = \begin{cases} 1 & \text{valor } X_1 \\ 0 & \text{valor } X_2 \end{cases}$

Excel: JERARQUIA(Observación 1 ; Observaciones 1 : 25 ; Poisición 1 : 25)

Poisición	Observaciones	Jerarquía	R_i	c_i	$c_i \cdot R_i$
1	6	1	1	0	0
2	7	2	2	0	0
3	8	3	3	0	0
4	9	4	5	0	0
5	9	4	5	0	0
6	9	4	5	1	5
7	10	7	7	1	7
8	11	8	8,5	0	0
9	11	8	8,5	1	8,5
10	12	10	10,5	0	0
11	12	10	10,5	1	10,5
12	13	12	13	0	0
13	13	12	13	0	0
14	13	12	13	1	13
15	14	15	15	1	15
16	15	16	16,5	0	0
17	15	16	16,5	1	16,5
18	16	18	18,5	0	0
19	16	18	18,5	1	18,5
20	17	20	20	1	20
21	18	21	21	1	21
22	19	22	22	1	22
23	21	23	23	1	23
24	22	24	24,5	0	0
25	22	24	24,5	1	24,5
					204,5

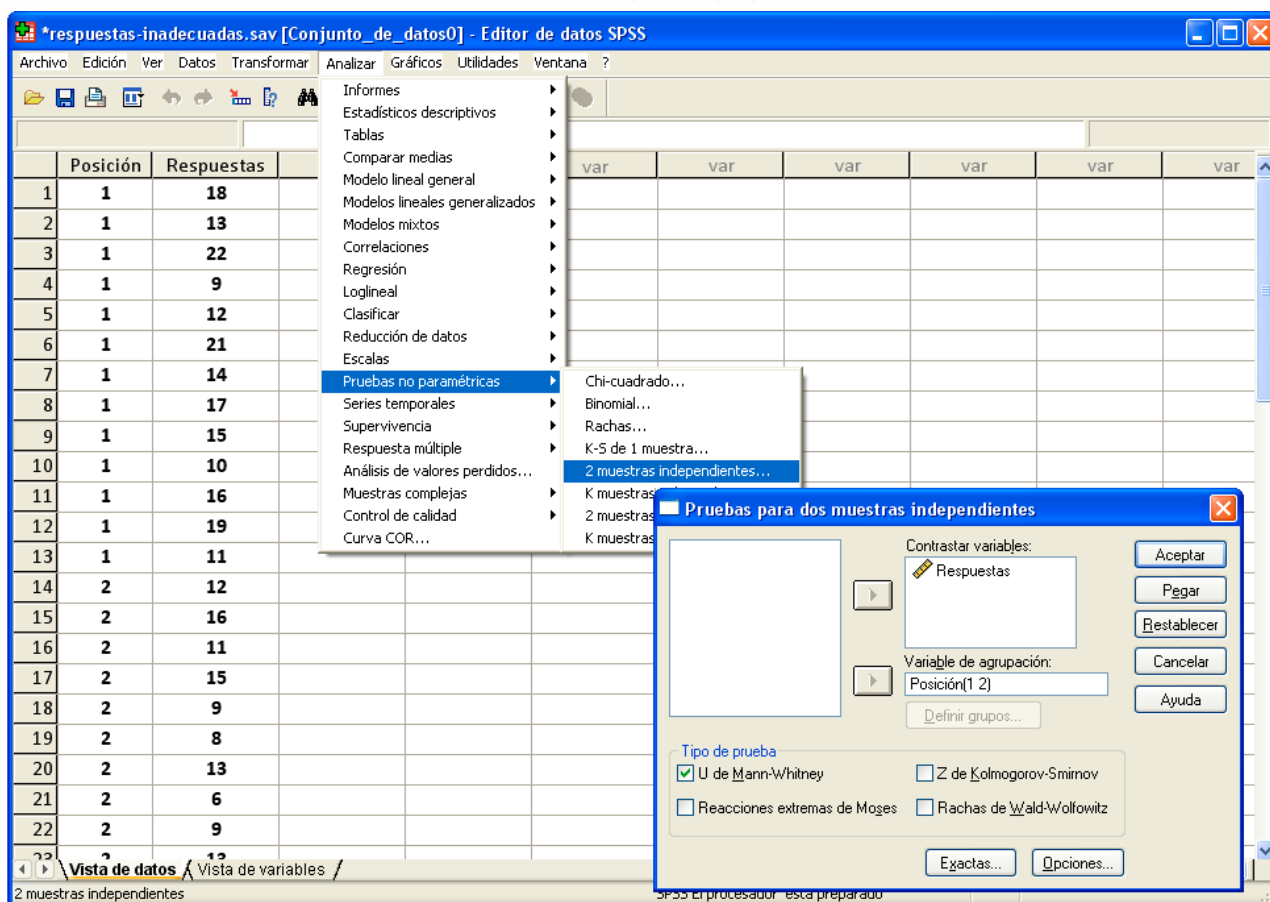
$$W_{x_1} \equiv \sum_{i=1}^{25} c_i \cdot R_i = 204,5$$

En consecuencia,

$$U = U_{x_A} = n_A \cdot n_B + \frac{n_A \cdot (n_A + 1)}{2} - W_{x_1} = 13 \cdot 12 + \frac{13 \cdot 14}{2} - 204,5 = 42,5$$

$$\text{Estadístico contraste: } z_p = \frac{U - \frac{n_A \cdot n_B}{2}}{\sqrt{\frac{n_A \cdot n_B \cdot (n_A + n_B + 1)}{12}}} = \frac{42,5 - 78}{18,3848} = -1,9309$$

Como $z_p = -1,9309 > -2,33 = z_{0,01}$ se acepta la hipótesis nula con una significación de 0,051.



*respuestas-inadecuadas.sav [Conjunto_de_datos0] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

Informes
Estadísticos descriptivos
Tablas
Comparar medias
Modelo lineal general
Modelos lineales generalizados
Modelos mixtos
Correlaciones
Regresión
Loglineal
Clasificar
Reducción de datos
Escalas
Pruebas no paramétricas
Series temporales
Supervivencia
Respuesta múltiple
Análisis de valores perdidos...
Muestras complejas
Control de calidad
Curva COR...

Chi-cuadrado...
Binomial...
Rachas...
K-S de 1 muestra...
2 muestras independientes...
K muestras...
2 muestras...
K muestras...

Pruebas para dos muestras independientes

Contrastar variables:
Respuestas

Variable de agrupación:
Posición(1 2)

Definir grupos...

Tipo de prueba
☒ U de Mann-Whitney
☐ Z de Kolmogorov-Smirnov
☐ Reacciones extremas de Moises
☐ Rachas de Wald-Wolfowitz

Exactas... Opciones...

Aceptar Pegar Restablecer Cancelar Ayuda

Vista de datos Vista de variables /
2 muestras independientes

Estadísticos de contraste^b

	Respuestas
U de Mann-Whitney	42,500
W de Wilcoxon	120,500
Z	-1,936
Sig. asintót. (bilateral)	,053
Sig. exacta [2*(Sig. unilateral)]	42,500 ^a

a. No corregidos para los empates.

b. Variable de agrupación: Posición

En el contraste unilateral a la izquierda, el $p_valor = 1 - 0,053 / 2 = 0,9735 > 0,01 = \alpha$, se acepta la hipótesis nula de que el estrés es mayor en la ciudad A.

✓ *Utilizando el contraste de Kolmogorov-Smirnov para dos muestras*

Se plantea el contraste unilateral a la izquierda: $H_0: F(z) \geq G(z)$ $H_1: F(z) < G(z)$

Estadístico de contraste: $D_{n_1, n_2}^- = \max [G_{n_2}(x) - F_{n_1}(x)]$

$$\text{con } F_{n_1}(x) = \frac{N_1(x)}{n_1} \quad F_{n_2}(x) = \frac{N_2(x)}{n_2}$$

Se rechaza H_0 si $D_{n_1, n_2}^- > D_{\alpha, n_1, n_2}$

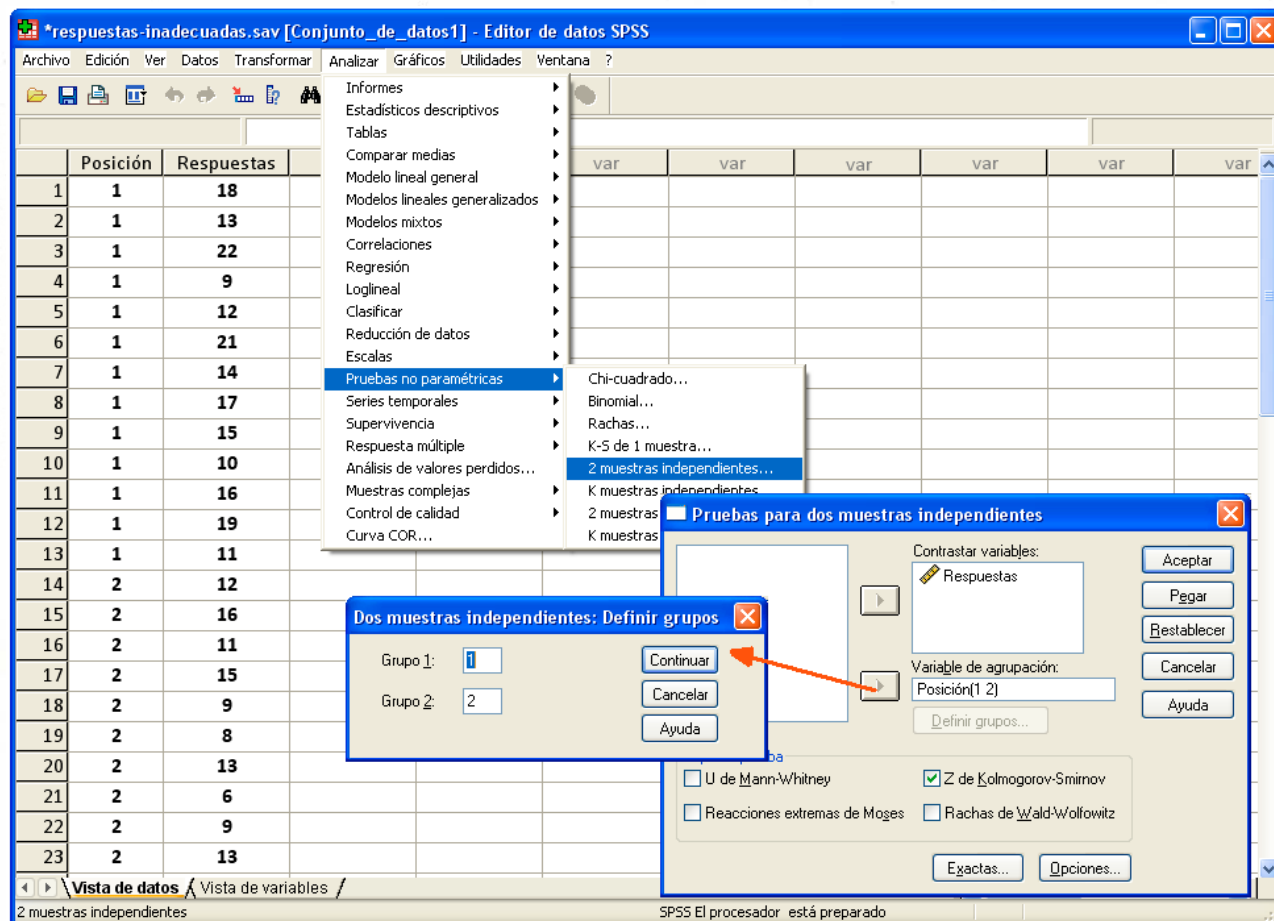
Se crea una tabla haciendo los cálculos necesarios para obtener el valor experimental del estadístico de contraste D_{n_1, n_2}

x_i	n_{x_i}	y_i	n_{y_i}	N_{x_i}	N_{y_i}	$F_{n_1}(x)$	$G_{n_2}(x)$	$G_{n_2}(x) - F_{n_1}(x)$
	0	6	1	0	1	0	0,0833	0,0833
	0	7	1	0	2	0	0,1667	0,1667
	0	8	1	0	3	0	0,2500	0,2500
9	1	9	2	1	5	0,0769	0,4167	0,3397
10	1		0	2	5	0,1538	0,4167	0,2628
11	1	11	1	3	6	0,2308	0,5000	0,2692
12	1	12	1	4	7	0,3077	0,5833	0,2756
13	1	13	2	5	9	0,3846	0,7500	0,3654
14	1		0	6	9	0,4615	0,7500	0,2885
15	1	15	1	7	10	0,5385	0,8333	0,2949
16	1	16	1	8	11	0,6154	0,9167	0,3013
17	1		0	9	11	0,6923	0,9167	0,2244
18	1		0	10	11	0,7692	0,9167	0,1474
19	1		0	11	11	0,8462	0,9167	0,0705
21	1		0	12	11	0,9231	0,9167	-0,0064
22	1	22	1	13	12	1	1	0
13			12					

$$D_{n_1, n_2}^- = \max [G_{n_2}(x) - F_{n_1}(x)] = 0,3654 \quad D_{n_1, n_2}^+ = \max [F_{n_1}(x) - G_{n_2}(x)] = -0,0064$$

$$\text{El valor crítico } D_{\alpha, n_1, n_2} = \sqrt{\frac{n_1 + n_2}{n_1 \cdot n_2}} \cdot C_{KS} = \sqrt{\frac{13 + 12}{13 \cdot 12}} \cdot 1,5174 = 0,60744$$

Siendo $D_{n_1, n_2}^- = 0,3654 < 0,60744 = D_{\alpha, n_1, n_2} \rightarrow$ Se acepta la hipótesis nula al 1% de significación. En consecuencia, el estrés es mayor en la ciudad A.



*respuestas-inadecuadas.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

Informes
Estadísticos descriptivos
Tablas
Comparar medias
Modelo lineal general
Modelos lineales generalizados
Modelos mixtos
Correlaciones
Regresión
Loglineal
Clasificar
Reducción de datos
Escalas
Pruebas no paramétricas
Series temporales
Supervivencia
Respuesta múltiple
Análisis de valores perdidos...
Muestras complejas
Control de calidad
Curva COR...

Chi-cuadrado...
Binomial...
Rachas...
K-S de 1 muestra...
2 muestras independientes...
K muestras independientes...
2 muestras
K muestras

Pruebas para dos muestras independientes

Contrastar variables:
Respuestas

Variable de agrupación:
Posición(1 2)

Definir grupos...

☐ U de Mann-Whitney ☒ Z de Kolmogorov-Smirnov
☐ Reacciones extremas de Moises ☐ Rachas de Wald-Wolfowitz

Exactas... Opciones...

Dos muestras independientes: Definir grupos

Grupo 1: 1 Continuar
Grupo 2: 2 Cancelar
Ayuda

Vista de datos Vista de variables /
2 muestras independientes SPSS El procesador está preparado

Estadísticos de contraste^a

		Respuestas
Diferencias más extremas	Absoluta	,3654
	Positiva	,0064
	Negativa	-,3654
Z de Kolmogorov-Smirnov		,9127
Sig. asintót. (bilateral)		,3754

a. Variable de agrupación: Posición

El $p_{\text{valor}} = 1 - 0,3754 / 2 = 0,8123 > 0,01 = \alpha$ del contraste unilateral a la izquierda $F(z) \geq G(z)$, con lo que se acepta la hipótesis nula, con una significación del 1%. En consecuencia, la ciudad A presenta mayor estrés.

✓ **Utilizando la prueba de la Mediana**

Se establece el contraste unilateral a la izquierda: $H_0: M_{e_x} \geq M_{e_y}$ $H_1: M_{e_x} < M_{e_y}$

donde M_{e_x} y M_{e_y} son las correspondientes medianas poblacionales.

Estadístico de contraste: V = Número de observaciones de X menores o iguales que la mediana de la muestra conjunta de $n_1 + n_2$ elementos.

Para calcular el valor experimental de V se calcula la mediana de la muestra conjunta, habiendo ordenado todas las observaciones de menor a mayor.

6	7	8	9	9	9	10	11	11	12	12	13	13	13	14	15	15	16	16	17	18	19	21	22	22
---	---	---	---	---	---	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

Como $n = n_1 + n_2 = 13 + 12 = 25$ es impar, la mediana será la observación que ocupa el lugar

$$\frac{(n+1)}{2} = \frac{(25+1)}{2} = 13, \text{ es decir, } m_e = 13$$

x_i	9	10	11	12	13	14	15	16	17	18	19	21	22
-------	---	----	----	----	----	----	----	----	----	----	----	----	----

→ $V = 5$

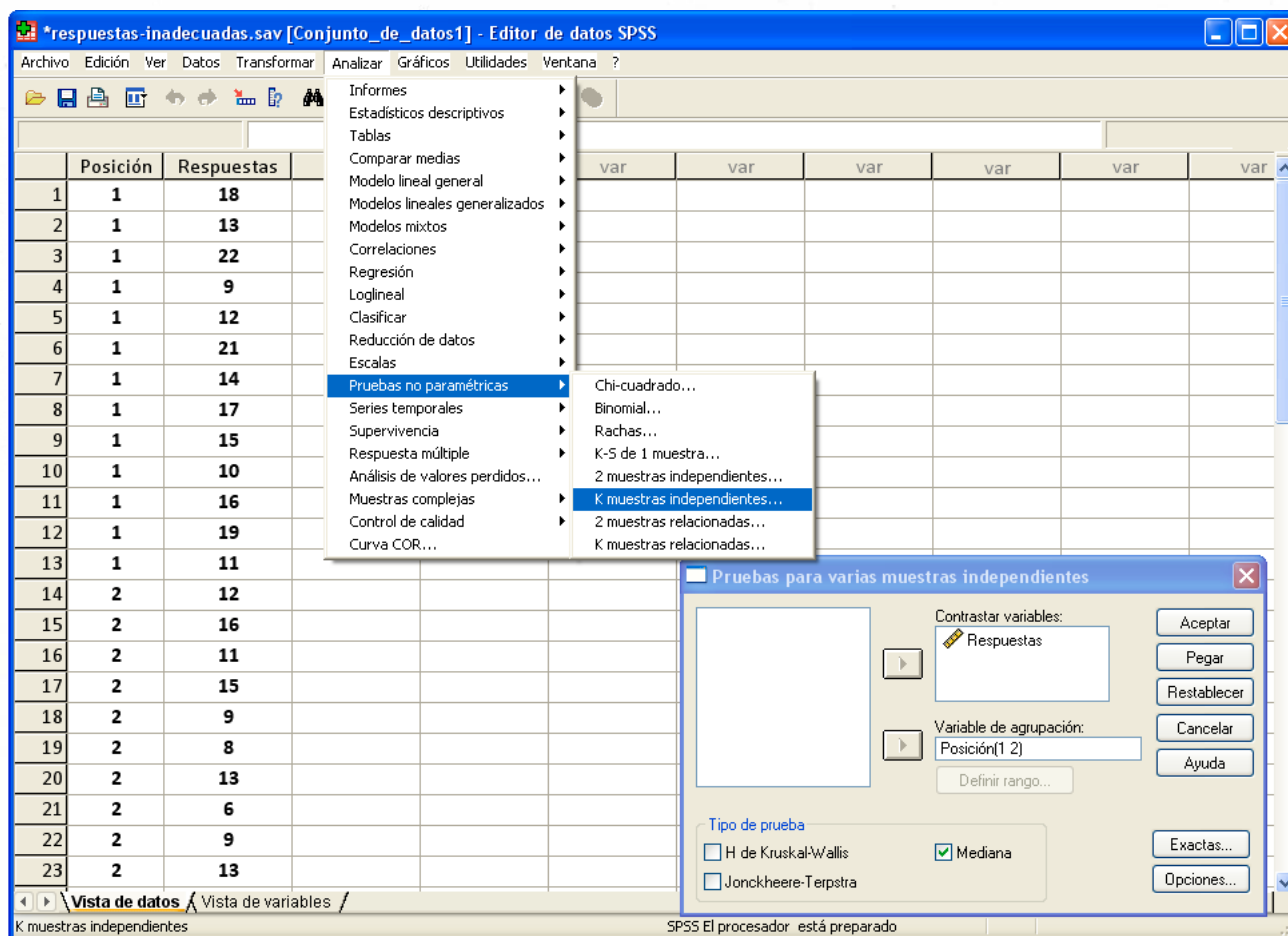
Siendo $n_1 = 13 > 10$ y $n_2 = 12 > 10$ se aproxima mediante una distribución normal:

$$V \sim N(E[V], \text{Var}[V]) \text{ con } n = 25 \text{ impar} \rightarrow k = \frac{n-1}{2} = 12$$

$$E[V] = k \cdot \frac{n_1}{n} = 12 \cdot \frac{13}{25} = 6,24 \quad \text{Var}[V] = k \cdot \frac{n_1}{n} \cdot \frac{n_2}{n} \cdot \frac{n-k}{n-1} = 12 \cdot \frac{13}{25} \cdot \frac{12}{25} \cdot \frac{25-12}{25-1} = 1,6224$$

$$\text{Estadístico contraste: } z_p = \frac{V - E[V]}{\sqrt{\text{Var}[V]}} = \frac{5 - 6,24}{\sqrt{1,6224}} = -0,9735$$

Como $z_p = -0,9735 > -2,33 = -z_{0,01}$ se acepta la hipótesis nula, por lo que se puede afirmar con una significación del 1%, que el estrés es mayor en la ciudad A.



*respuestas-inadecuadas.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

Informes
Estadísticos descriptivos
Tablas
Comparar medias
Modelo lineal general
Modelos lineales generalizados
Modelos mixtos
Correlaciones
Regresión
Loglineal
Clasificar
Reducción de datos
Escalas
Pruebas no paramétricas
Series temporales
Supervivencia
Respuesta múltiple
Análisis de valores perdidos...
Muestras complejas
Control de calidad
Curva COR...

Chi-cuadrado...
Binomial...
Rachas...
K-S de 1 muestra...
2 muestras independientes...
K muestras independientes...
2 muestras relacionadas...
K muestras relacionadas...

Pruebas para varias muestras independientes

Contrastar variables:
Respuestas

Variable de agrupación:
Posición(1 2)

Definir rango...

Tipo de prueba
☐ H de Kruskal-Wallis
☒ Mediana
☐ Jonckheere-Terpstra

Aceptar
Pegar
Restablecer
Cancelar
Ayuda
Exactas...
Opciones...

Vista de datos / Vista de variables /
K muestras independientes

SPSS El procesador está preparado

Prueba de la mediana

Frecuencias

		Posición	
		1	2
Respuestas	> Mediana	8	3
	<= Mediana	5	9

Estadísticos de contraste^a

	Respuestas
N	25
Mediana	13,00
Sig. exacta	,111

a. Variable de agrupación: Posición

El contraste unilateral a la izquierda tiene un $p_valor = 1 - 0,111 / 2 = 0,9445 > 0,01 = z_{0,01}$ por lo que se acepta la hipótesis nula de que el estrés es mayor en la ciudad A, con una significación del 1%.

 **187.** Durante los meses de declaración de la renta una sucursal bancaria decidió dar cita previa a los clientes para evitar largas esperas. A pesar de la media, los clientes esperan un tiempo medio de 5 minutos antes de ser atendidos. Determinado día se eligió al azar anotar los tiempos de espera de cada uno de los clientes, contabilizando los tiempos:

4 0 3 9 6,5 7,5 4 6 5,5

Para un nivel de significación del 5%, ¿se ajusta la muestra a una distribución exponencial?

Solución:

La distribución exponencial es una distribución de probabilidad continua con parámetro $\lambda > 0$, con función de densidad $f(x_i) = P(x_i) = \lambda \cdot e^{-\lambda \cdot x_i}$ y con función de distribución

$$F(x_i) = P(X \leq x_i) = 1 - e^{-\lambda \cdot x_i} \text{ para } x_i \geq 0: E(X) = \frac{1}{\lambda}, V(X) = \frac{1}{\lambda^2}$$

Siendo la variable aleatoria $X \equiv$ Tiempo de espera de cada uno de los clientes, $X \sim \exp(\lambda = 5)$

$$\text{En consecuencia, } E(X) = \frac{1}{\lambda} = 5 \rightarrow \lambda = \frac{1}{5} = 0,2$$

Se plantea el contraste bilateral: $H_0: X \sim \exp(0,2)$ $H_1: X \not\sim \exp(0,2)$

$$H_0: F_n(x) = F_0(x) \quad H_1: F_n(x) \neq F_0(x) \quad \text{siendo} \quad \begin{cases} F_n(x) \equiv \text{verdadera distribución de la variable } X \\ F_0(x) \equiv \text{función de distribución de } \exp(0,2) \end{cases}$$

Utilizando el *test de Kolmogorov-Smirnov*

$$\text{Estadístico de contraste: } D_n = \max |F_n(x_i) - F_0(x_i)| = \max(a_i, b_i)$$

$$a_i = |F_n(x_i) - F_0(x_i)| \quad b_i = |F_n(x_{i-1}) - F_0(x_i)|$$

Se acepta H_0 si $D_n = \max |F_n(x_i) - F_0(x_i)| < D_{\alpha, n}$ (valor crítico tabla Kolmogorov-Smirnov)

Se construye una tabla ordenando las observaciones y calculando $F_0(x_i)$, $F_n(x_i)$, $|F_n(x_i) - F_0(x_i)|$ y $|F_n(x_{i-1}) - F_0(x_i)|$

$F_0(x_i) \equiv$ Función de distribución: DISTR.EXP

$$F_n(x_i) \equiv \text{Función de distribución empírica de la muestra: } F_n(x_i) = \frac{\text{nº observaciones}}{n} = \frac{N(x_i)}{n}$$

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

x_i	n_i	N_i	$F_n(x_i)$	$F_0(x_i)$	$ F_n(x_i) - F_0(x_i) $	$ F_n(x_{i-1}) - F_0(x_i) $
0	1	1	0,1111	0	0,1111	0
3	1	2	0,2222	0,4512	0,2290	0,3401
4	2	4	0,4444	0,5507	0,1062	0,3284
5,5	1	5	0,5556	0,6671	0,1116	0,2227
6	1	6	0,6667	0,6988	0,0321	0,1433
6,5	1	7	0,7778	0,7275	0,0503	0,0608
7,5	1	8	0,8889	0,7769	0,1120	0,0009
9	1	9	1	0,8347	0,1653	0,0542

$$F_n(0) = P(X \leq 0) = 1 - e^{-0,2 \cdot 0} = 0$$

$$F_n(3) = P(X \leq 3) = 1 - e^{-0,2 \cdot 3} = 0,4512$$

$$F_n(4) = P(X \leq 4) = 1 - e^{-0,2 \cdot 4} = 0,5507$$

$$\vdots = \vdots = \vdots = \vdots$$

$$F_n(9) = P(X \leq 9) = 1 - e^{-0,2 \cdot 9} = 0,8347$$

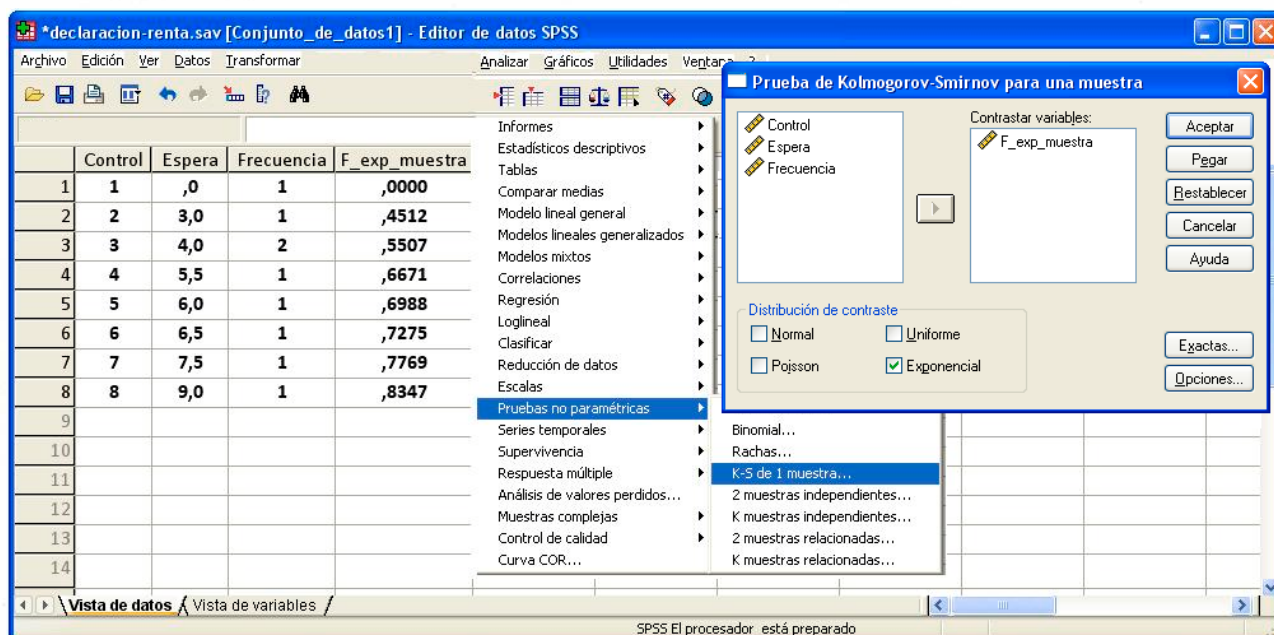
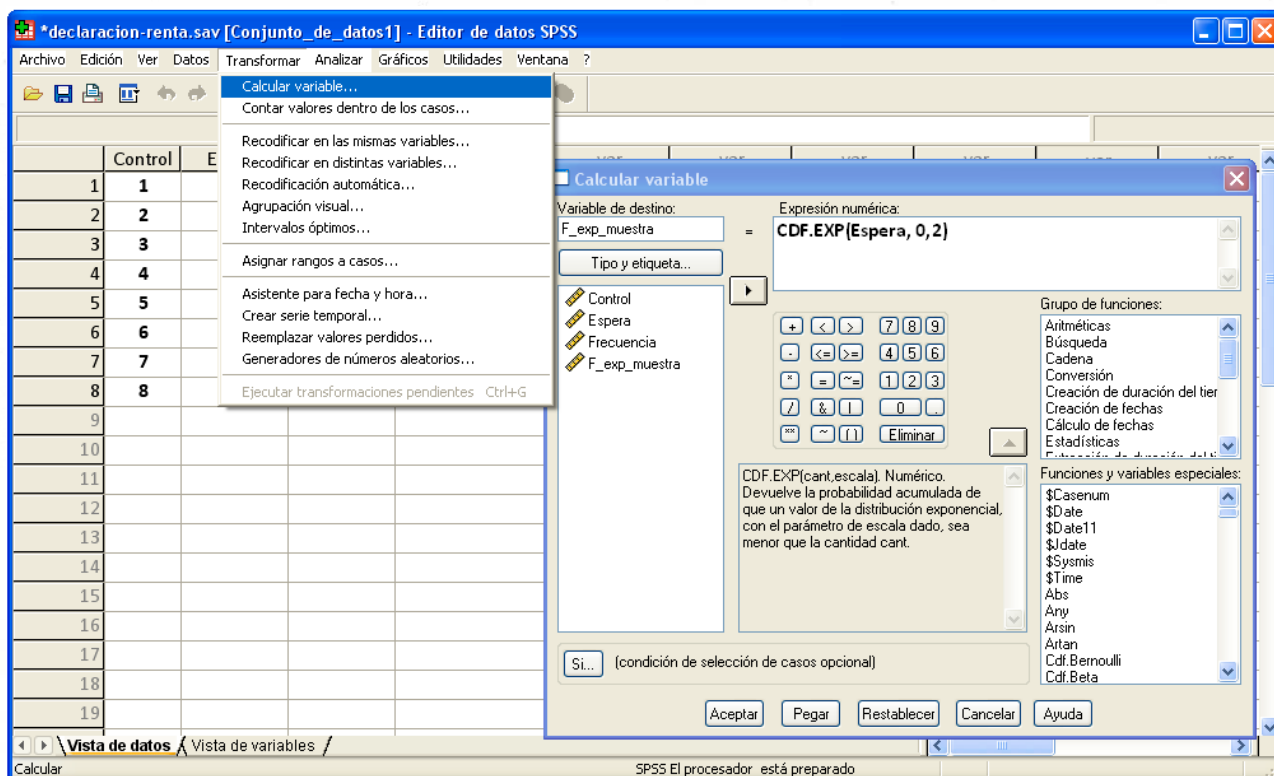
$$D_n = \max |F_n(x_i) - F_0(x_i)| = \max(a_i, b_i) = 0,3401$$

En la tabla de Kolmogorov-Smirnov: $D_{0,05,9} = 0,387$

siendo, $D_n = 0,3401 < 0,387 = D_{0,05,9} \rightarrow$ Se acepta la hipótesis nula, con un 5% de significación, afirmando que la espera de los clientes seguía una distribución exponencial de parámetro $\lambda = 0,2$

Valores críticos del test de Kolmogorov-Smirnov para una muestra

Tamaño muestra	$\alpha = 0,10$	Test unilateral			
		0,05	0,025	0,01	0,005
6	0,410	0,468	0,519	0,577	0,617
7	0,381	0,436	0,483	0,538	0,576
8	0,358	0,410	0,454	0,507	0,542
9	0,339	0,387	0,430	0,480	0,513
10	0,323	0,369	0,409	0,457	0,489



Prueba de Kolmogorov-Smirnov para una muestra

	F_exp_muestra
N	9 ^c
Parámetro exponencial. ^{a,b}	
Media	,6572
Diferencias más extremas	
Absoluta	,406
Positiva	,406
Negativa	
Z de Kolmogorov-Smirnov	1,148
Sig. asintót. (bilateral)	,143

$p_{\text{valor}} = 0,143 > 0,05 = \alpha \rightarrow$ Se acepta la hipótesis nula, concluyendo que los datos se ajustan a una distribución exponencial de parámetro $\lambda = 0,2$

188. En un test de coeficiente de inteligencia se anotaron las respuestas incorrectas que figuran en la tabla siguiente:

Intervalos $L_{i-1} - L_i$	50 – 60	60 – 70	70 – 80	80 – 90	90 – 100	100 – 110
Frecuencia n_i	30	25	120	10	10	5

Con un nivel de significación del 5%, ¿se ajusta la muestra a una distribución exponencial? Utilizar el test de Kolmogorov-Smirnov y la prueba de la Chi-cuadrado.

Solución:

✓ Utilizando el test de Kolmogorov-Smirnov

Siendo la variable aleatoria $X \equiv$ Respuestas inadecuadas en test de inteligencia, $X \sim \exp(\lambda)$

Se plantea el contraste bilateral: $H_0: X \sim \exp(\lambda)$ $H_1: X \not\sim \exp(\lambda)$

$H_0: F_n(x) = F_0(x)$ $H_1: F_n(x) \neq F_0(x)$ siendo $\begin{cases} F_n(x) \equiv \text{verdadera distribución de la variable } X \\ F_0(x) \equiv \text{función de distribución de } \exp(\lambda) \end{cases}$

Estadístico de contraste: $D_n = \max |F_n(x_i) - F_0(x_i)| = \max(a_i, b_i)$

$a_i = |F_n(x_i) - F_0(x_i)|$ $b_i = |F_n(x_{i-1}) - F_0(x_i)|$

Se acepta H_0 si $D_n = \max |F_n(x_i) - F_0(x_i)| < D_{\alpha, n}$ (valor crítico tabla Kolmogorov-Smirnov)

$F_0(x_i) \equiv$ Función de distribución: $\text{DISTR.EXP}(x_i; (1/71); x_1: x_6)$

$F_n(x_i) \equiv$ Función de distribución empírica de la muestra: $F_n(x_i) = \frac{\text{nº observaciones}}{n} = \frac{N(x_i)}{n}$

Se construye una tabla ordenando las observaciones y calculando $F_0(x_i)$, $F_n(x_i)$, $|F_n(x_i) - F_0(x_i)|$ y $|F_n(x_{i-1}) - F_0(x_i)|$

Se utiliza el estimador de máxima verosimilitud del parámetro λ : $\hat{\lambda} = \bar{x}$

La función de densidad y la función de distribución que se emplea, respectivamente, será:

$$f(x) = \frac{1}{\lambda} \cdot e^{-(1/\lambda) \cdot x} \quad \text{y} \quad F(x) = P(X \leq x) = 1 - e^{-(1/\lambda) \cdot x}$$

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

x_i	n_i	$x_i \cdot n_i$	N_i	$F_n(x_i)$	$F_0(x_i)$	$ F_n(x_i) - F_0(x_i) $	$ F_n(x_{i-1}) - F_0(x_i) $
55	30	1650	30	0,300	0,5391	0,2391	0,5391
65	25	1625	55	0,550	0,5997	0,0497	0,2997
75	20	1500	75	0,750	0,6523	0,0977	0,1023
85	10	850	85	0,850	0,6980	0,1520	0,0520
95	10	950	95	0,950	0,7376	0,2124	0,1124
105	5	525	100	1,000	0,7721	0,2279	0,1779
	100	7100					

$$\hat{\lambda} = \bar{x} = \frac{1}{100} \sum_{i=1}^6 x_i \cdot n_i = \frac{7100}{100} = 71 \rightarrow f(x) = \frac{1}{71} \cdot e^{-(1/71) \cdot x} \quad y \quad F(x) = P(X \leq x) = 1 - e^{-(1/71) \cdot x}$$

$$F_n(55) = P(X \leq 55) = 1 - e^{-(1/71) \cdot 55} = 0,5391$$

$$F_n(65) = P(X \leq 65) = 1 - e^{-(1/71) \cdot 65} = 0,5997$$

$$F_n(75) = P(X \leq 75) = 1 - e^{-(1/71) \cdot 75} = 0,6523$$

$$\vdots = \vdots = \vdots = \vdots$$

$$F_n(105) = P(X \leq 105) = 1 - e^{-(1/71) \cdot 105} = 0,7721$$

$$D_n = \max |F_n(x_i) - F_0(x_i)| = \max(a_i, b_i) = 0,5391$$

En la tabla de Kolmogorov-Smirnov: $D_{0,05, 100} = 0,13581$

siendo, $D_n = 0,5391 > 0,13581 = D_{0,05, 9} \rightarrow$ Se rechaza la hipótesis nula, con un 5% de significación, afirmando que las respuestas inadecuadas del test de inteligencia no siguen una distribución exponencial de parámetro $\lambda = 1/71$

Valores críticos del test de Kolmogorov-Smirnov para una muestra

Tamaño muestra	Test bilateral				
	$\alpha = 0,20$	0,10	0,050	0,02	0,010
Aproximación $n > 40$	$\frac{1,0730}{\sqrt{n}}$	$\frac{1,2239}{\sqrt{n}}$	$\frac{1,3581}{\sqrt{n}}$	$\frac{1,5174}{\sqrt{n}}$	$\frac{1,6276}{\sqrt{n}}$

✓ **Prueba de Chi-cuadrado**

Se plantea el contraste bilateral:

H_0 : Las respuestas inadecuadas siguen una distribución $\exp(\lambda)$

H_1 : Las respuestas inadecuadas no siguen una distribución $\exp(\lambda)$

La función de densidad que se emplea será: $f(x) = \frac{1}{71} \cdot e^{-(x/71)}$

Intervalos $L_{i-1} - L_i$	50 – 60	60 – 70	70 – 80	80 – 90	90 – 100	100 – 110
x_i	55	65	75	85	95	105
$p_i = P(x = x_i)$	0,0649	0,0564	0,0490	0,0425	0,0370	0,0321
$e_i = n \cdot p_i$	6,49	5,64	4,90	4,26	3,70	3,21

$$p_1 = P(50 \leq X \leq 60) = \int_{50}^{60} \frac{1}{71} e^{-x/71} dx = -e^{-60/71} + e^{-50/71} = 0,0649$$

$$p_2 = P(60 \leq X \leq 70) = \int_{60}^{70} \frac{1}{71} e^{-x/71} dx = -e^{-70/71} + e^{-60/71} = 0,0564$$

$$p_3 = P(70 \leq X \leq 80) = \int_{70}^{80} \frac{1}{71} e^{-x/71} dx = -e^{-80/71} + e^{-70/71} = 0,0490$$

$$p_4 = P(80 \leq X \leq 90) = \int_{80}^{90} \frac{1}{71} e^{-x/71} dx = -e^{-90/71} + e^{-80/71} = 0,0425$$

$$p_5 = P(90 \leq X \leq 100) = \int_{90}^{100} \frac{1}{71} e^{-x/71} dx = -e^{-100/71} + e^{-90/71} = 0,0370$$

$$p_6 = P(100 \leq X \leq 110) = \int_{100}^{110} \frac{1}{71} e^{-x/71} dx = -e^{-110/71} + e^{-100/71} = 0,0321$$

$$P(X = x_i) = c_i \cdot f(x_i) = \frac{10}{71} \cdot e^{-(x_i/71)}$$

$$P(X = 55) = 10 \cdot f(55) = \frac{10}{71} \cdot e^{-(55/71)} = 0,0649$$

$$P(X = 65) = 10 \cdot f(65) = \frac{10}{71} \cdot e^{-(65/71)} = 0,0564$$

$$P(X = 75) = 10 \cdot f(75) = \frac{10}{71} \cdot e^{-(75/71)} = 0,0490$$

$$P(X = 85) = 10 \cdot f(85) = \frac{10}{71} \cdot e^{-(85/71)} = 0,0425$$

$$P(X = 95) = 10 \cdot f(95) = \frac{10}{71} \cdot e^{-(95/71)} = 0,0370$$

$$P(X = 105) = 10 \cdot f(105) = \frac{10}{71} \cdot e^{-(105/71)} = 0,0321$$

Las frecuencias esperadas $e_i = n \cdot p_i = 100 \cdot p_i$ de las modalidades no pueden ser menor que 5, en caso contrario se agrupan dos o más modalidades contiguas en una sola hasta lograr que la nueva frecuencia sea mayor que cinco.

En este sentido, se agrupan las modalidades de forma que queden los intervalos [70 – 90] y [90 – 110]

Intervalos $L_{i-1} - L_i$	50 – 60	60 – 70	70 – 90	90 – 110
x_i	55	65	80	100
n_i	30	25	30	15
$p_i = P(x = x_i)$	0,0649	0,0564	0,0915	0,0691
$e_i = n \cdot p_i$	6,49	5,64	9,15	6,91

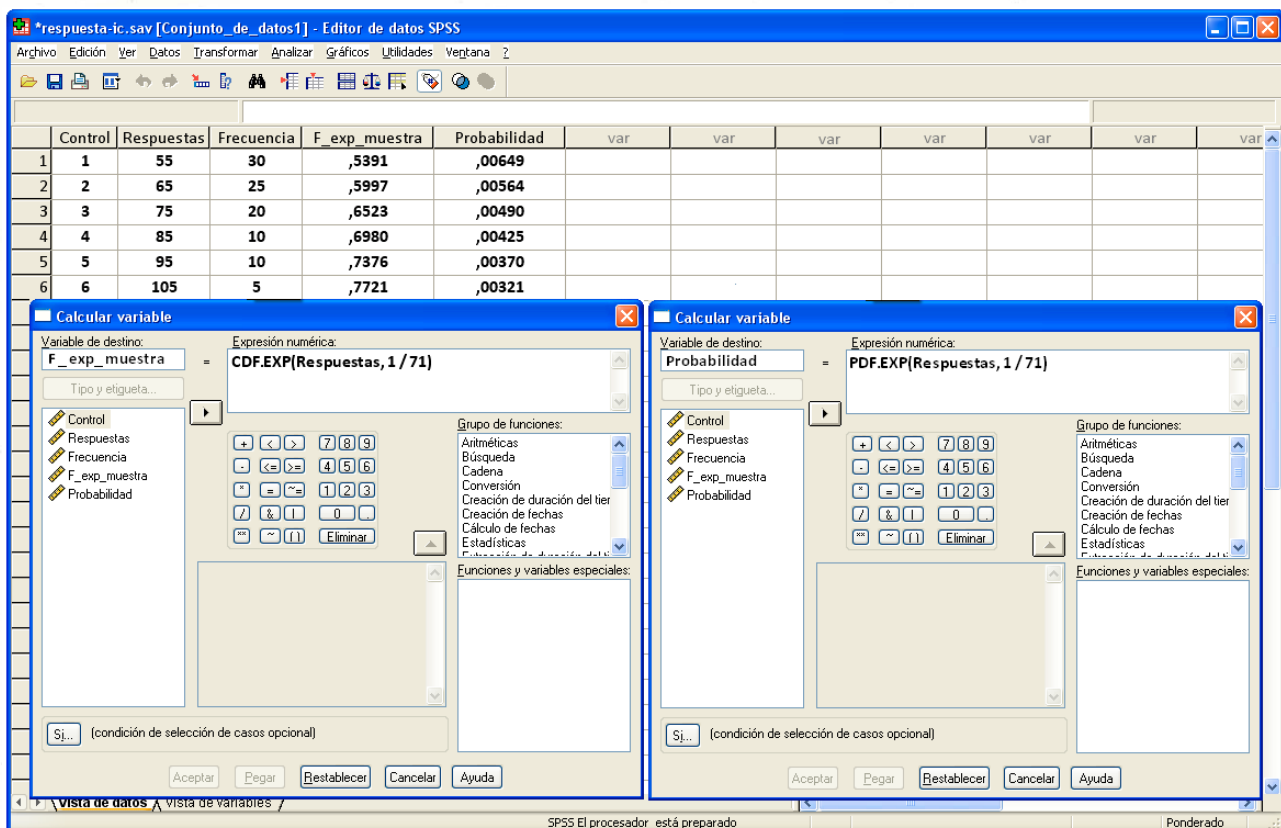
En este caso, los grados de libertad pasan a ser $(k - p - 1)$, con $k = 4$ modalidades, $p = 1$ parámetro λ que ha habido que calcular y -1 por ser modalidades excluyentes.

El valor que se obtiene para la Chi-cuadrado de las observaciones es:

$$\chi^2_2 = \sum_{i=1}^4 \frac{(O_i - e_i)^2}{e_i} = \frac{(30 - 6,49)^2}{6,49} + \frac{(25 - 5,64)^2}{5,64} + \frac{(30 - 9,15)^2}{9,15} + \frac{(15 - 6,91)^2}{6,91} = 208,6026$$

Estadístico teórico: $\chi^2_{\alpha, k-p-1} = \chi^2_{0,05, 2} = 5,991$

Como $\chi^2_2 = 208,603 > 5,991 = \chi^2_{0,05, 2}$ se rechaza la hipótesis nula, deduciendo que las respuestas inadecuadas del test de inteligencia no siguen una distribución exponencial de parámetro $\lambda = 1/71$, con una significación del 5%.



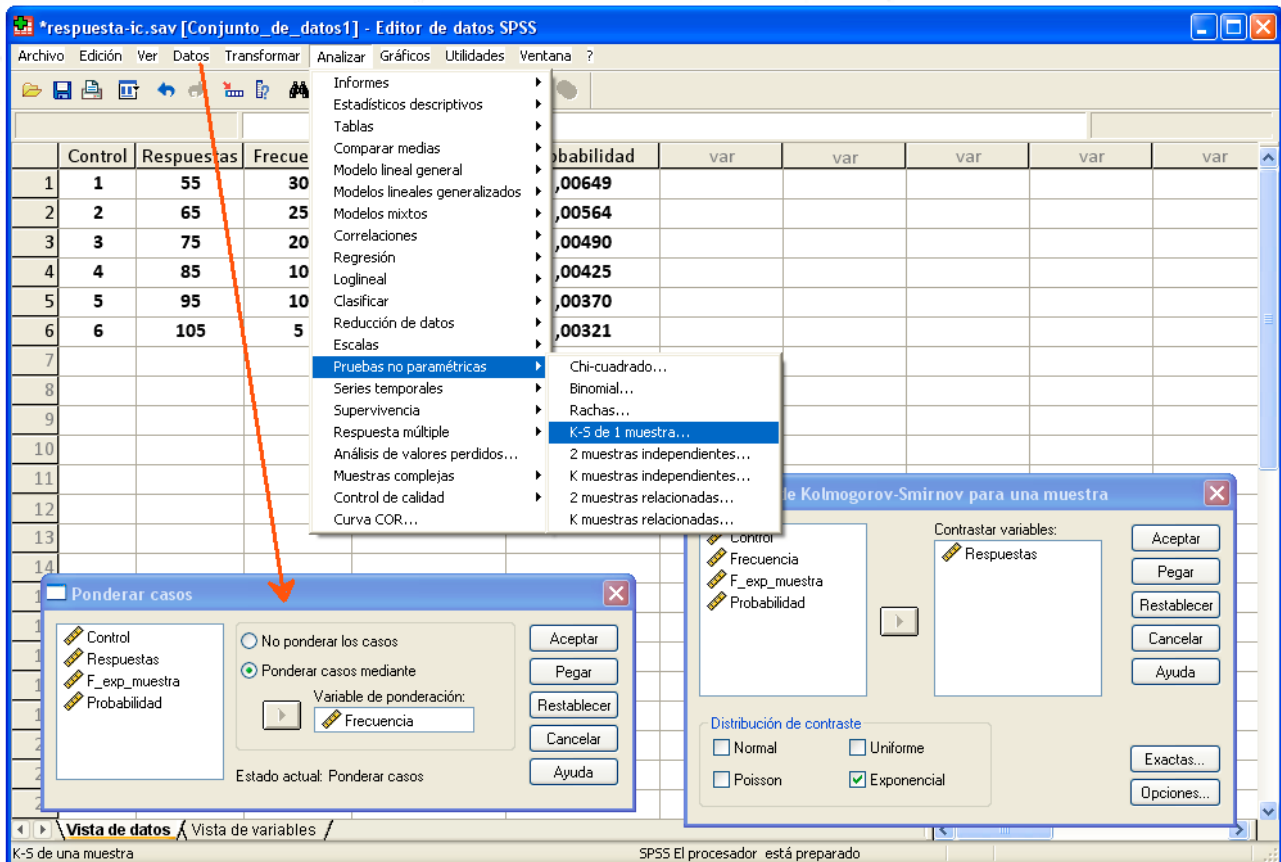
The screenshot shows the SPSS 'Editor de datos' window for a file named '*respuesta-ic.sav'. The data table has columns: Control, Respuestas, Frecuencia, F_exp_muestra, Probabilidad, and several empty 'var' columns. The data rows are as follows:

	Control	Respuestas	Frecuencia	F_exp_muestra	Probabilidad
1	1	55	30	,5391	,00649
2	2	65	25	,5997	,00564
3	3	75	20	,6523	,00490
4	4	85	10	,6980	,00425
5	5	95	10	,7376	,00370
6	6	105	5	,7721	,00321

Two 'Calcular variable' dialog boxes are open in the foreground:

- Left Dialog:** Variable de destino: F_exp_muestra. Expresión numérica: CDF.EXP(Respuestas, 1 / 71).
- Right Dialog:** Variable de destino: Probabilidad. Expresión numérica: PDF.EXP(Respuestas, 1 / 71).

Both dialogs show a list of variables on the left (Control, Respuestas, Frecuencia, F_exp_muestra, Probabilidad) and a list of functions on the right (Aritméticas, Búsqueda, Cadena, Conversión, etc.).



The screenshot shows the SPSS Editor de datos SPSS window. The 'Ponderar casos' dialog box is open, showing the 'Control' variable selected for weighting. The 'Pruebas no paramétricas' menu is also open, showing the 'K-S de 1 muestra...' option selected. The 'Pruebas no paramétricas' menu is open, showing the 'K-S de 1 muestra...' option selected. The 'Pruebas no paramétricas' menu is open, showing the 'K-S de 1 muestra...' option selected.

Prueba de Kolmogorov-Smirnov para una muestra

		Respuestas
N		100
Parámetro exponencial. ^{a,b}	Media	71,00
Diferencias más extremas	Absoluta	,539
	Positiva	,228
	Negativa	-,539
Z de Kolmogorov-Smirnov		5,391
Sig. asintót. (bilateral)		,000

a. La distribución de contraste es exponencial.

b. Se han calculado a partir de los datos.

$p_{\text{valor}} = 0,00 < 0,05 = \alpha \rightarrow$ Se rechaza la hipótesis nula. En consecuencia, ñas respuestas inadecuadas del test no siguen una distribución exponencial.

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

189. En un vuelo se adjudican los asientos a pasajeros que proceden de dos agencias de viajes (A y B) de la forma que se presenta:

Asiento	Agencia viajes
1	A
2	B
3	B
4	B
5	A
6	B
7	A
8	B
9	B
10	B
11	B
12	A
13	A
14	B
15	B
16	A
17	B
18	A
19	A
20	A
21	B
22	B
23	B
24	B
25	A
26	A
27	B
28	A
29	B
30	B

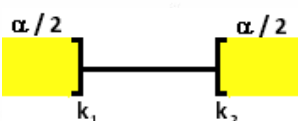
¿Se han asignado los asientos aleatoriamente, con un nivel de significación del 5%?

Solución:

Pra verificar si los asientos se han adjudicado aleatoriamente, se realiza el contraste bilateral:

H_0 : La muestra es aleatoria H_1 : La muestra no es aleatoria

Estadístico de prueba: R = "Número de rachas en la muestra"

La región crítica  $P(R \leq k_1 | H_0) = \alpha / 2$ $P(R \geq k_2 | H_0) = \alpha / 2$

Los valores críticos k_1 y k_2 se pueden obtener con las tablas de Wald-Wolfowitz

Se cuentan el número de rachas:

Agencia	A	BBB	A	B	A	BBBB	AA	BB	A	B	AAA	BBBB	AA	B	A	BB
Rachas	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

$R = 16$ rachas Asientos Agencia A = $n_1 = 12$ Asientos Agencia B = $n_2 = 18$

Valores críticos de la prueba de rachas nivel de significación del 5%

$$P[R \leq k_1 / H_0] = 0,25 \quad P[R \geq k_2 / H_0] = 0,25$$

n ₁	n ₂																			
	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
12 K ₂					13	14	16	16	17	18	19	19	20	20	21	21	21	22	22	
K ₁	2	2	3	4	4	5	6	6	7	7	7	8	8	8	9	9	9	10	10	

$k_1 = 9$ y $k_2 = 21 \rightarrow k_1 = 9 < R = 16 < 21 = k_2$ con lo que el estadístico R se encuentra en la región de aceptación y, en consecuencia, la muestra puede considerarse aleatoria.

▪ Siendo $n_1 = 12 > 10$ y $n_2 = 18 > 10$ se puede aplicar la distribución asintótica normal de Wald y Wolfowitz, denotando por $\gamma = \frac{n_1}{n}$ la proporción de observaciones de una categoría en la muestra de tamaño n se verifica que: $R \xrightarrow{d} N[2\gamma(1-\gamma)n, 2\gamma(1-\gamma)\sqrt{n}]$

En esta línea, $\gamma = \frac{n_1}{n} = \frac{12}{30} = 0,4$, quedando la distribución normal:

$$R \sim N[2 \cdot 0,4 \cdot (1 - 0,4) \cdot 30, 2 \cdot 0,4 \cdot (1 - 0,4) \sqrt{30}] \equiv N(14,4, 2,629)$$

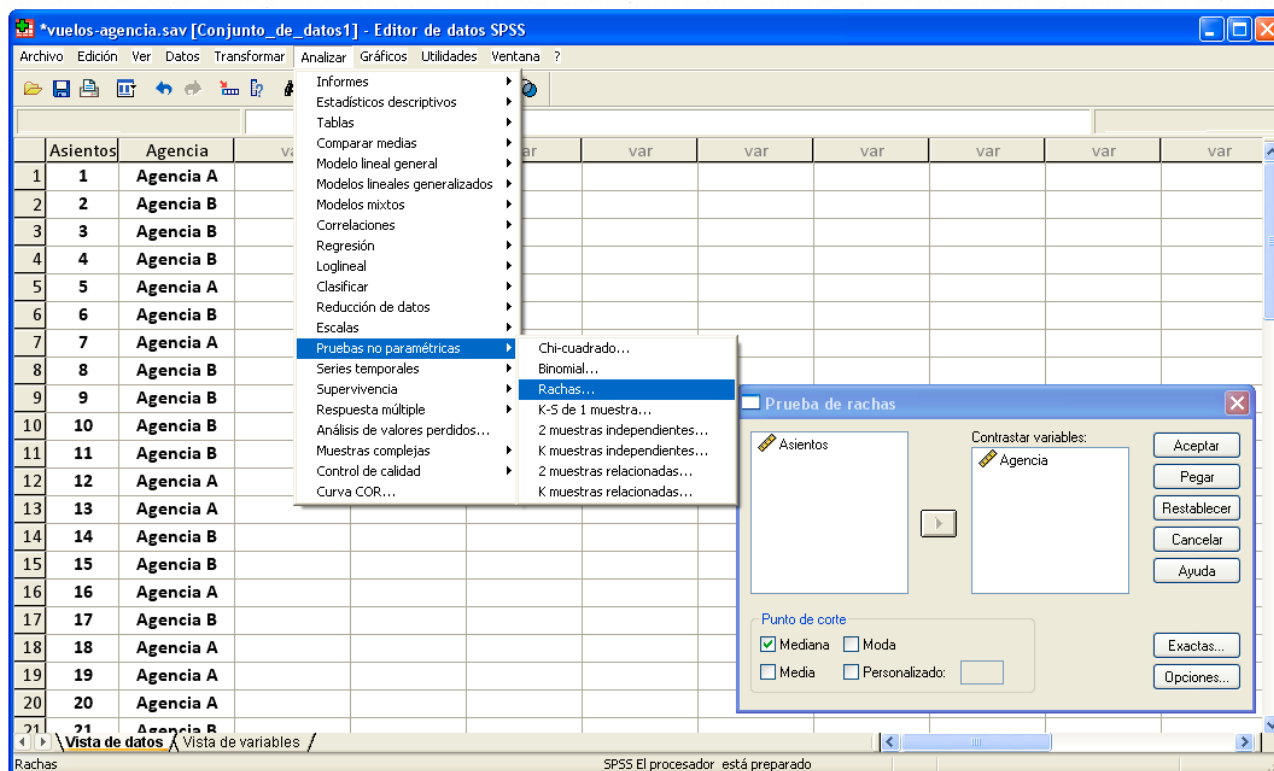
Los valores críticos k_1 y k_2 se determinan con el valor de significación $\alpha = 0,05$:

$$\begin{aligned} \alpha &= P[ET1] = P[\text{Rechazar } H_0 \mid H_0 \text{ Cierta}] = P[R \mid N(14,4, 2,629)] = \\ &= P[(R \leq k_1) \cup (R \geq k_2) \mid N(14,4, 2,629)] = P\left[\frac{R - 14,4}{2,629} \leq \frac{k_1 - 14,4}{2,629}\right] + P\left[\frac{R - 14,4}{2,629} \geq \frac{k_2 - 14,4}{2,629}\right] = \\ &= P\left[z \leq \frac{k_1 - 14,4}{2,629}\right] + P\left[z \geq \frac{k_2 - 14,4}{2,629}\right] = 0,025 + 0,025 \end{aligned}$$

$$P\left[z \leq \frac{k_1 - 14,4}{2,629}\right] = P\left[z \geq \frac{-k_1 + 14,4}{2,629}\right] = 0,025 \rightarrow \frac{-k_1 + 14,4}{2,629} = 1,96 \rightarrow k_1 = 9,247$$

$$P\left[z \geq \frac{k_2 - 14,4}{2,629}\right] = 0,025 \rightarrow \frac{k_2 - 14,4}{2,629} = 1,96 \rightarrow k_2 = 19,553$$

El número de rachas $R = 16$ se encuentra en la región de aceptación: $9,247 < R = 16 < 19,553$ por lo que se acepta la hipótesis nula de aleatoriedad de los asientos del vuelo con una significación del 5%.



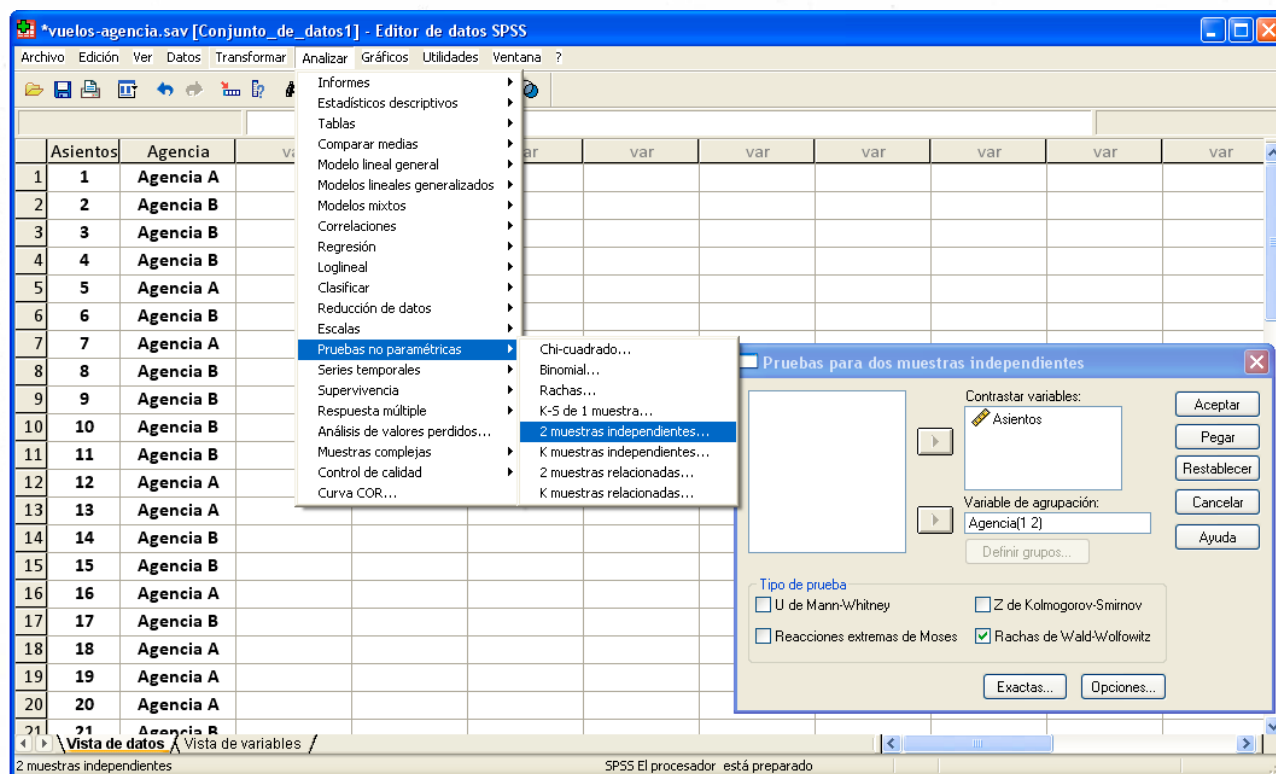
Prueba de rachas

	Agencia
Valor de prueba ^a	2
Casos < Valor de prueba	12
Casos ≥ Valor de prueba	18
Casos en total	30
Número de rachas	16
Z	,039
Sig. asintót. (bilateral)	,969

$$p_{\text{valor}} (\text{Sig. asintótica bilateral}) = 0,969 > 0,05 = \alpha$$

Por tanto, se admite la hipótesis nula H_0 de aleatoriedad de la muestra.

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....



*vuelos-agencia.sav [Conjunto_de_datos1] - Editor de datos SPSS

Archivo Edición Ver Datos Transformar Analizar Gráficos Utilidades Ventana ?

Informes
Estadísticos descriptivos
Tablas
Comparar medias
Modelo lineal general
Modelos lineales generalizados
Modelos mixtos
Correlaciones
Regresión
Loglineal
Clasificar
Reducción de datos
Escalas

Pruebas no paramétricas
Series temporales
Supervivencia
Respuesta múltiple
Análisis de valores perdidos...
Muestras complejas
Control de calidad
Curva COR...

Chi-cuadrado...
Binomial...
Rachas...
K-S de 1 muestra...
2 muestras independientes...
K muestras independientes...
2 muestras relacionadas...
K muestras relacionadas...

Pruebas para dos muestras independientes

Contrastar variables:
Asientos

Variable de agrupación:
Agencia(1 2)

Definir grupos...

Tipo de prueba
☐ U de Mann-Whitney
☐ Z de Kolmogorov-Smirnov
☐ Reacciones extremas de Moses
☒ Rachas de Wald-Wolfowitz

Exactas... Opciones...

Aceptar Pegar Restablecer Cancelar Ayuda

Vista de datos Vista de variables /

2 muestras independientes SPSS El procesador está preparado

Prueba de Wald-Wolfowitz

Frecuencias

	Agencia	N
Asientos	Agencia A	12
	Agencia B	18
	Total	30

Estadísticos de contraste^{b,c}

		Número de rachas	Z	Sig. exacta (unilateral)
Asientos	Número exacto de rachas	16 ^a	,426	,662

a. No se han hallado empates intra-grupo.

b. Prueba de Wald-Wolfowitz

c. Variable de agrupación: Agencia

El p_{valor} (Sig. exacta unilateral) = $0,662 > 0,05 = \alpha \rightarrow$ Se admite la hipótesis nula H_0 de aleatoriedad de la muestra.

190. La asociación de mutuas de accidentes analiza el número de bajas ocurridas en el último mes debidas a accidentes laborales, en empresas constructoras que realizan su actividad en la capital de provincia. Los datos recogidos aparecen en la tabla siguiente:

Intervalos $L_{i-1} - L_i$	0 - 2	2 - 4	4 - 6	6 - 8	8 - 12
Frecuencia n_i	22	35	32	3	3

Con un nivel de significación del 5%, ¿se ajusta la muestra a una distribución normal?. Utilizar la prueba de la Chi-cuadrado y la prueba de Lilliefors.

Solución:

✓ **Prueba de la Chi-cuadrado**

Contraste bilateral: H_0 : El número de bajas sigue una distribución normal
 H_1 : El número de bajas no sigue una distribución normal

El método de aplicación de la prueba de ajuste de normalidad de la distribución de frecuencias por intervalos es como sigue en la tabla siguiente:

$L_{i-1} - L_i$	x_i	n_i	N_i	$x_i \cdot n_i$	$x_i^2 \cdot n_i$	p_i	$e_i = p_i \cdot n$
0 - 2	1	22	22	22	22	0,1802	17,11
2 - 4	3	35	57	105	315	0,3694	35,09
4 - 6	5	32	89	160	800	0,3017	28,66
6 - 8	7	3	92	21	147	0,0980	9,31
8 - 12	10	3	95	30	300	0,0132	1,25
		95		338	1584		

Se calculan las estimaciones de μ y σ , empleando los estimadores que se obtienen por el método de máxima verosimilitud: $\hat{\mu} = \bar{x} = 3,56$, $\hat{\sigma} = \sigma_x = 2$

$$\hat{\mu} = \alpha_1 = \bar{x} = \frac{1}{n} \sum_{i=1}^5 x_i \cdot n_i = \frac{338}{95} = 3,56 \quad \alpha_2 = \frac{1}{n} \sum_{i=1}^5 x_i^2 \cdot n_i = \frac{1584}{95} = 16,674$$

$$\hat{\sigma}_x^2 = \alpha_2 - \alpha_1^2 = 16,674 - 3,56^2 = 4 \quad \hat{\sigma}_x = \sqrt{4} = 2$$

$$\hat{s}_x^2 = \frac{n \cdot \hat{\sigma}_x^2}{(n-1)} = \frac{95 \cdot 4}{94} = 4,043 \quad \hat{s}_x = \sqrt{4,043} = 2,01$$

Mediante la tabla normal se hallan las probabilidades de cada uno de los intervalos, utilizando una distribución normal $N(3,56, 2)$

$$p_1 = P(0 \leq x < 2) = P\left[\frac{0-3,56}{2} \leq \frac{x-3,56}{2} < \frac{2-3,56}{2}\right] = P(-1,78 \leq z < -0,78) =$$

$$= P(0,78 \leq z < 1,78) = P(z \geq 0,78) - P(z > 1,78) = 0,2177 - 0,0375 = 0,1802$$

$$p_2 = P(2 \leq x < 4) = P\left[\frac{2-3,56}{2} \leq \frac{x-3,56}{2} < \frac{4-3,56}{2}\right] = P(-0,78 \leq z < 0,22) =$$

$$= P(z \geq -0,78) - P(z > 0,22) = P(z \leq 0,78) - P(z > 0,22) = 1 - P(z \geq 0,78) - P(z > 0,22) =$$

$$= 1 - 0,2177 - 0,4129 = 0,3694$$

$$p_3 = P(4 \leq x < 6) = P\left[\frac{4-3,56}{2} \leq \frac{x-3,56}{2} < \frac{6-3,56}{2}\right] = P(0,22 \leq z < 1,22) =$$

$$= P(z \geq 0,22) - P(z > 1,22) = 0,4129 - 0,1112 = 0,3017$$

$$p_4 = P(6 \leq x < 8) = P\left[\frac{6-3,56}{2} \leq \frac{x-3,56}{2} < \frac{8-3,56}{2}\right] = P(1,22 \leq z < 2,22) =$$

$$= P(z \geq 1,22) - P(z > 2,22) = 0,1112 - 0,0132 = 0,0980$$

$$p_5 = P(8 \leq x < 12) = P\left[\frac{8-3,56}{2} \leq \frac{x-3,56}{2} < \frac{12-3,56}{2}\right] = P(2,22 \leq z < 4,22) =$$

$$= P(z \geq 2,22) - P(z > 4,22) = 0,0132 - 0,000013 = 0,0132$$

Como la frecuencia esperada $e_5 = p_5 \cdot n = 0,0132 \cdot 95 = 1,254 < 5$ es necesario agrupar dos o más modalidades contiguas en una sola hasta lograr que la nueva frecuencia esperada sea mayor que cinco. En este sentido, se agrupan los intervalos 6 – 12 que pasa tener una frecuencia esperada $e_4 = 9,31 + 1,25 = 10,56$

La tabla queda, Excel: $p_i = F(b) - F(a) = \text{DISTR.NORM}(L_i; \hat{\mu}; \hat{\sigma}; 1) - \text{DISTR.NORM}(L_{i-1}; \hat{\mu}; \hat{\sigma}; 1)$

$L_{i-1} - L_i$	x_i	n_i	p_i	$e_i = p_i \cdot n$	n_i^2	n_i^2 / e_i
0 – 2	1	22	0,1802	17,11	484	28,279
2 – 4	3	35	0,3694	35,09	1225	34,910
4 – 6	5	32	0,3017	28,66	1024	35,727
6 – 12	9	6	0,1112	10,56	36	3,407
		95				102,324

Los grados de libertad son $(k - p - 1)$ siendo $k = 4$ el número de modalidades diferentes, $p = 2$ ha sido necesario calcular 2 parámetros $(\hat{\mu}, \hat{\sigma})$ y (-1) por ser las modalidades excluyentes.

$$\text{Estadístico contraste: } \chi_{4-2-1}^2 = \sum_{i=1}^4 \frac{(n_i - e_i)^2}{e_i} = \sum_{i=1}^4 \frac{n_i^2}{e_i} - n = 102,324 - 95 = 7,324$$

Siendo $\chi_1^2 = 7,324 > 3,841 = \chi_{0,05,1}^2 \rightarrow$ Se rechaza la hipótesis nula, concluyendo que el número de bajas ocurridas en el último mes debidas a accidentes laborales no sigue una distribución normal, con una significación del 5%.

✓ Prueba de Lilliefors

Contraste bilateral: H_0 : La muestra procede de una distribución normal
 H_1 : La muestra no procede de una distribución normal

Estadístico contraste: $D_n = \max |F_n(z) - F_0(z)| = \max(a_i, b_i)$

Se acepta H_0 si $D_n = \max |F_n(z) - F_0(z)| = \max(a_i, b_i) < D_{\alpha, n}$ (valor crítico tabla de Lilliefors)

Para determinar el estadístico de contraste se construye una tabla donde se ordenan los datos, tipificando cada valor, se calcula $F_0(z_i)$, $F_n(z_i)$ y se crean las columnas $a_i = |F_0(z_i) - F_n(z_i)|$ y

$$b_i = |F_0(z_i) - F_n(z_{i-1})|$$

Valores tipificados $z_i = \frac{x_i - \hat{\mu}}{\hat{s}_x} = \frac{x_i - 3,56}{2,01}$

$F_0(z_i) \equiv$ Función de distribución de una $N(0, 1)$: DISTR.NORM.ESTAND(z_i)

$F_n(z_i) \equiv$ Función de distribución empírica de la muestra tipificada: $F_n(z_i) = \frac{n^\circ \text{ observaciones}}{n} = \frac{N(z)}{n}$

x_i	z_i	$F_0(z_i)$	$F_n(z_i)$	$a_i = F_0(z_i) - F_n(z_i) $	$b_i = F_0(z_i) - F_n(z_{i-1}) $
1	-1,27	0,1021	0,2316	0,1295	0,1020
3	-0,28	0,3897	0,6000	0,2103	0,1581
5	0,72	0,7642	0,9368	0,1726	0,1642
7	1,71	0,9564	0,9684	0,0121	0,0196
10	3,20	0,9993	1	0,0007	0,0309

$$F_0(x_1) = P(X \leq 1) = P\left[\frac{X - \hat{\mu}}{\hat{s}_x} \leq \frac{1 - 3,56}{2,01}\right] = P(z \leq -1,27) = P(z \geq 1,27) = 0,1021$$

$$F_0(x_2) = P(X \leq 3) = P\left[\frac{X - \hat{\mu}}{\hat{s}_x} \leq \frac{3 - 3,56}{2,01}\right] = P(z \leq -0,28) = P(z \geq 0,28) = 0,3897$$

$$F_0(x_3) = P(X \leq 5) = P\left[\frac{X - \hat{\mu}}{\hat{s}_x} \leq \frac{5 - 3,56}{2,01}\right] = P(z \leq 0,72) = 1 - P(z \geq 0,72) = 1 - 0,2358 = 0,7642$$

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

$$F_0(x_4) = P(X \leq 7) = P\left[\frac{X - \hat{\mu}}{\hat{s}_x} \leq \frac{7 - 3,56}{2,01}\right] = P(z \leq 1,71) = 1 - P(z \geq 1,71) = 1 - 0,0436 = 0,9564$$

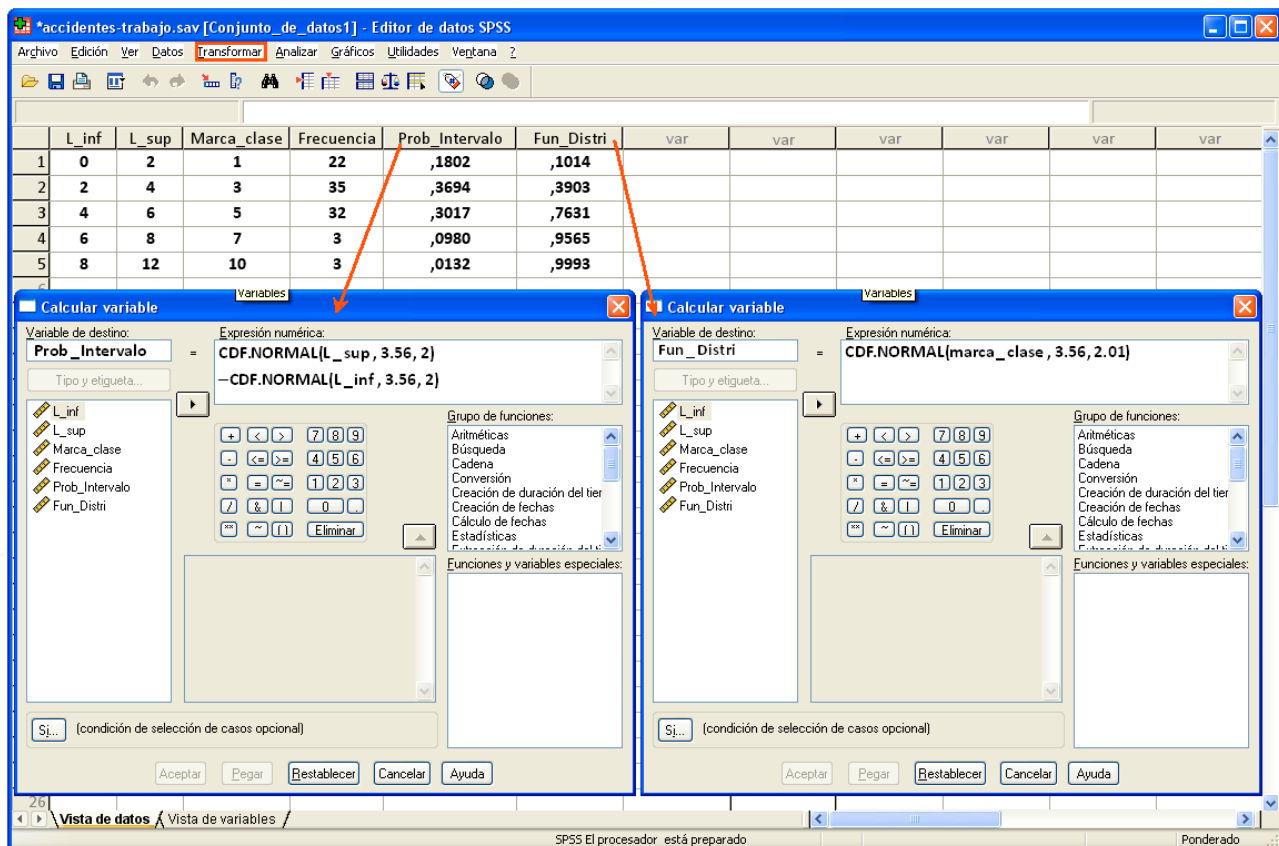
$$F_0(x_4) = P(X \leq 10) = P\left[\frac{X - \hat{\mu}}{\hat{s}_x} \leq \frac{10 - 3,56}{2,01}\right] = P(z \leq 3,20) = 1 - P(z \geq 3,20) = 1 - 0,0007 = 0,9993$$

$$D_n = \max(a_i, b_i) = 0,2103 > 0,09 = \frac{0,886}{\sqrt{95}} = D_{0,05, 95} \text{ (valor crítico tabla de Lilliefors)}$$

Por lo que se rechaza la hipótesis nula de normalidad para la muestra, con un nivel de significación del 5%

Valores críticos del test de Lilliefors de normalidad

	Nivel de significación α				
	0,20	0,15	0,10	0,05	0,01
Aproximación $n > 30$	$\frac{0,736}{\sqrt{n}}$	$\frac{0,768}{\sqrt{n}}$	$\frac{0,805}{\sqrt{n}}$	$\frac{0,886}{\sqrt{n}}$	$\frac{1,031}{\sqrt{n}}$



Variables

	L_inf	L_sup	Marca_clase	Frecuencia	Prob_Intervalo	Fun_Distri	var	var	var	var	var	var
1	0	2	1	22	,1802	,1014						
2	2	4	3	35	,3694	,3903						
3	4	6	5	32	,3017	,7631						
4	6	8	7	3	,0980	,9565						
5	8	12	10	3	,0132	,9993						

Calcular variable

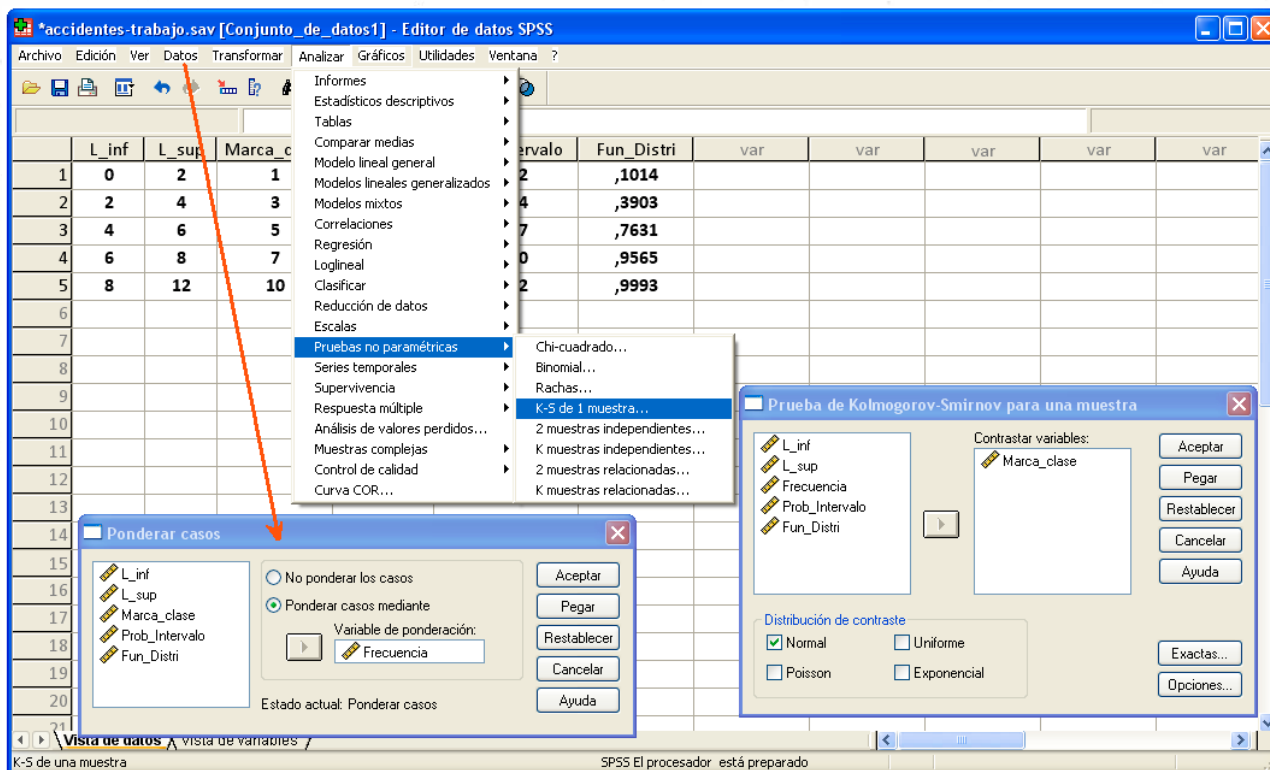
Variable de destino: Prob_Intervalo

Expresión numérica: CDF.NORMAL(L_sup, 3.56, 2) - CDF.NORMAL(L_inf, 3.56, 2)

Calcular variable

Variable de destino: Fun_Distri

Expresión numérica: CDF.NORMAL(marca_clase, 3.56, 2.01)



Prueba de Kolmogorov-Smirnov para una muestra

		Marca_clase
N		95
Parámetros normales^{a,b}	Media	3,56
	Desviación típica	2,014
	Diferencias más extremas Absoluta	,209
	Positiva	,209
Z de Kolmogorov-Smirnov	Negativa	-,163
		2,038
Sig. asintót. (bilateral)		,000

a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

$p_valor(\text{Sig. asintótica bilateral}) = 0 < 0,05 = \alpha \rightarrow$ Se rechaza la hipótesis nula de normalidad, con un nivel de significación del 5%

191. Para tratar de evaluar las diferentes capacidades por sexo, se somete a un examen lógico a una muestra de hombres y mujeres, obteniendo las siguientes puntuaciones:

Hombres	52,48	72,98	73,8	22,14	36,9	82	73,8	40,18	24,6	57,4	123
Mujeres	65,6	48,38	82	123	143,5	73,8	24,6	26,24	49,2	65,6	131,2

Con una significación del 5%, ¿existen diferencias significativas entre hombres y mujeres en cuanto a la dispersión de las puntuaciones?. Utilizar el contraste de Siegel-Tukey.

Solución:

Se establece el contraste bilateral: $H_0: \sigma_H = \sigma_M$ $H_1: \sigma_H \neq \sigma_M$

Estadístico contraste: $R_n = \sum_{i=1}^n a_i \cdot c_i$ donde, $c_i = \begin{cases} 1 & \text{valor de H} \\ 0 & \text{valor de M} \end{cases}$

Para calcular el estadístico de contraste R_n es necesario calcular los rangos a_i , ordenando las puntuaciones muestrales conjuntas de menor a mayor, asignando primero al coeficiente instrumental $c_i = 1$ para las puntuaciones de hombres y después $c_i = 0$ para las puntuaciones de mujeres.

$a_i \equiv$ Coeficientes obtenidos al asignar los rangos según el método de Siegel-Tukey: La clasificación se realiza alternando extremos (el rango 1 es el más bajo, 2 y 3 son los dos más altos, 4 y 5 son los dos más bajos, y así sucesivamente).

Cuando el número total de observaciones es un número impar, se ignorará la observación central.

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

	a	a_i	c_i	$a_i \cdot c_i$
H	22,14	1	1	1
H	24,6	4	1	4
M	24,6	5	0	0
M	26,24	8	0	0
H	36,9	9	1	9
H	40,18	12	1	12
M	48,38	13	0	0
M	49,2	16	0	0
H	52,48	17	1	17
H	57,4	20	1	20
M	65,6	21	0	0
M	65,6	22	0	0
H	72,98	19	1	19
H	73,8	18	1	18
H	73,8	15	1	15
M	73,8	14	0	0
H	82	11	1	11
M	82	10	0	0
H	123	7	1	7
M	123	6	0	0
M	131,2	3	0	0
M	143,5	2	0	0
				133

$$n_1 = 11 \quad n_2 = 11 \quad n = n_1 + n_2 = 22 \quad R_{22} = \sum_{i=1}^{22} a_i \cdot c_i = 133$$

Como las muestras son grandes ($n_1 = 11 > 10$ y $n_2 = 11 > 10$) se emplea la aproximación asintótica a una distribución normal $N(0, 1)$:

$$z_p = \frac{|R_n - E[R_n]|}{\sqrt{V[R_n]}} \quad \text{donde, } E[R_n] = \frac{n_1 \cdot (n+1)}{2} \quad V[R_n] = \frac{n_1 \cdot n_2 \cdot (n+1)}{12}$$

Se acepta H_0 si $z_p \leq z_{\alpha/2}$

$$E[R_{22}] = \frac{11 \cdot (22+1)}{2} = 126,5 \quad V[R_{22}] = \frac{n_1 \cdot n_2 \cdot (n+1)}{12} = \frac{11 \cdot 11 \cdot (23)}{12} = 231,92$$

$$z_p = \frac{|R_{22} - E[R_{22}]|}{\sqrt{V[R_{22}]}} = \frac{|133 - 126,5|}{\sqrt{231,92}} = 0,4268 < 1,96 = z_{\alpha/2} = z_{0,025}$$

En consecuencia, se acepta la hipótesis nula, concluyendo que no existen diferencias significativas en cuanto a la dispersión de las puntuaciones, con un nivel de significación del 5%.

192. En un estudio sobre las sanciones que recaen sobre empresas por incumplimiento de normas de seguridad en el trabajo se encontró que las sanciones medias eran las mismas en dos comunidades autónomas (X e Y). En cada comunidad se eligieron al azar expedientes sancionadores, presentando en la tabla adjunta los importes (expresados en cientos de euros).

X	78,4	119	32,2	89,6	105	128,8	63	53,2	93,8	56	102,2	18,2	72,8
Y	84	81,2	70	44,8	96,6	116,2	50,4	67,2	77	123,2	98	28	

Con un 10% de significación, ¿Podría afirmarse que existen diferencias significativas en cuanto a la dispersión de las sanciones de seguridad en el trabajo en las dos comunidades?

Utilizar el contraste de Siegel-Tukey.

Solución:

Se establece el contraste bilateral: $H_0: \sigma_X = \sigma_Y$ $H_1: \sigma_X \neq \sigma_Y$

Estadístico contraste: $R_n = \sum_{i=1}^n a_i \cdot c_i$ donde, $c_i = \begin{cases} 1 & \text{valor de X} \\ 0 & \text{valor de Y} \end{cases}$

Para calcular el estadístico de contraste R_n es necesario calcular los rangos a_i , ordenando las puntuaciones muestrales conjuntas de menor a mayor, asignando primero al coeficiente instrumental $c_i = 1$ para las sanciones de la comunidad X y después $c_i = 0$ para las sanciones de la comunidad Y.

$a_i \equiv$ Coeficientes obtenidos al asignar los rangos según el método de Siegel-Tukey: La clasificación se realiza alternando extremos (el rango 1 es el más bajo, 2 y 3 son los dos más altos, 4 y 5 son los dos más bajos, y así sucesivamente).

En este caso el número total de observaciones $n = n_X + n_Y = 13 + 12 = 25$ es un número impar, se ignora la obseración central.

Como las muestras son grandes ($n_X = 13 > 10$ y $n_Y = 12 > 10$) se emplea la aproximación asintótica a una distribución normal $N(0, 1)$:

$$z_p = \frac{|R_n - E[R_n]|}{\sqrt{V[R_n]}} \quad \text{donde, } E[R_n] = \frac{n_X \cdot (n+1)}{2} \quad V[R_n] = \frac{n_X \cdot n_Y \cdot (n+1)}{12}$$

Se acepta H_0 sí $z_p \leq z_{\alpha/2}$

Comunidades	a	a_i	c_i	$a_i \cdot c_i$
X	18,2	1	1	1
Y	28	4	0	0
X	32,2	5	1	5
Y	44,8	8	0	0
Y	50,4	9	0	0
X	53,2	12	1	12
X	56	13	1	13
X	63	16	1	16
Y	67,2	17	0	0
Y	70	20	0	0
X	72,8	21	1	21
Y	77	24	0	0
X	78,4	-----	1	0
Y	81,2	23	0	0
Y	84	22	0	0
X	89,6	19	1	19
X	93,8	18	1	18
Y	96,6	15	0	0
Y	98	14	0	0
X	102,2	11	1	11
X	105	10	1	10
Y	116,2	7	0	0
X	119	6	1	6
Y	123,2	3	0	0
X	128,8	2	1	2
				134

$$R_{25} = \sum_{i=1}^{25} a_i \cdot c_i = 134$$

$$E[R_{25}] = \frac{n_x \cdot (n+1)}{2} = \frac{13 \cdot (25+1)}{2} = 169 \quad V[R_{25}] = \frac{n_x \cdot n_y \cdot (n+1)}{12} = \frac{13 \cdot 12 \cdot (25+1)}{12} = 338$$

$$z_p = \frac{|R_{25} - E[R_{25}]|}{\sqrt{V[R_{25}]}} = \frac{|134 - 169|}{\sqrt{338}} = 1,9037 > 1,645 = z_{\alpha/2} = z_{0,05}$$

En consecuencia, se rechaza la hipótesis nula, concluyendo que existen diferencias significativas en cuanto a la dispersión de las sanciones de seguridad en el trabajo en las dos comunidades, con un nivel de significación del 10%.

193. Se tienen cuatro razas de gatos. Se toman tres muestras elegidas al azar de forma que en cada una de ellas hay un gato de cada raza. Se dispone de tres tipos de vitaminas distintas y se administra una de ellas solamente a cada muestra. Se elige como variable de respuesta el peso ganado por cada uno de los gatos, información que se recoge en la tabla:

Razas	Tratamientos		
	Vitamina A	Vitamina B	Vitamina C
1	80	70	82
2	60	32	35
3	90	82	85
4	95	89	78

Con una significación del 5%, se quiere saber:

- 1) Si la ganancia media de pesos es equivalente para las tres vitaminas
- 2) Si el peso medio ganado es significativamente distinto de una raza a otra

Solución:

1) Se supone que se cumplen las hipótesis de normalidad, homocedasticidad y elección al azar de las muestras. Se aplica el modelo ANOVA con dos factores de variación sin interacción (cuando el efecto de un factor sobre el peso depende de cual sea el nivel del otro factor):

Factor A: Razas

Factor B: Vitaminas

El contraste de hipótesis:

H_0 : Las tres vitaminas proporcionan una ganancia en peso medio equivalente

H_1 : Las ganancias en peso medio proporcionadas por las distintas vitaminas no son iguales

La descomposición de la varianza total sin interacción: $SCT = SCF_A + SCF_B + SCR$

Razas	Tratamientos			
	Vitamina A	Vitamina B	Vitamina C	TOTAL
1	80	70	82	$Y_{1.} = 232$
2	60	32	35	$Y_{2.} = 127$
3	90	82	85	$Y_{3.} = 257$
4	95	89	78	$Y_{4.} = 262$
TOTAL	$Y_{.1} = 325$	$Y_{.2} = 273$	$Y_{.3} = 280$	$Y_{..} = 878$

$n = 1$, $G = 4$, $J = 3$, $N = G \cdot J = 12$

$$SCT = \sum_{g=1}^4 \sum_{j=1}^3 Y_{gj}^2 - \frac{Y_{..}^2}{N} = 68952 - \frac{878^2}{12} = 68952 - 64240,333 = 4711,667$$

$$SCF_A = \frac{\sum_{g=1}^G Y_{i.}^2}{n \cdot J} - \frac{Y_{..}^2}{N} = \frac{\sum_{g=1}^4 Y_{i.}^2}{3} - \frac{Y_{..}^2}{12} = \left[\frac{232^2}{3} + \frac{127^2}{3} + \frac{257^2}{3} + \frac{262^2}{3} \right] - \frac{878^2}{12} = 3975$$

$$SCF_B = \frac{\sum_{j=1}^J Y_{.j}^2}{n \cdot G} - \frac{Y_{..}^2}{N} = \frac{\sum_{j=1}^3 Y_{.j}^2}{4} - \frac{Y_{..}^2}{12} = \left[\frac{325^2}{4} + \frac{273^2}{4} + \frac{280^2}{4} \right] - \frac{878^2}{12} = 64638,5 - 64240,333 = 398,167$$

$$SCT = SCF_A + SCF_B + SCR \rightarrow SCR = SCT - SCF_A - SCF_B = 4711,667 - 3975 - 398,167 = 338,5$$

$$MCF_A = \frac{SCF_A}{G-1} = \frac{3975}{3} = 1325 \quad MCR = \frac{SCR}{(G-1)(J-1)} = \frac{338,5}{6} = 56,4167 \quad F_A = \frac{MCF_A}{MCR} = 23,486$$

$$MCF_B = \frac{SCF_B}{J-1} = \frac{398,167}{2} = 199,083 \quad F_B = \frac{MCF_B}{MCR} = 3,529 \quad MCT = \frac{SCT}{N-1} = \frac{338,5}{11} = 30,773$$

Fuente variación	Suma de cuadrados	Grados libertad	Media cuadrática	Estadístico
Intergrupos	$SCF_A = 3975$	$G - 1 = 3$	$MCF_A = \frac{SCF_A}{G-1} = 1325$	$F_A = \frac{MCF_A}{MCR} = 23,486$
Intersujetos	$SCF_B = 398,167$	$J - 1 = 2$	$MCF_B = \frac{SCF_B}{J-1} = 199,083$	$F_B = \frac{MCF_B}{MCR} = 3,529$
Error	$SCR = 338,5$	$(G-1)(J-1) = 6$	$MCR = \frac{SCR}{(G-1)(J-1)} = 56,417$	
Total	$SCT = 4711,667$	$N - 1 = 11$	$MCT = \frac{SCT}{N-1} = 30,773$	

El estadístico observado $F_B = 3,529$ contrasta si las vitaminas proporcionan o no ganancia de peso medio en los gatos. El estadístico teórico $F_{0,05; (J-1), (G-1)(J-1)} = F_{0,05; 2, 6} = 5,1433$

Siendo $F_B = 3,529 < 5,1433 = F_{0,05; 2, 6} \rightarrow$ Se acepta la hipótesis nula de que no hay diferencia de ganancia de pesos debida a las vitaminas.

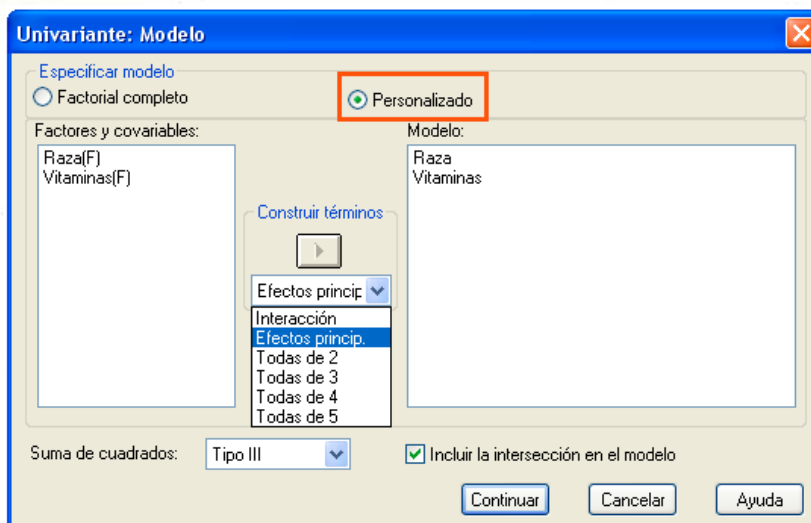
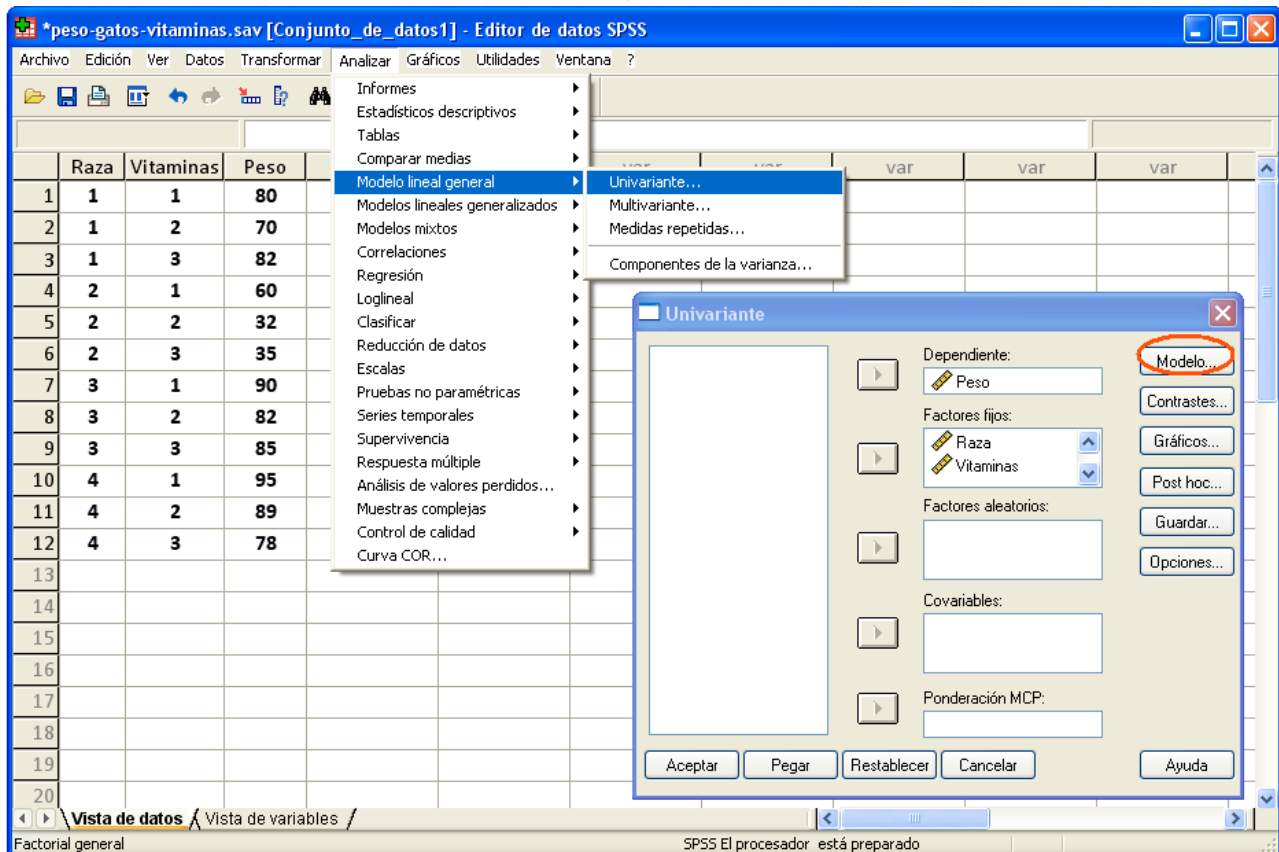
2) Para contrastar si el peso medio ganado es significativamente distinto por razas, se establece el contraste:

H_0 : El peso medio ganado es independiente de las razas

H_1 : El peso medio ganado depende de las razas

El estadístico observado $F_A = 23,486$ contrasta si las razas proporcionan o no ganancia de peso medio en los gatos. El estadístico teórico $F_{0,05; (G-1), (G-1)(J-1)} = F_{0,05; 3, 6} = 4,7571$

Como $F_A = 23,486 > 4,7571 = F_{0,05; 3, 6} \rightarrow$ Se rechaza la hipótesis nula de que no hay diferencia de ganancia de pesos debida a las razas, concluyendo que el peso medio ganado depende de la raza de los gatos.



Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

Pruebas de los efectos inter-sujetos

Variable dependiente: Peso

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	4373,167 ^a	5	874,633	15,503	,002
Intersección	64240,333	1	64240,333	1138,68	,000
Raza	3975,000	3	1325,000	23,486	,001
Vitaminas	398,167	2	199,083	3,529	,097
Error	338,500	6	56,417		
Total	68952,000	12			
Total corregida	4711,667	11			

a. R cuadrado = ,928 (R cuadrado corregida = ,868)

Al contrastar la hipótesis nula H_0 : El peso medio ganado es independiente de las razas.

El p_valor (Significación) = $0,001 < 0,05 = \alpha \rightarrow$ Se rechaza la hipótesis nula, concluyendo que el peso medio ganado depende de la raza de los gatos.

♦ Al contrastar la hipótesis nula H_0 : Las vitaminas proporcionan una ganancia en peso medio equivalente, se observa un p_valor (Significación) = $0,097 > 0,05 = \alpha \rightarrow$ Se acepta la hipótesis nula y, en consecuencia, no hay diferencia de ganancia de pesos debida a las vitaminas.

194. Se usaron tres clases de gasolina en cuatro marcas de automóviles. ¿Indican los valores de la tabla (kilómetros recorridos por litro) una diferencia significativa entre las clases de gasolina?. ¿Y entre las marcas de automóviles?. Significación del 5%.

Gasolina	Automóviles			
	A1	A2	A3	A4
G1	6,290	6,460	6,418	6,205
G2	6,503	6,290	7,225	6,843
G3	7,310	7,183	7,438	7,055

Solución:

1) Se supone que se cumplen las hipótesis de normalidad, homocedasticidad y elección al azar de las muestras. Se aplica el modelo ANOVA con dos factores de variación sin interacción (cuando el efecto de un factor depende de cual sea el nivel del otro factor):

Factor A: Tipos de gasolina

Factor B: Marcas de automóviles

♦ Para analizar si existe diferencia significativa entre las clases de gasolina, se establece el contraste de hipótesis:

H_0 : No hay diferencia significativa entre los efectos medios de las distintas gasolinas

H_1 : Si hay diferencia significativa entre los efectos medios de las distintas gasolinas

♦ Para analizar si existe diferencia significativa entre los efectos medios producidos por las distintas marcas de automóviles, se establece el contraste:

H_0 : No hay diferencia significativa entre marcas de automóviles

H_1 : Si hay diferencia significativa entre marcas de automóviles

La descomposición de la varianza total sin interacción: $SCT = SCF_A + SCF_B + SCR$

Gasolina	Automóviles				
	A1	A2	A3	A4	TOTAL
G1	6,290	6,460	6,418	6,205	$Y_{1\cdot} = 25,373$
G2	6,503	6,290	7,225	6,843	$Y_{2\cdot} = 26,861$
G3	7,310	7,183	7,438	7,055	$Y_{4\cdot} = 28,986$
TOTAL	$Y_{\cdot 1} = 20,103$	$Y_{\cdot 2} = 19,933$	$Y_{\cdot 3} = 21,081$	$Y_{\cdot 4} = 20,103$	$Y_{\cdot\cdot} = 81,22$

$$n = 1, G = 3, J = 4, N = G \cdot J = 12$$

$$SCT = \sum_{g=1}^3 \sum_{j=1}^4 Y_{gj}^2 - \frac{Y_{\cdot\cdot}^2}{N} = 551,997 - \frac{81,22^2}{12} = 551,997 - 549,724 = 2,273$$

$$SCF_A = \frac{\sum_{g=1}^G Y_{i\cdot}^2}{n \cdot J} - \frac{Y_{\cdot\cdot}^2}{N} = \frac{\sum_{g=1}^3 Y_{i\cdot}^2}{4} - \frac{Y_{\cdot\cdot}^2}{12} = \left[\frac{25,373^2}{4} + \frac{26,861^2}{4} + \frac{28,986^2}{4} \right] - \frac{81,22^2}{12} = 1,649$$

$$SCF_B = \frac{\sum_{j=1}^J Y_{\cdot j}^2}{n \cdot G} - \frac{Y_{\cdot\cdot}^2}{N} = \frac{\sum_{j=1}^4 Y_{\cdot j}^2}{3} - \frac{Y_{\cdot\cdot}^2}{12} = \left[\frac{20,103^2}{3} + \frac{19,933^2}{3} + \frac{21,081^2}{3} + \frac{20,103^2}{3} \right] - \frac{81,22^2}{12} = 549,998 - 549,724 = 0,274$$

$$SCT = SCF_A + SCF_B + SCR \rightarrow SCR = SCT - SCF_A - SCF_B = 2,273 - 1,649 - 0,274 = 0,351$$

$$MCF_A = \frac{SCF_A}{G-1} = \frac{1,649}{2} = 0,824 \quad MCR = \frac{SCR}{(G-1)(J-1)} = \frac{0,351}{6} = 0,0585 \quad F_A = \frac{MCF_A}{MCR} = 14,108$$

$$MCF_B = \frac{SCF_B}{J-1} = \frac{0,274}{3} = 0,091 \quad F_B = \frac{MCF_B}{MCR} = 1,563 \quad MCT = \frac{SCT}{N-1} = \frac{2,273}{11} = 0,207$$

Fuente variación	Suma de cuadrados	Grados libertad	Media cuadrática	Estadístico
Intergrupos	$SCF_A = 1,649$	$G - 1 = 2$	$MCF_A = \frac{SCF_A}{G-1} = 0,824$	$F_A = \frac{MCF_A}{MCR} = 14,108$
Intersujetos	$SCF_B = 0,274$	$J - 1 = 3$	$MCF_B = \frac{SCF_B}{J-1} = 0,091$	$F_B = \frac{MCF_B}{MCR} = 1,563$
Error	$SCR = 0,351$	$(G-1)(J-1) = 6$	$MCR = \frac{SCR}{(G-1)(J-1)} = 0,058$	
Total	$SCT = 2,273$	$N - 1 = 11$	$MCT = \frac{SCT}{N-1} = 0,207$	

⇒ ¿Existe diferencia significativa entre las clases de gasolina?

Estadístico observado o empírico: $F_A = \frac{MCF_A}{MCR} = 14,108$

Estadístico teórico: $F_{\alpha, (G-1), (G-1)(J-1)} = F_{0,05, 2, 6} = 5,1433$

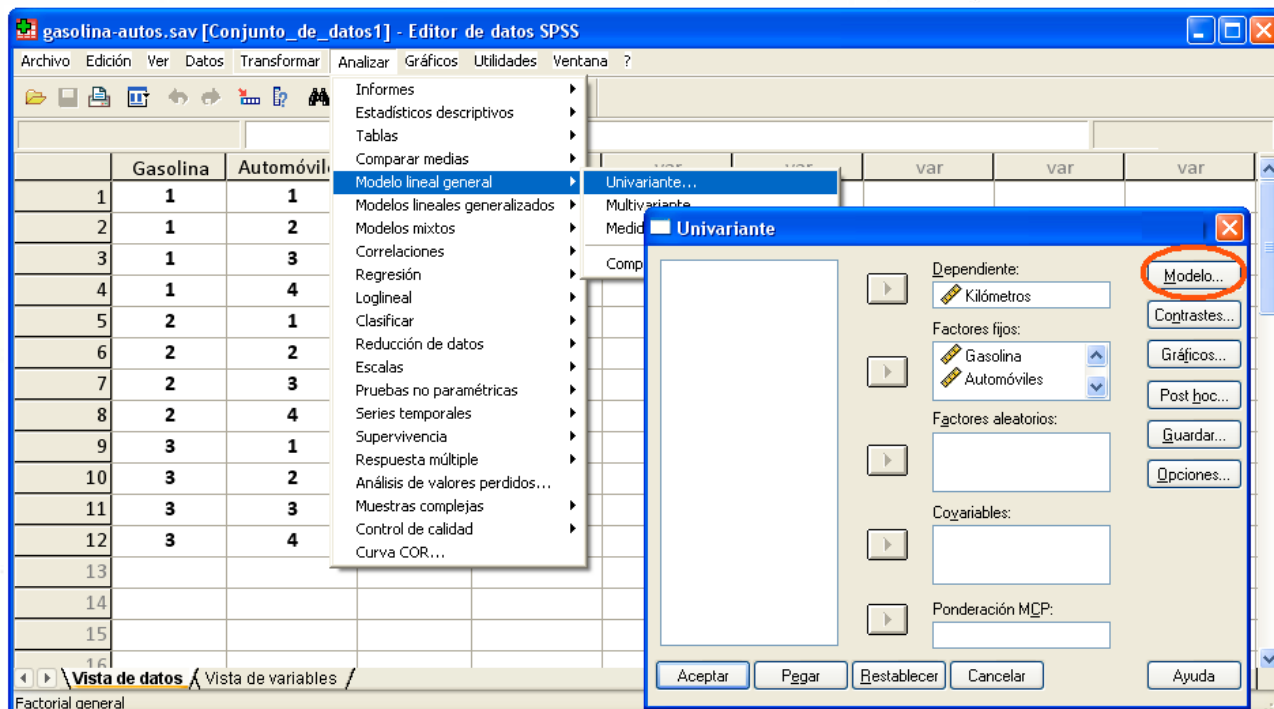
Siendo $F_A = 14,108 > 5,1433 = F_{0,05, 2, 6}$ se rechaza la hipótesis nula. En consecuencia, si hay diferencia significativa entre los efectos medios de las distintas gasolinas.

⇒ ¿Existe diferencia significativa entre los efectos medios producidos por las distintas marcas de automóviles?

Estadístico observado o empírico: $F_B = \frac{MCF_B}{MCR} = 1,563$

Estadístico teórico: $F_{\alpha, (J-1), (G-1)(J-1)} = F_{0,05, 3, 6} = 4,7571$

Como $F_B = 1,563 < 4,7571 = F_{0,05, 3, 6}$ se acepta la hipótesis nula. Por tanto, no hay diferencia significativa entre las marcas de automóviles.



Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

Univariante: Modelo

Especificar modelo

☐ Factorial completo ☒ Personalizado

Factores y covariables:

Gasolina(F)
Automóviles(F)

Modelo:

Gasolina
Automóviles

Construir términos

Efectos princip...

Suma de cuadrados: Tipo III ☒ Incluir la intersección en el modelo

Continuar Cancelar Ayuda

Pruebas de los efectos inter-sujetos

Variable dependiente: Kilómetros

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	1,923 ^a	5	,385	6,581	,020
Intersección	549,724	1	549,724	9408,45	,000
Gasolina	1,649	2	,824	14,108	,005
Automóviles	,274	3	,091	1,563	,293
Error	,351	6	,058		
Total	551,997	12			
Total corregida	2,273	11			

a. R cuadrado = ,846 (R cuadrado corregida = ,717)

El factor Gasolina tiene un $p_{\text{valor}} = 0,05 \leq 0,05 = \alpha \rightarrow$ Se rechaza la hipótesis nula, concluyendo que hay diferencia significativa entre los efectos medios de las distintas gasolinas.

El factor Automóviles tiene un $p_{\text{valor}} = 0,293 > 0,05 = \alpha \rightarrow$ Se acepta la hipótesis nula, con lo que no hay diferencia significativa entre las marcas de automóviles.

195. Analizar el efecto de las variables Edad y Fumar sobre la Ansiedad Social, con un nivel de significación del 5%.

Edad	Fumar	
	Si	No
15 - 25	3,91	4,83
	5,01	3,95
	4,44	4,04
	3,33	3,66
	4,71	9,42
25 - 35	5,65	9,66
	6,49	7,68
	5,47	9,57
	5,72	7,93
	5,44	7,39
35 45	4,94	5,92
	7,13	5,45
	5,54	5,19
	5,92	6,12
	6,16	4,45

Solución:

En cada caso, las hipótesis a contrastar son las siguientes:

▪ Efecto A: Categoría edad

$$\begin{cases} H_0: \text{La media de edad en cada categoría es la misma} \\ H_1: \text{Las medias no son iguales} \end{cases}$$

▪ Efecto B: Fumar

$$\begin{cases} H_0: \text{La media no difiere entre fumadores y no fumadores} \\ H_1: \text{Las medias no son iguales} \end{cases}$$

▪ Efecto interacción: Edad x Fumar

Bajo los supuestos de Independencia, Normalidad y Homocedasticidad, se calculan los estadísticos del contraste.

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

Fumar			
Edad	Si (1)	No (2)	TOTAL
1	3,91	4,83	$Y_{1.} = 47,30$ $\bar{Y}_{1.} = 4,73$
	5,01 $Y_{11} = 21,40$	3,95 $Y_{12} = 25,90$	
	4,44 $\bar{Y}_{11} = 4,28$	4,04 $\bar{Y}_{12} = 5,18$	
	3,33	3,66	
	4,71	9,42	
2	5,65	9,66	$Y_{2.} = 71$ $\bar{Y}_{2.} = 7,1$
	6,49	7,68	
	5,47 $Y_{21} = 28,77$	9,57 $Y_{22} = 42,23$	
	5,72 $\bar{Y}_{21} = 5,754$	7,93 $\bar{Y}_{22} = 8,446$	
	5,44	7,39	
3	4,94	5,92	$Y_{3.} = 56,82$ $\bar{Y}_{3.} = 5,682$
	7,13	5,45	
	5,54 $Y_{31} = 29,69$	5,19 $Y_{32} = 27,13$	
	5,92 $\bar{Y}_{31} = 5,938$	6,12 $\bar{Y}_{32} = 5,426$	
	6,16	4,45	
TOTAL	$Y_{.1} = 79,86$ $\bar{Y}_{.1} = 5,324$	$Y_{.2} = 95,26$ $\bar{Y}_{.2} = 6,351$	$Y_{..} = 175,12$ $\bar{Y}_{..} = 5,8373$

$$G = 3 \quad J = 2 \quad n = 5 \quad \sum_{i=1}^5 \sum_{g=1}^3 \sum_{j=1}^2 Y_{igj}^2 = 1106,281$$

$$SCT = \sum_{i=1}^5 \sum_{g=1}^3 \sum_{j=1}^2 Y_{igj}^2 - \frac{Y_{..}^2}{N} = 1106,28 - \frac{175,12^2}{30} = 84,047$$

$$SCF_A = \frac{\sum_{g=1}^G Y_{.g}^2}{n \cdot J} - \frac{Y_{..}^2}{N} = \frac{\sum_{g=1}^3 Y_{.g}^2}{5 \cdot 2} - \frac{Y_{..}^2}{30} = \left[\frac{47,30^2}{5 \cdot 2} + \frac{71^2}{5 \cdot 2} + \frac{56,82^2}{5 \cdot 2} \right] - \frac{175,12^2}{30} = 28,446$$

$$SCF_B = \frac{\sum_{j=1}^J Y_{.j}^2}{n \cdot G} - \frac{Y_{..}^2}{N} = \frac{\sum_{j=1}^2 Y_{.j}^2}{5 \cdot 3} - \frac{Y_{..}^2}{30} = \left[\frac{79,86^2}{15} + \frac{95,26^2}{15} \right] - \frac{175,12^2}{30} = 7,905$$

$$SCR = \sum_{i=1}^n \sum_{g=1}^G \sum_{j=1}^J Y_{igj}^2 - \frac{\sum_{g=1}^G \sum_{j=1}^J Y_{gj}^2}{n} = \sum_{i=1}^5 \sum_{g=1}^3 \sum_{j=1}^2 Y_{igj}^2 - \frac{\sum_{g=1}^3 \sum_{j=1}^2 Y_{gj}^2}{n} =$$

$$= 1106,281 - \left[\frac{21,40^2}{5} + \frac{25,90^2}{5} + \frac{28,77^2}{5} + \frac{42,23^2}{5} + \frac{29,69^2}{5} + \frac{27,13^2}{5} \right] = 34,803$$

$$SCT = SCF_A + SCF_B + SCF_{AB} + SCR \rightarrow SCF_{AB} = SCT - SCF_A - SCF_B - SCR$$

$$SCF_{AB} = SCT - SCF_A - SCF_B - SCR = 84,04 - 28,44 - 7,90 - 34,80 = 12,892$$

$$\text{O bien, } SCF_{AB} = \frac{\sum_{g=1}^G \sum_{j=1}^J Y_{gj}^2}{n} - \frac{\sum_{g=1}^G Y_{g\cdot}^2}{n \cdot J} - \frac{\sum_{j=1}^J Y_{\cdot j}^2}{n \cdot G} + \frac{Y_{\cdot\cdot}^2}{N} = \frac{\sum_{g=1}^3 \sum_{j=1}^2 Y_{gj}^2}{5} - \frac{\sum_{g=1}^3 Y_{g\cdot}^2}{5 \cdot 2} - \frac{\sum_{j=1}^2 Y_{\cdot j}^2}{5 \cdot 3} + \frac{Y_{\cdot\cdot}^2}{30} =$$

$$= 1071,478 - 1050,68 - 1030,139 + 1022,234 = 12,892$$

donde,

$$\frac{\sum_{g=1}^3 \sum_{j=1}^2 Y_{gj}^2}{n} = \frac{21,40^2 + 25,90^2 + 28,77^2 + 42,23^2 + 29,69^2 + 27,13^2}{5} = 1071,478$$

$$\frac{\sum_{g=1}^3 Y_{g\cdot}^2}{10} = \frac{47,30^2 + 71^2 + 56,82^2}{10} = 1050,680$$

$$\frac{\sum_{j=1}^2 Y_{\cdot j}^2}{15} = \frac{79,86^2 + 95,26^2}{15} = 1030,139 \quad \frac{Y_{\cdot\cdot}^2}{30} = \frac{175,12^2}{30} = 1022,234$$

Análisis de la varianza con dos factores

Fuente variación	Suma de cuadrados	Grados libertad	Media cuadrática	Estadístico F
Factor A	$SCF_A = 28,44$	$(G-1) = 3-1 = 2$	$MCF_A = \frac{SCF_A}{G-1} = 14,22$	$F = \frac{MCF_A}{MCR} = 9,80$
Factor B	$SCF_B = 7,90$	$(J-1) = 2-1 = 1$	$MCF_B = \frac{SCF_B}{J-1} = 7,90$	$F = \frac{MCF_B}{MCR} = 5,44$
Interacción	$SCF_{AB} = 12,89$	$(G-1)(J-1) = 2$	$MCF_{AB} = \frac{SCF_{AB}}{(G-1)(J-1)} = 6,44$	$F = \frac{MCF_{AB}}{MCR} = 4,44$
Residual	$SCR = 34,80$	$N - G \cdot J = 24$	$MCR = \frac{SCR}{N - G \cdot J} = 1,45$	
Total	$SCT = 84,04$	$N - 1 = 29$	$MCT = \frac{SCT}{N-1} = 2,89$	

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....

$$MCF_A = \frac{SCF_A}{G-1} = \frac{28,44}{2} = 14,22 \quad MCF_B = \frac{SCF_B}{J-1} = 7,90 \quad MCF_{AB} = \frac{SCF_{AB}}{(G-1)(J-1)} = \frac{12,89}{2} = 6,44$$

$$MCR = \frac{SCR}{N-GJ} = \frac{34,80}{24} = 1,45 \quad MCT = \frac{SCT}{N-1} = \frac{84,04}{29} = 2,89$$

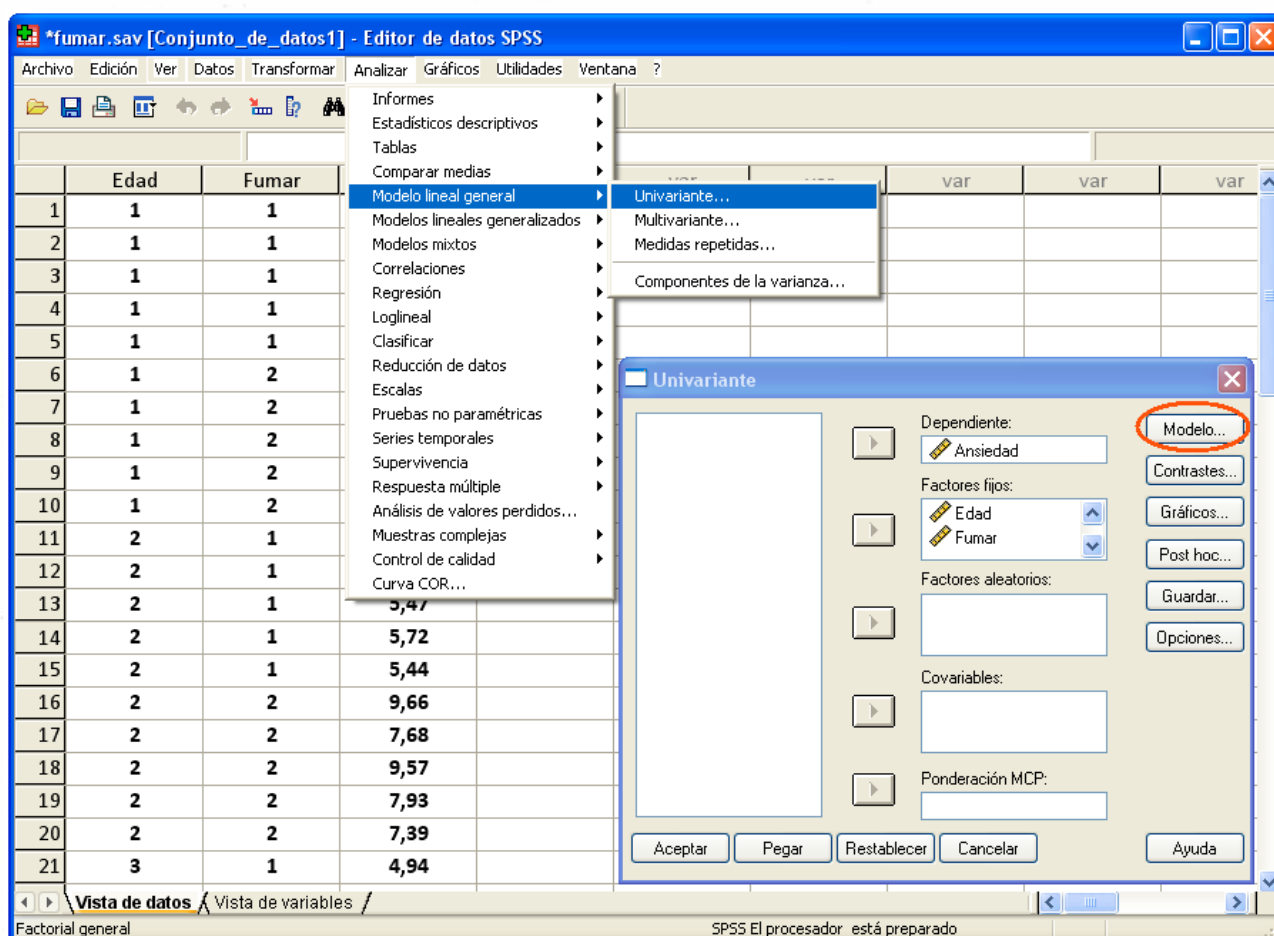
$$F = \frac{MCF_A}{MCR} = \frac{14,22}{1,45} = 9,80 \quad F = \frac{MCF_B}{MCR} = \frac{7,90}{1,45} = 5,44 \quad F = \frac{MCF_{AB}}{MCR} = \frac{6,44}{1,45} = 4,44$$

El efecto de cada factor (edad, fumar) es:

$$\eta_A^2 = \frac{SCF_A}{SCF_A + SCR} = \frac{28,44}{28,44 + 34,80} = 0,45$$

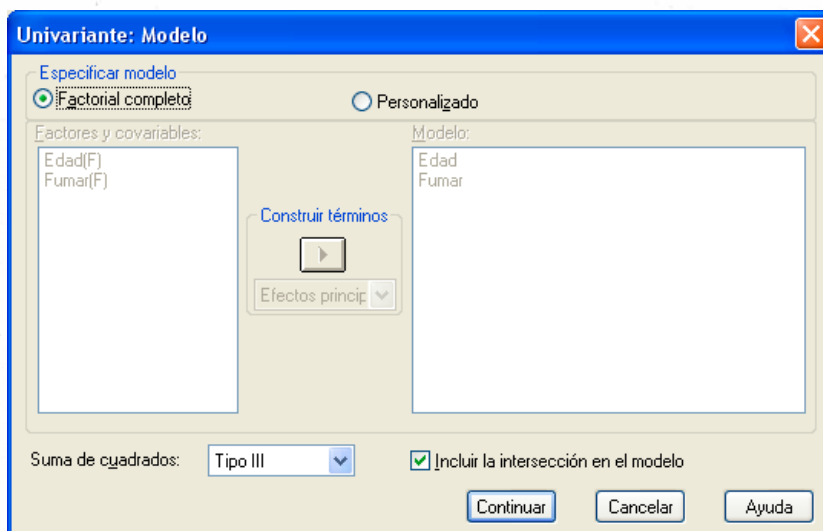
$$\eta_B^2 = \frac{SCF_B}{SCF_B + SCR} = \frac{7,90}{7,90 + 34,80} = 0,185$$

$$\eta_{AB}^2 = \frac{SCF_{AB}}{SCF_{AB} + SCR} = \frac{12,89}{12,89 + 34,80} = 0,270$$



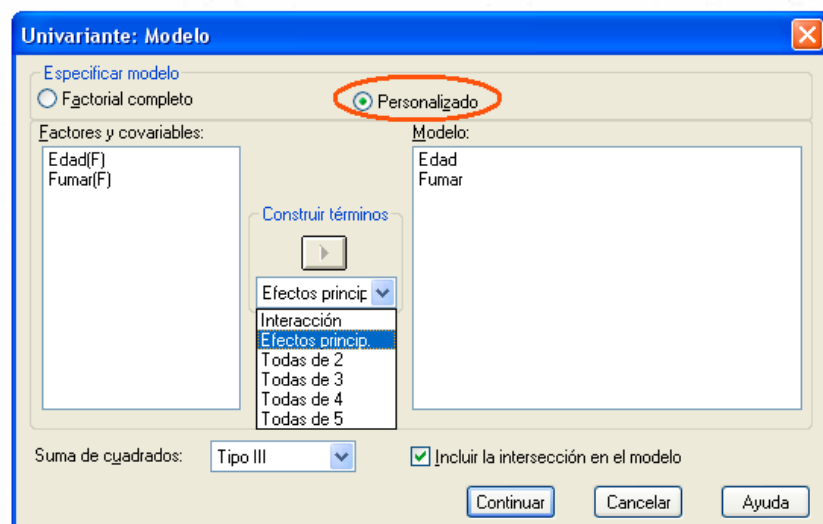
The screenshot shows the SPSS Editor de datos window for a file named *fumar.sav. The data table has columns 'Edad' and 'Fumar'. The 'Univariate' dialog box is open, showing 'Ansiedad' as the dependent variable and 'Edad' and 'Fumar' as fixed factors. The 'Modelo...' button is circled in red.

	Edad	Fumar	
1	1	1	
2	1	1	
3	1	1	
4	1	1	
5	1	1	
6	1	2	
7	1	2	
8	1	2	
9	1	2	
10	1	2	
11	2	1	
12	2	1	
13	2	1	5,47
14	2	1	5,72
15	2	1	5,44
16	2	2	9,66
17	2	2	7,68
18	2	2	9,57
19	2	2	7,93
20	2	2	7,39
21	3	1	4,94



Se puede construir el modelo de interés introduciendo los factores tratamiento y sus interacciones a criterio del investigador.

Por defecto está el modelo *Factorial completo*, que incluye a todos los factores tratamiento introducidos previamente y a todas sus iteraciones.



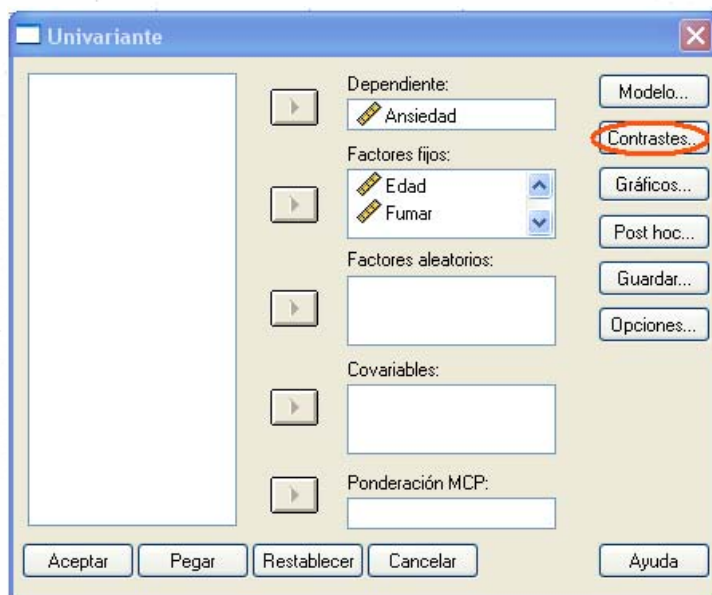
Si se desea un modelo alternativo, por ejemplo un modelo de dos vías principales, es decir, sin interacciones, se marca *Personalizado*, con lo que se activa el campo *Construir términos*. Se continua pasando los Factores y covariables al modelo.

Se deja por defecto *Suma de cuadrados Tipo III*, es el procedimiento más utilizado. Proporciona la descomposición de las sumas de cuadrados tal y como se ha visto anteriormente.

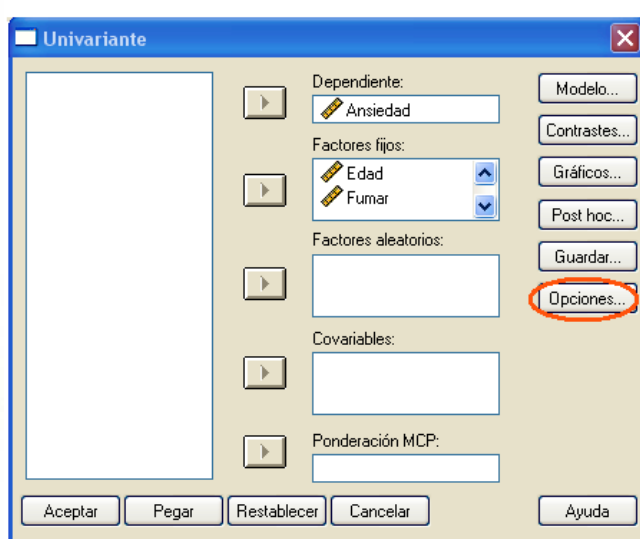
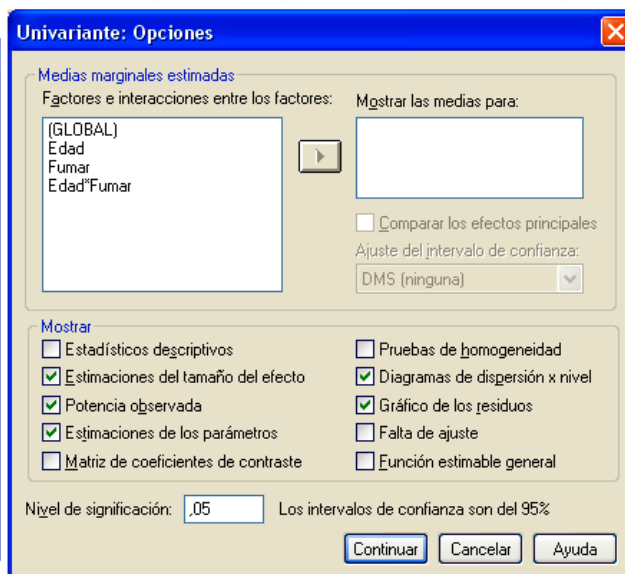
La suma de cuadrados Tipo III explicada por el factor Edad es igual a la diferencia entre la suma de cuadrados residual del modelo completo (con todos los factores) sin el factor Edad y la suma de cuadrados residual del modelo completo.

El orden de introducción de los factores de tratamiento es independiente y produce una descomposición ortogonal, de modo que las sumas de cuadrados suman la suma de cuadrados total.

En ocasiones, también es de utilidad la *Suma de cuadrados Tipo I*, se conoce como el método de descomposición jerárquica de la suma de cuadrados. La Suma de cuadrados tipo I explicada por el factor Edad es igual a la diferencia entre la suma de cuadrados residual del modelo construido con los factores incluidos hasta ese momento menos el factor Edad y la suma de cuadrados residual del modelo con el factor Edad incluido. Cuando el diseño es balanceado (mismo número de observaciones en cada casilla) la descomposición en *Suma de cuadrados Tipo III* coincide con la descomposición en *Suma de cuadrados Tipo I*.




Se elige el factor tratamiento sobre cuyos niveles se ejecutarán los contrastes. En el subcuadro se selecciona la familia de contrastes de interés.

Estadísticos descriptivos proporciona la media muestral, desviación típica muestral y tamaño para cada nivel y combinación de niveles.

Estimaciones del tamaño del efecto añaden al cuadro ANOVA el coeficiente η^2 (eta al cuadrado) para cada fuente de variación incluida en el modelo. Con objeto de medir el efecto de cada factor (y de la interacción) separadamente se utilizan estadísticos eta cuadrados parciales, que se definen:

$$\eta^2_{\text{Edad}} = \frac{SCF_{\text{Edad}}}{SCF_{\text{Edad}} + SCR} = 0,45 \quad \eta^2_{\text{Fumar}} = \frac{SCF_{\text{Fumar}}}{SCF_{\text{Fumar}} + SCR} = 0,185 \quad \eta^2_{\text{Edad*Fumar}} = \frac{SCF_{\text{Edad*Fumar}}}{SCF_{\text{Edad*Fumar}} + SCR} = 0,270$$

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

Potencia observada añade al cuadrado ANOVA una nueva columna con el valor de la potencia del *F-test* para cada fuente de variación incluida en el modelo.

El valor de la potencia se entenderá como la capacidad de la prueba de hipótesis para detectar una diferencia poblacional real entre los niveles de la fuente de variación igual a la diferencia observada en las muestras, con un nivel de significación igual al prefijado.

La potencia de un contraste se define como la probabilidad de rechazar la hipótesis nula cuando es falsa.

Prueba de los efectos inter-sujetos: **Modelo Factorial completo**

Variable dependiente: Ansiedad

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación	Eta al cuadrado parcial	Parámetro de no centralidad	Potencia observada ^a
Modelo corregido	49,244 ^b	5	9,849	6,792	,000	,586	33,958	,990
Intersección	1022,234	1	1022,234	704,919	,000	,967	704,919	1,000
Edad	28,446	2	14,223	9,808	,001	,450	19,616	,969
Fumar	7,905	1	7,905	5,451	,028	,185	5,451	,611
Edad * Fumar	12,892	2	6,446	4,445	,023	,270	8,890	,708
Error	34,803	24	1,450					
Total	1106,281	30						
Total corregida	84,047	29						

a. Calculado con alfa = ,05

b. R cuadrado = ,586 (R cuadrado corregida = ,500)

Se observan coeficientes parciales η^2 bajos para todas las fuentes de variación, siendo el mayor el correspondiente a la intersección (0,967).

En todos los casos la potencia es alta lo que sugiere una alta capacidad del test para rechazar la hipótesis nula con diferencias reales iguales a las observadas.

Estimaciones de los parámetros

Variable dependiente: Ansiedad

Parámetro	B	Error típ.	t	Significación	Intervalo de confianza al 95%.		Eta al cuadrado parcial	Parámetro de no centralidad	Potencia observada ^a
					Límite inferior	Límite superior			
Intersección	5,426	,539	10,075	,000	4,315	6,537	,809	10,075	1,000
[Edad=1]	-,246	,762	-,323	,749	-1,818	1,326	,004	,323	,061
[Edad=2]	3,020	,762	3,965	,001	1,448	4,592	,396	3,965	,967
[Edad=3]	0 ^b	.	.	.	.	.	.	.	.
[Fumar=1]	,512	,762	,672	,508	-1,060	2,084	,018	,672	,099
[Fumar=2]	0 ^b	.	.	.	.	.	.	.	.
[Edad=1] * [Fumar=1]	-1,412	1,077	-1,311	,202	-3,635	,811	,067	1,311	,242
[Edad=1] * [Fumar=2]	0 ^b	.	.	.	.	.	.	.	.
[Edad=2] * [Fumar=1]	-3,204	1,077	-2,975	,007	-5,427	-,981	,269	2,975	,814
[Edad=2] * [Fumar=2]	0 ^b	.	.	.	.	.	.	.	.
[Edad=3] * [Fumar=1]	0 ^b	.	.	.	.	.	.	.	.
[Edad=3] * [Fumar=2]	0 ^b	.	.	.	.	.	.	.	.

a. Calculado con alfa = ,05

b. Al parámetro se le ha asignado el valor cero porque es redundante.

La información puede esquematizarse como sigue:

Variable dependiente: Ansiedad

Parámetro	B
Intersección	5,426
[Edad=1]	-,246
[Edad=2]	3,020
[Edad=3]	0 ^b
[Fumar=1]	,512
[Fumar=2]	0 ^b
[Edad=1] * [Fumar=1]	-1,412
[Edad=1] * [Fumar=2]	0 ^b
[Edad=2] * [Fumar=1]	-3,204
[Edad=2] * [Fumar=2]	0 ^b
[Edad=3] * [Fumar=1]	0 ^b
[Edad=3] * [Fumar=2]	0 ^b

- Parámetros que se igualan a 0 por redundantes:

$$\alpha_3 = \beta_2 = (\alpha\beta)_{12} = (\alpha\beta)_{22} = (\alpha\beta)_{31} = (\alpha\beta)_{32} = 0$$

- Parámetros que se estiman:

$$\hat{\mu} = 5,426$$

$$\hat{\alpha}_1 = -0,246 \quad \hat{\alpha}_2 = 3,020$$

$$\hat{\beta}_1 = 0,512$$

$$(\alpha\beta)_{11} = -1,412 \quad (\alpha\beta)_{21} = -3,204$$

El modelo estimado se resume (empleando variables dummy) como:

$$Y = 5,426 - 0,246 \cdot X_1 + 3,020 \cdot X_2 + 0,512 \cdot Z_1 - 1,412 \cdot X_1 \cdot Z_1 - 3,204 \cdot X_2 \cdot Z_1 + \varepsilon$$

siendo:

$X_1 = 1$ si la edad es del tipo 1 y $X_1 = 0$ en otro caso

$X_2 = 1$ si la edad es del tipo 2 y $X_2 = 0$ en otro caso

$Z_1 = 1$ si fuma y $Z_1 = 0$ en caso de no fumar

- Excepto en los casos $\alpha_2 = 3,020$ y $(\alpha\beta)_{21} = -3,204$, los parámetros no son significativamente distintos de cero y por tanto no tienen peso significativo en el modelo estimado.

- La Tabla proporciona además los intervalos de confianza para cada parámetro y los coeficientes η^2 y potencia observada para cada contraste de los parámetros.

Contraste de Levene sobre la igualdad de las
varianzas error

Variable dependiente: Ansiedad

F	gl1	gl2	Significación
2,706	5	24	,045

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

a. Diseño: Intersección+Edad+Fumar+Edad * Fumar

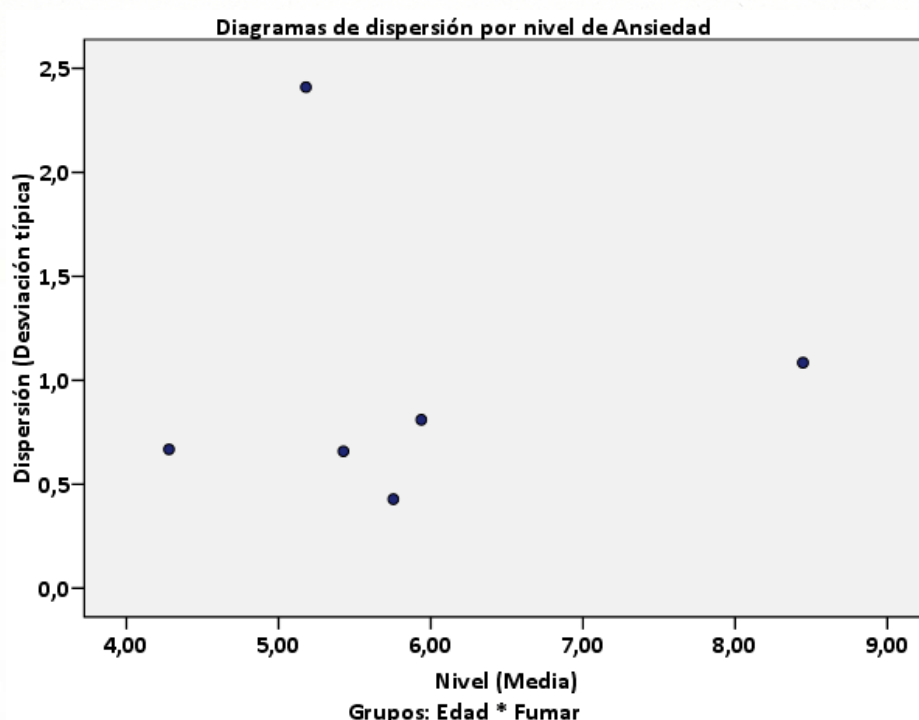
- La prueba de Levene conduce a un estadístico igual a 2,706 que en una $F_{5,24}$ deja a su derecha una cola de probabilidad 0,045.

Siendo $p_valor = 0,045 < \alpha = 0,05$, no se acepta la hipótesis nula de igualdad de varianzas (*criterio de homocedasticidad*).

- **Diagramas de dispersión x nivel:** Aporta información gráfica sobre la homogeneidad de la varianza, completando la información recogida en Pruebas de homogeneidad.

El diagrama consiste en un gráfico de puntos: cada uno representa una población (combinación de niveles) donde la media se representa en la abscisa y la desviación típica (o varianza) en la ordenada.

El objetivo es detectar si la varianza depende de la media (es frecuente comprobar que la respuesta crezca cuando lo hace la media).



- **Gráfico de los residuos:** Es un gráfico matricial de puntos enfrentando *Valores Observados* de la variable respuesta, *Valores Pronosticados* por el modelo estimado para la variable respuesta, y *Residuos* (valores observados menos valores pronosticados por el modelo estimado).

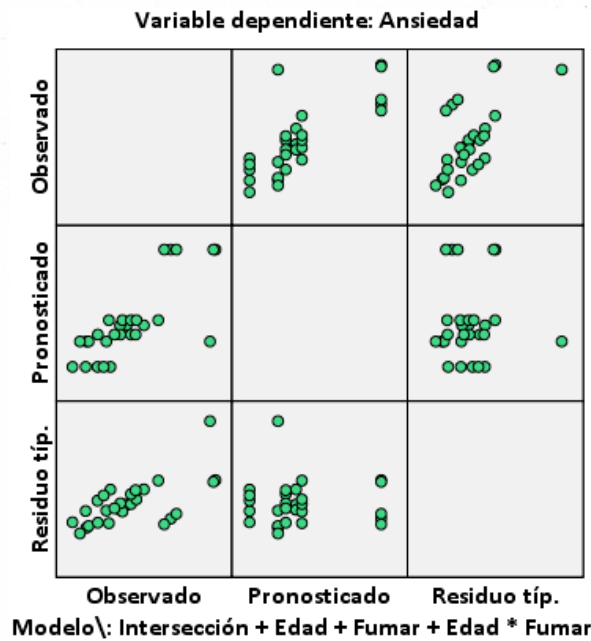
Los residuos se muestran tipificados (divididos por la raíz cuadrada del error cuadrático medio estimado).

Este gráfico matricial da una idea si se verifican los supuestos de independencia, homogeneidad de varianzas y falta de ajuste. En efecto:

- ✓ Si existe Independencia, el gráfico de dispersión Residuos-Observados no debería mostrar pauta de variación (línea, curva, ...). Gráfico de coordenadas (1, 3) o (3, 1) en la matriz.
- ✓ Si existe Homocedasticidad, el gráfico de dispersión Residuos-Pronosticados es de interés puesto que la dispersión de los residuos debe de ser similar para todos los valores pronosticados. Gráfico de coordenadas (2, 3) o (3, 2) en la matriz.

Asignatura..... Grupo.....
 Apellidos..... Nombre.....
 Ejercicio del día.....

- ✓ En un modelo con buen ajuste a los datos, la nube de puntos Pronosticados-Observados debiera mostrar un perfil cercano a la linealidad (a más linealidad mejor ajuste). Gráfico de coordenadas (1, 2) o (2, 1) en la matriz.



196. Se realiza un experimento para estudiar el efecto del nivel del agua sobre la longitud del tallo de dos tipos de plantas de guisantes, utilizando una significación del 5%. Para ello se utilizaron tres niveles de agua, los datos obtenidos se reflejan en la tabla adjunta, figurando en rojo el orden temporal de la toma de datos.

		Nivel del agua		
		Nivel I	Nivel II	Nivel III
Tipo Planta	Tipo 1	71,3 (1)	107,5 (2)	123,1 (11)
		75,1 (4)	96,1 (8)	125,2 (16)
		69,0 (5)	103,6 (18)	125,7 (19)
		73,2 (7)	100,7 (21)	121,0 (20)
		74,4 (22)	102,3 (26)	122,9 (30)
	Tipo 2	70,4 (13)	88,1 (6)	109,0 (3)
		73,2 (15)	85,8 (10)	101,1 (9)
		71,1 (23)	86,0 (17)	103,2 (12)
		71,2 (24)	87,5 (25)	109,7 (14)
		69,2 (27)	81,0 (29)	106,1 (28)

Solución:

Existen dos factores tratamiento: 'Niveles de agua' (con tres niveles de efectos fijos) y 'Tipo de planta' (con dos niveles de efectos fijos).

Los niveles se cruzan formando un total de 6 observaciones o condiciones experimentales diferentes.

Para cada tratamiento se obtienen 5 respuestas de la variable 'Crecimiento de la longitud del tallo' de otras tantas unidades experimentales. Se crea así un diseño balanceado (de 5 réplicas), aleatorio y de tamaño 30.

Al disponer de réplicas es posible contrastar la existencia de interacción entre los niveles de los dos factores tratamiento, de manera que el modelo matemático es el propio de un diseño completo de dos vías.

Nivel del agua

	Nivel I	Nivel II	Nivel III	Total
Tipo planta 1	71,3 (1)	107,5 (2)	123,1 (11)	$Y_{1\cdot} = 1491,1$ $\bar{Y}_{1\cdot} = 99,41$
	75,1 (4) $Y_{11} = 363$	96,1 (8) $Y_{12} = 510,2$	125,2 (16) $Y_{13} = 617,9$	
	69,0 (5) $\bar{Y}_{11} = 72,60$	103,6 (18) $\bar{Y}_{12} = 120,04$	125,7 (19) $\bar{Y}_{13} = 123,58$	
	73,2 (7)	100,7 (21)	121,0 (20)	
	74,4 (22)	102,3 (26)	122,9 (30)	
Tipo planta 2	70,4 (13)	88,1 (6)	109,0 (3)	$Y_{2\cdot} = 1312,6$ $\bar{Y}_{2\cdot} = 87,51$
	73,2 (15) $Y_{21} = 355,1$	85,8 (10) $Y_{22} = 428,4$	101,1 (9) $Y_{23} = 529,1$	
	71,1 (23) $\bar{Y}_{21} = 71,02$	86,0 (17) $\bar{Y}_{22} = 85,68$	103,2 (12) $\bar{Y}_{23} = 105,82$	
	71,2 (24)	87,5 (25)	109,7 (14)	
	69,2 (27)	81,0 (29)	106,1 (28)	
Total	$Y_{\cdot 1} = 718,1$	$Y_{\cdot 2} = 938,6$	$Y_{\cdot 3} = 1147$	$Y_{\cdot\cdot} = 2803,7$
	$\bar{Y}_{\cdot 1} = 71,81$	$\bar{Y}_{\cdot 2} = 102,86$	$\bar{Y}_{\cdot 3} = 114,7$	$\bar{Y}_{\cdot\cdot} = 93,46$

$$G = 2 \quad J = 3 \quad n = 5 \quad N = 5 \cdot 2 \cdot 3 = 30 \quad \sum_{i=1}^5 \sum_{g=1}^2 \sum_{j=1}^3 Y_{igj}^2 = 272890,99$$

$$SCT = \sum_{i=1}^5 \sum_{g=1}^2 \sum_{j=1}^3 Y_{igj}^2 - \frac{Y_{\cdot\cdot}^2}{N} = 272890,99 - \frac{2803,7^2}{30} = 10866,534$$

$$SCF_A = \frac{\sum_{g=1}^G Y_{i\cdot}^2}{n \cdot J} - \frac{Y_{\cdot\cdot}^2}{N} = \frac{\sum_{g=1}^2 Y_{i\cdot}^2}{5 \cdot 3} - \frac{Y_{\cdot\cdot}^2}{30} = \left[\frac{1491,1^2}{5 \cdot 3} + \frac{1312,6^2}{5 \cdot 3} \right] - \frac{2803,7^2}{30} = 1062,075$$

$$SCF_B = \frac{\sum_{j=1}^J Y_{\cdot j}^2}{n \cdot G} - \frac{Y_{\cdot\cdot}^2}{N} = \frac{\sum_{j=1}^3 Y_{\cdot j}^2}{5 \cdot 2} - \frac{Y_{\cdot\cdot}^2}{30} = \left[\frac{718,1^2}{5 \cdot 2} + \frac{938,6^2}{5 \cdot 2} + \frac{1147^2}{5 \cdot 2} \right] - \frac{2803,7^2}{30} = 9200,201$$

$$SCR = \sum_{i=1}^n \sum_{g=1}^G \sum_{j=1}^J Y_{igj}^2 - \frac{\sum_{g=1}^G \sum_{j=1}^J Y_{gj}^2}{n} = \sum_{i=1}^5 \sum_{g=1}^2 \sum_{j=1}^3 Y_{igj}^2 - \frac{\sum_{g=1}^2 \sum_{j=1}^3 Y_{gj}^2}{n} =$$

$$= 272890,99 - \left[\frac{363^2}{5} + \frac{510,2^2}{5} + \frac{617,9^2}{5} + \frac{355,1^2}{5} + \frac{428,4^2}{5} + \frac{529,1^2}{5} \right] = 202,424$$

$$SCT = SCF_A + SCF_B + SCF_{AB} + SCR \rightarrow SCF_{AB} = SCT - SCF_A - SCF_B - SCR$$

$$SCF_{AB} = SCT - SCF_A - SCF_B - SCR = 10866,534 - 1062,075 - 9200,201 - 202,424 = 401,834$$

$$\text{o bien, } SCF_{AB} = \frac{\sum_{g=1}^G \sum_{j=1}^J Y_{gj}^2}{n} - \frac{\sum_{g=1}^G Y_{g\cdot}^2}{n \cdot J} - \frac{\sum_{j=1}^J Y_{\cdot j}^2}{n \cdot G} + \frac{Y_{\cdot\cdot}^2}{N} = \frac{\sum_{g=1}^2 \sum_{j=1}^3 Y_{gj}^2}{5} - \frac{\sum_{g=1}^2 Y_{g\cdot}^2}{5 \cdot 3} - \frac{\sum_{j=1}^3 Y_{\cdot j}^2}{5 \cdot 2} + \frac{Y_{\cdot\cdot}^2}{30}$$

$$R^2 = \frac{SCF}{SCT} = \frac{SCF_A + SCF_B + SCF_{AB}}{SCT} = 0,981$$

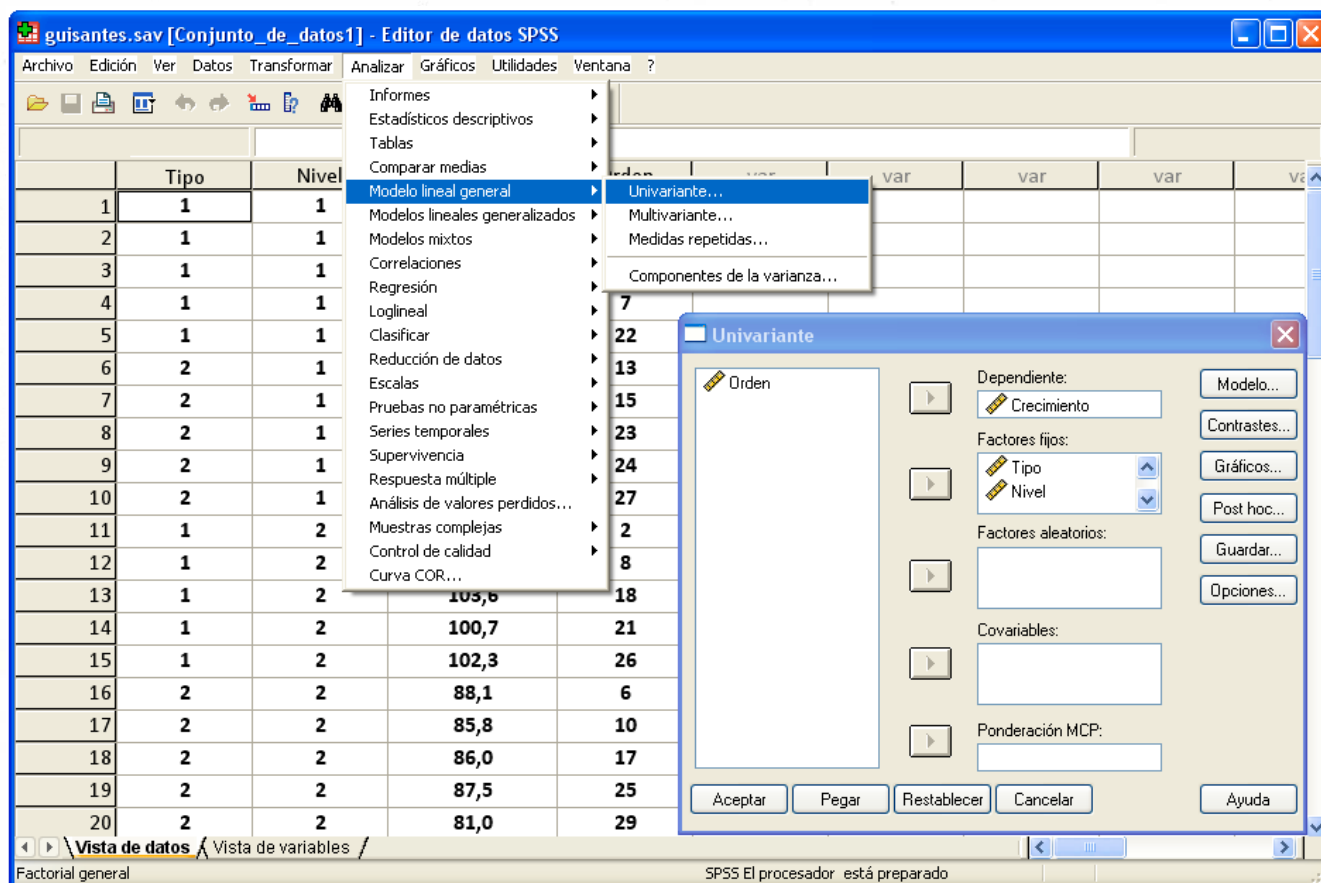
Análisis de la varianza con dos factores

Fuente variación	Suma de cuadrados	Grados libertad	Media cuadrática	Estadístico F
Factor A	$SCF_A = 1062,075$	$(G-1) = 1$	$MCF_A = \frac{SCF_A}{G-1} = 1062,075$	$F_A = \frac{MCF_A}{MCR} = 125,928$
Factor B	$SCF_B = 9200,201$	$(J-1) = 2$	$MCF_B = \frac{SCF_B}{J-1} = 4600,10$	$F_B = \frac{MCF_B}{MCR} = 545,423$
Interacción	$SCF_{AB} = 401,834$	$(G-1)(J-1) = 2$	$MCF_{AB} = \frac{SCF_{AB}}{(G-1)(J-1)} = 200,917$	$F_{AB} = \frac{MCF_{AB}}{MCR} = 23,822$
Residual	$SCR = 202,424$	$N - GJ = 24$	$MCR = \frac{SCR}{N - GJ} = 8,434$	$R^2 = \frac{SCF}{SCT} = 0,981$
Total	$SCT = 10866,534$	$N - 1 = 29$	$MCT = \frac{SCT}{N - 1} = 374,708$	

$$\eta_A^2 = \frac{SCF_A}{SCF_A + SCR} = \frac{1062,075}{1062,075 + 202,424} = 0,840$$

$$\eta_B^2 = \frac{SCF_B}{SCF_B + SCR} = \frac{9200,201}{9200,201 + 202,424} = 0,978$$

$$\eta_{AB}^2 = \frac{SCF_{AB}}{SCF_{AB} + SCR} = \frac{401,834}{401,834 + 202,424} = 0,665$$



Dependiente: Se introduce la variable respuesta (necesariamente cuantitativa y unidimensional). En el ejemplo, la variable *crecimiento de los tallos de guisantes*.

Factores fijos: Se introducen las variables conteniendo los niveles de los factores de tratamiento con efectos fijos (ya que sus efectos sobre la respuesta desean ser comparados y son el objeto de la investigación). En el ejemplo, *variables tipo y nivel*.

Factores aleatorios: Se introducen los niveles de los factores tratamiento con efectos aleatorios (los niveles son una muestra aleatoria de una población mayor y por ello no son el objetivo de la investigación ya que la inferencia se realiza sobre la población y no sobre la muestra). Introducir tantas variables como factores de tratamiento. En el ejercicio, ninguna variable es de efectos aleatorios por lo que queda vacío.

Covariables: Introducir las covariables (factor de control en el modelo no categórico sino continuo). En el ejemplo no hay covariables por lo que este campo queda vacío.

Ponderación MCP: Variable de pesos para computar los estimadores mínimo cuadráticos de manera ponderada. De utilidad cuando no se tiene homocedasticidad.

Factores inter-sujetos

		N
Tipo	1	15
	2	15
Nivel	1	10
	2	10
	3	10

El cuadro Factores inter-sujetos proporciona un resumen de las etiquetas de valor de cada nivel y el número de observaciones de cada nivel.

Modelo Factorial completo

Pruebas de los efectos inter-sujetos

Variable dependiente: Crecimiento

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación
Modelo corregido	10664,110 ^a	5	2132,822	252,874	,000
Intersección	262024,456	1	262024,456	31066,4	,000
Tipo	1062,075	1	1062,075	125,923	,000
Nivel	9200,201	2	4600,100	545,402	,000
Tipo * Nivel	401,834	2	200,917	23,821	,000
Error	202,424	24	8,434		
Total	272890,990	30			
Total corregida	10866,534	29			

a. R cuadrado = ,981 (R cuadrado corregida = ,977)

$$R^2 = \frac{SCF}{SCT} = \frac{SCF_A + SCF_B + SCF_{AB}}{SCT} = \frac{1062,075 + 9200,201 + 401,834}{10866,534} = \frac{10664,110}{10866,534} = 0,981$$

La tabla Prueba de los efectos inter-sujetos proporciona el cuadro ANOVA. Lo proporciona para dos posibles descomposiciones de la suma de cuadrados global según que en el modelo la respuesta aparezca en bruto (Suma de Cuadrados Total (Total = SCT) o con la constante sustraída (Suma de Cuadrados Corregida).

El modelo en bruto es: $Y_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}$

Equivalente a: $SC \text{ Total} = SC \text{ Intersección} + SC \text{ Tipo} + SC \text{ Nivel} + SC \text{ Tipo} \times \text{Nivel} + SC \text{ Error}$

El modelo corregido (sustrayendo la constante): $Y_{ij} - \mu = \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ij}$

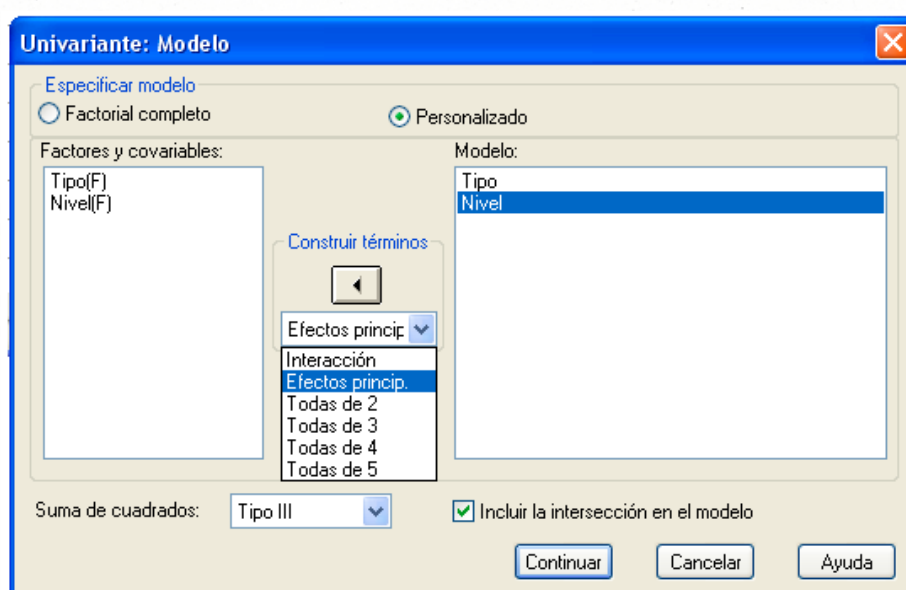
Notación: $\frac{SC \text{ Total corregido}}{SCT} = \frac{SC \text{ Tipo}}{SCF_A} + \frac{SC \text{ Nivel}}{SCF_B} + \frac{SC \text{ Tipo} * \text{Nivel}}{SCF_{AB}} + \frac{SC \text{ Error}}{SCR}$

MODELO: Cuando se desea un modelo alternativo, por ejemplo un modelo de *dos vías principales* (esto es, sin interacción) entonces:

Marcar Personalizado, activando los campos que le siguen.

En Construir términos se selecciona Efectos Principales, marcar tipo (F) en el campo Factores y covariables y pulsar la flecha Construir términos. El factor tipo (F) ya formará parte del modelo al aparecer en el campo Modelo.

Se hace lo mismo con el factor nivel (F).



Se deja por defecto Suma de cuadrados Tipo III, porque es el procedimiento más utilizado. Proporciona la descomposición de las sumas de cuadrados tal y como ya se ha visto.

La Suma de cuadrados Tipo III explicada por un factor A es igual a la diferencia entre la suma de cuadrados residual del modelo completo (con todos los factores) sin el factor A y la suma de cuadrados residual del modelo completo. Es independiente del orden de introducción de los factores tratamiento y produce una descomposición ortogonal de modo que las sumas de cuadrados suman la suma de cuadrados total.

En ocasiones también es de utilidad la Suma de cuadrados de Tipo I, conocida como el método de descomposición jerárquica de la suma de cuadrados.

La suma de cuadrados de Tipo I explicada por un factor A es igual a la diferencia entre la suma de cuadrados residual del modelo construido con los factores incluidos hasta ese momento menos el factor A y la suma de cuadrados residual del modelo con A incluido.

Cuando el diseño es balanceado, la descomposición en suma de cuadrados Tipo III coincide con la descomposición en suma de cuadrados Tipo I.

CONTRASTES: En el cuadro de Factores se elige el factor tratamiento sobre cuyos niveles se ejecutarán los contrastes. Por ejemplo, marcar Nivel (Ninguno).



Contrastes Helmert: Familia de $(J-1)$ contrastes comparando el efecto de cada nivel del factor (menos el último) con el efecto promedio de los niveles subsiguientes.

En este caso, los contrastes de Helmert son dos:

$$\left\{ \begin{array}{l} H_0^{(1)} : \beta_1 - \frac{1}{2}(\beta_2 + \beta_3) = 0 \quad \text{frente} \quad H_1^{(1)} : \beta_1 - \frac{1}{2}(\beta_2 + \beta_3) \neq 0 \\ H_0^{(2)} : \beta_2 - \beta_3 = 0 \quad \text{frente} \quad H_1^{(2)} : \beta_2 - \beta_3 \neq 0 \end{array} \right.$$

Siendo β_j el efecto marginal del nivel del agua j-ésimo.

La resolución particular de cada uno de ellos se realiza bajo el supuesto de independencia y normalidad, mediante la t de Student con un número de grados de libertad igual al empleado para estimar el error.

En general, se rechaza $H_0 : \sum_j \alpha_j \beta_j = 0$ (con $\sum_j \beta_j = 0$) al nivel de significación α si:

$$\left| \frac{\sum_j \alpha_j \hat{\beta}_j}{\sqrt{\text{SCMR} \sum_j \frac{\alpha_j^2}{n_j}}} \right| > t_{\alpha/2, \text{gl}(\text{SCMR})}$$

Siendo $\hat{\beta}_j$ la media muestral de las n_j observaciones en el j-ésimo nivel y SCMR la suma de cuadrados media residual.

El criterio para rechazar la hipótesis nula se puede escribir: $\frac{\left(\sum_j \alpha_j \hat{\beta}_j / \sum_j \frac{\alpha_j^2}{n_j} \right)^2}{\text{SCMR}} > F_{\alpha/2, 1}$

Resultados del contraste (matriz K)

Nivel contraste de Helmert		Variable dependiente	
		Crecimiento	
Nivel 1 - Anterior	Estimación del contraste	$\beta_1 - \frac{1}{2}(\beta_2 + \beta_3)$	-32,470
	Valor hipotetizado	0	
	Diferencia (Estimado - Hipotetizado)		-32,470
	Error típ.		1,125
	Significación		,000
	Intervalo de confianza al 95 %	Límite inferior	-34,791
		Límite superior	-30,149
Nivel 2 - Nivel 3	Estimación del contraste	$\beta_2 - \beta_3$	-20,840
	Valor hipotetizado	0	
	Diferencia (Estimado - Hipotetizado)		-20,840
	Error típ.		1,299
	Significación		,000
	Intervalo de confianza al 95 %	Límite inferior	-23,521
		Límite superior	-18,159

La estimación del contraste $\beta_1 - \frac{1}{2}(\beta_2 + \beta_3)$ presenta un $p_valor = 0 < 0,05$, rechazando la

hipótesis nula $H_0^{(1)} : \beta_1 - \frac{1}{2}(\beta_2 + \beta_3) = 0$.

De otra parte, el intervalo de confianza no cubre el 0, lo que conduce a la misma conclusión.

La estimación del contraste $\beta_2 - \beta_3$ presenta un $p_valor = 0 < 0,05$, rechazando la hipótesis nula $H_0^{(2)} : \beta_2 - \beta_3 = 0$.

El intervalo de confianza no cubre el 0, lo que conduce a la misma conclusión.

Resultados de la prueba

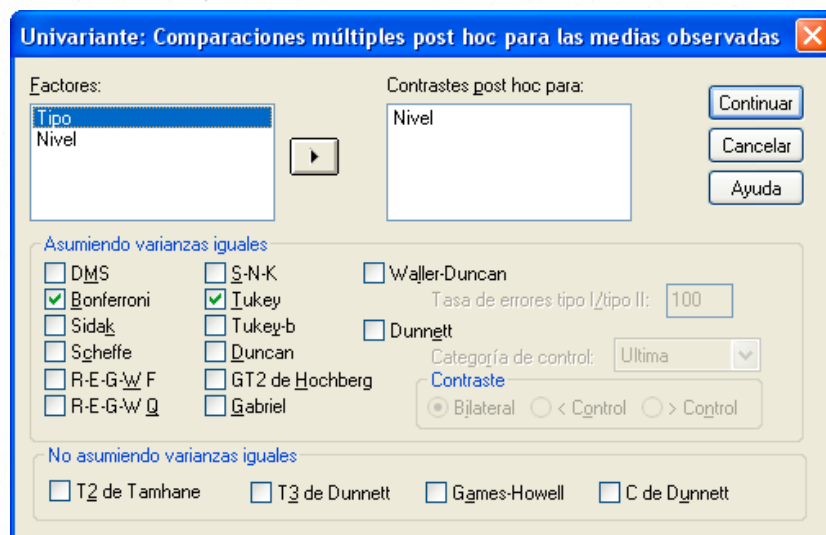
Variable dependiente: Crecimiento

Fuente	Suma de cuadrados	gl	Media cuadrática	F	Significación
Contraste	9200,201	2	4600,100	545,402	,000
Error	202,424	24	8,434		

El cuadro de *Resultados de la prueba* muestra la suma de cuadrados explicada por los dos contrastes de Helmert y el correspondiente F-test con dos grados de libertad (uno por contraste) y 24 grados de libertad que muestra la significación estadística conjunta de ambos contrastes. Siendo $p_valor = 0 < 0,05$ no se acepta la hipótesis nula de igualdad de varianzas (*criterio de homocedasticidad*)

Post Hoc (CONTRASTES DE RANGO MÚLTIPLE) : Las comparaciones Post hoc para los efectos marginales sólo deben realizarse en el supuesto de no interacción.

Se solicitan los procedimientos de Tukey y Bonferroni para los niveles del factor tratamiento 'nivel del agua' y ninguno para el otro factor.



Univariate: Comparaciones múltiples post hoc para las medias observadas

Factores:
Tipo
Nivel

Contrastes post hoc para:
Nivel

Continuar
Cancelar
Ayuda

Asumiendo varianzas iguales

☐ DMS ☐ S-N-K ☐ Waller-Duncan
☒ Bonferroni ☒ Tukey
☐ Sidak ☐ Tukey-b
☐ Scheffe ☐ Duncan
☐ R-E-G-W F ☐ GT2 de Hochberg
☐ R-E-G-W Q ☐ Gabriel

Tasa de errores tipo I/tipo II: 100

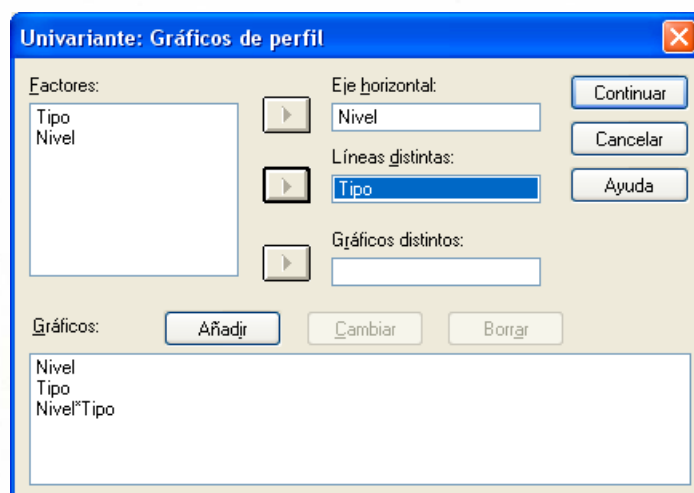
Categoría de control: Ultima

Contraste
☒ Bilateral ☐ < Control ☐ > Control

No asumiendo varianzas iguales

☐ T2 de Tamhane ☐ T3 de Dunnett ☐ Games-Howell ☐ C de Dunnett

Gráfico de las tres medidas estimadas



Univariate: Gráficos de perfil

Factores:
Tipo
Nivel

Eje horizontal:
Nivel

Líneas distintas:
Tipo

Gráficos distintos:

Gráficos:
Nivel
Tipo
Nivel*Tipo

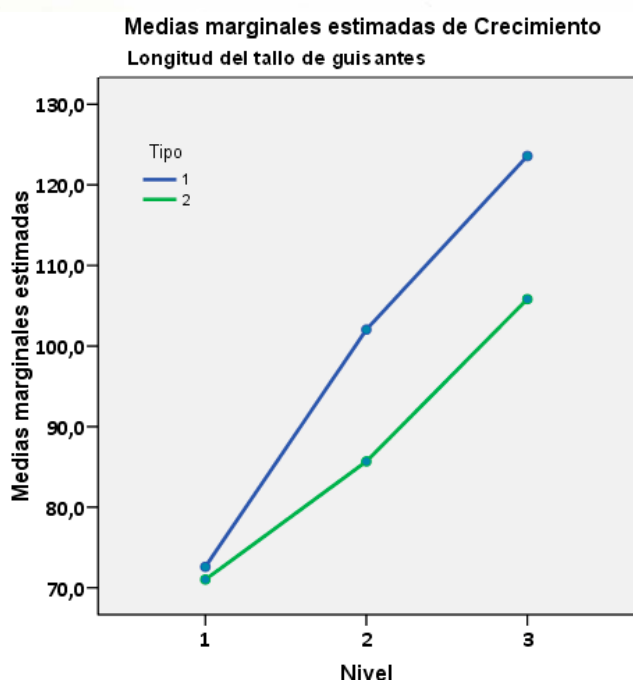
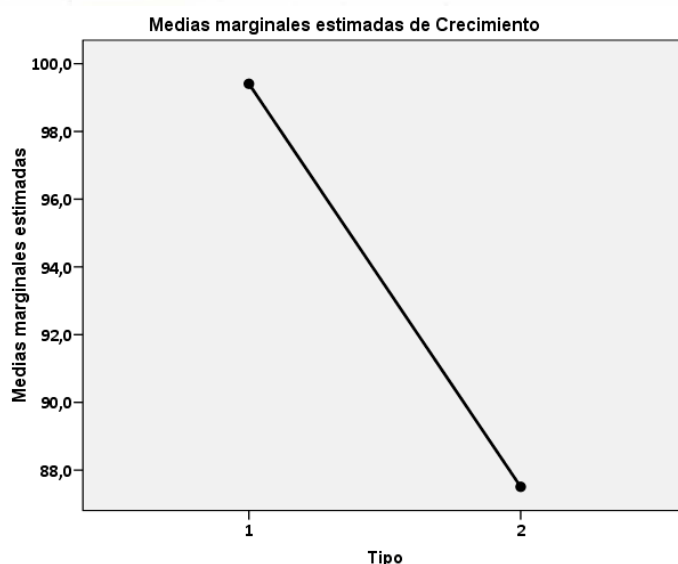
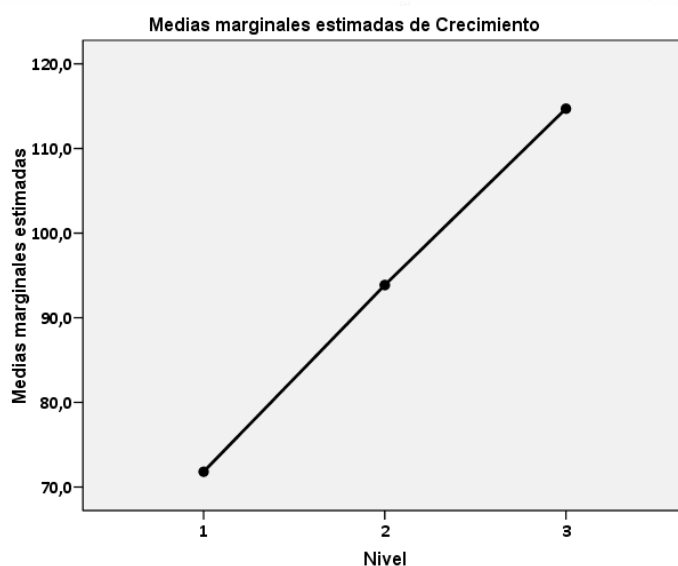
Añadir Cambiar Borrar

Continuar
Cancelar
Ayuda

Asignatura..... Grupo.....

Apellidos..... Nombre.....

Ejercicio del día.....



En el gráfico de interacción se observa que aparentemente no existe interacción entre tipos de planta y los dos últimos niveles de agua:

El crecimiento del tipo 1 ha sido superior en igual longitud promedio tanto con el nivel 2 como con el nivel 3 de agua y, análogamente, el nivel 3 de agua fue igual o mejor para el crecimiento con independencia del tipo de planta.

Sin embargo, este comportamiento no se ha mantenido con el nivel 1 de agua, estando aquí la posible interacción. Probablemente el nivel 1 de agua sea el menos exitoso, especialmente con un comportamiento malo para el tipo 1 de plantas.

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

Pruebas Post hoc**Nivel****Comparaciones múltiples****Variable dependiente: Crecimiento**

Variable dependiente: crecimiento						Intervalo de confianza al 95%.	
	(I) Nivel	(J) Nivel	Diferencia entre medias (I-J)	Error típ.	Significación	Límite inferior	Límite superior
DHS de Tukey	1	2	-22,050*	1,2988	,000	-25,293	-18,807
		3	-42,890*	1,2988	,000	-46,133	-39,647
	2	1	22,050*	1,2988	,000	18,807	25,293
		3	-20,840*	1,2988	,000	-24,083	-17,597
	3	1	42,890*	1,2988	,000	39,647	46,133
		2	20,840*	1,2988	,000	17,597	24,083
Bonferroni	1	2	-22,050*	1,2988	,000	-25,393	-18,707
		3	-42,890*	1,2988	,000	-46,233	-39,547
	2	1	22,050*	1,2988	,000	18,707	25,393
		3	-20,840*	1,2988	,000	-24,183	-17,497
	3	1	42,890*	1,2988	,000	39,547	46,233
		2	20,840*	1,2988	,000	17,497	24,183

Basado en las medias observadas.

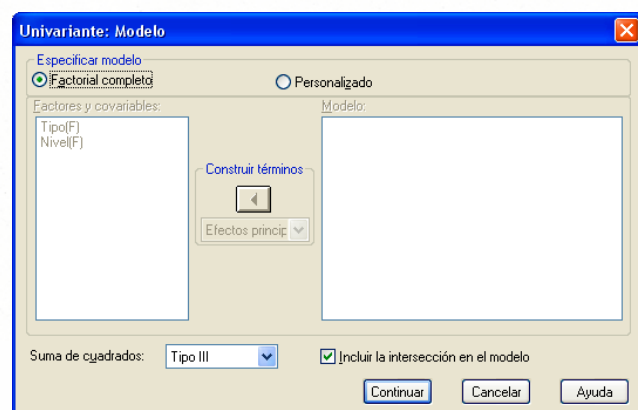
*. La diferencia de medias es significativa al nivel ,05.

Hay $c = \frac{J(J-1)}{2} = \frac{3 \cdot 2}{2} = 3$ comparaciones que se pueden realizar por niveles del agua. En el contraste de Bonferroni (LSD) y el contraste DHS de Tukey la hipótesis nula establece que el crecimiento medio de las plantas de guisantes es el mismo por niveles de agua.

Se observa en todos los casos, que bien por el $p_valor = 0,000 < 0,05$, o bien observando los intervalos de confianza que no cubren el 0, que no se verifica la hipótesis nula, habiendo una diferencia de medias significativa.

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

Opciones: Modelo Factorial completo



Univariante: Modelo

Especificar modelo:
☒ Factorial completo
☐ Personalizado

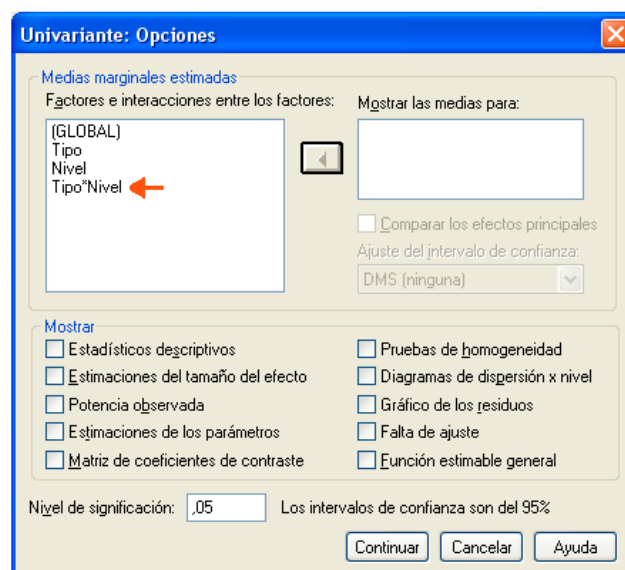
Factores y covariables:
 Tipo(F)
 Nivel(F)

Modelo:

Construir términos
 Efectos princip. ▼

Suma de cuadrados: Tipo III ▼ ☒ Incluir la intersección en el modelo

Continuar Cancelar Ayuda



Univariante: Opciones

Medias marginales estimadas
 Factores e interacciones entre los factores: Mostrar las medias para:

(GLOBAL)
 Tipo
 Nivel
 Tipo*Nivel

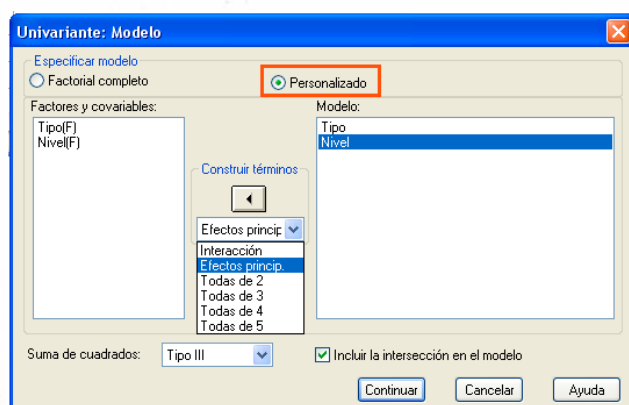
☐ Comparar los efectos principales
 Ajuste del intervalo de confianza:
 DMS (ninguna) ▼

Mostrar
☐ Estadísticos descriptivos
☐ Estimaciones del tamaño del efecto
☐ Potencia observada
☐ Estimaciones de los parámetros
☐ Matriz de coeficientes de contraste
☐ Pruebas de homogeneidad
☐ Diagramas de dispersión x nivel
☐ Gráfico de los residuos
☐ Falta de ajuste
☐ Función estimable general

Nivel de significación: .05 Los intervalos de confianza son del 95%

Continuar Cancelar Ayuda

Opciones: Modelo Personalizado, efectos principales



Univariante: Modelo

Especificar modelo:
☐ Factorial completo
☒ Personalizado

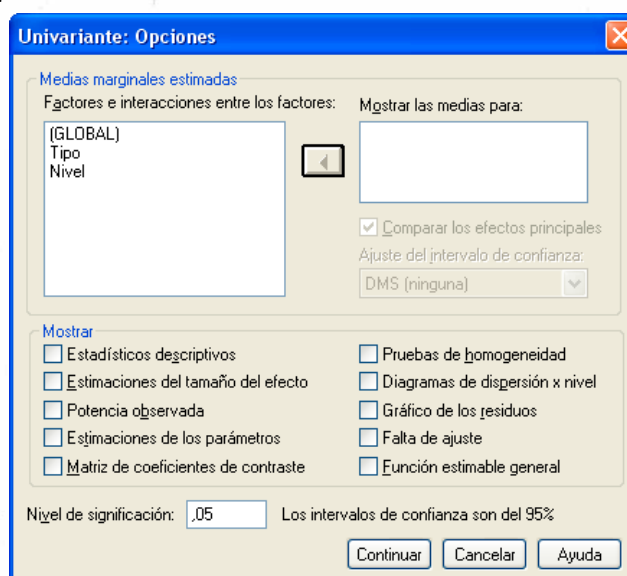
Factores y covariables:
 Tipo(F)
 Nivel(F)

Modelo:
 Tipo
 Nivel

Construir términos
 Efectos princip. ▼
 Interacción
 Efectos princip.
 Todas de 2
 Todas de 3
 Todas de 4
 Todas de 5

Suma de cuadrados: Tipo III ▼ ☒ Incluir la intersección en el modelo

Continuar Cancelar Ayuda



Univariante: Opciones

Medias marginales estimadas
 Factores e interacciones entre los factores: Mostrar las medias para:

(GLOBAL)
 Tipo
 Nivel

☒ Comparar los efectos principales
 Ajuste del intervalo de confianza:
 DMS (ninguna) ▼

Mostrar
☐ Estadísticos descriptivos
☐ Estimaciones del tamaño del efecto
☐ Potencia observada
☐ Estimaciones de los parámetros
☐ Matriz de coeficientes de contraste
☐ Pruebas de homogeneidad
☐ Diagramas de dispersión x nivel
☐ Gráfico de los residuos
☐ Falta de ajuste
☐ Función estimable general

Nivel de significación: .05 Los intervalos de confianza son del 95%

Continuar Cancelar Ayuda

Univariante: Opciones

Medias marginales estimadas

Factores e interacciones entre los factores:

(GLOBAL)
Tipo
Nivel
Tipo*Nivel

Mostrar las medias para:

Tipo
Nivel
Tipo*Nivel

☒ Comparar los efectos principales

Ajuste del intervalo de confianza:

DMS (ninguna)

Mostrar

☐ Estadísticos descriptivos ☒ Pruebas de homogeneidad

☐ Estimaciones del tamaño del efecto ☒ Diagramas de dispersión x nivel

☐ Potencia observada ☒ Gráfico de los residuos

☒ Estimaciones de los parámetros ☐ Falta de ajuste

☐ Matriz de coeficientes de contraste ☐ Función estimable general

Nivel de significación: .05 Los intervalos de confianza son del 95%

Continuar Cancelar Ayuda

Señalar que, como ejemplo, el cuadro de medias estimadas para los niveles de la interacción Tipo*Nivel no es el mismo para un *modelo completo* que para un *modelo de efectos principales*.

Al señalar *Comparar los efectos principales* se realizan las comparaciones de medias dos a dos (pairwise comparisons) de todos los niveles de los efectos principales con tres posibles criterios que se seleccionarán en la lista desplegable de Ajustes del intervalo de confianza.

DMS (Mínima Diferencia Significativa) consiste en utilizar el criterio de la t de Student para comparar dos muestras independientes, no considerando por tanto la tasa de error tipo I global.

Bonferroni y Sidak suponen diferentes correcciones para controlar la tasa de error tipo I, siendo recomendables cuando hay que realizar un número muy grande de comparaciones dos a dos.

En el *Modelo factorial completo*, se tiene:

Tipo de planta

Estimaciones

Variable dependiente: Crecimiento

Tipo	Media	Error típ.	Intervalo de confianza al 95%.	
			Límite inferior	Límite superior
1	99,407	,750	97,859	100,954
2	87,507	,750	85,959	89,054

Comparaciones por pares

Variable dependiente: Crecimiento

(I) Tipo	(J) Tipo	Diferencia entre medias (I-J)	Error típ.	Significación ^a	Intervalo de confianza al 95 % para la diferencia ^a	
					Límite inferior	Límite superior
1	2	11,900*	1,060	,000	9,711	14,089
2	1	-11,900*	1,060	,000	-14,089	-9,711

Basadas en las medias marginales estimadas.

*. La diferencia de las medias es significativa al nivel ,05.

a. Ajuste para comparaciones múltiples: Diferencia menos significativa (equivalente a la ausencia de ajuste).

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

Contrastes univariados

Variable dependiente: Crecimiento

	Suma de cuadrados	gl	Media cuadrática	F	Significación
Contraste	1062,075	1	1062,075	125,923	,000
Error	202,424	24	8,434		

Cada prueba F contrasta el efecto simple de Tipo en cada combinación de niveles del resto de los efectos mostrados. Estos contrastes se basan en las comparaciones por pares, linealmente independientes, entre las medias marginales estimadas.

Nivel de agua

Estimaciones

Variable dependiente: Crecimiento

Nivel	Media	Error típ.	Intervalo de confianza al 95%.	
			Límite inferior	Límite superior
1	71,810	,918	69,915	73,705
2	93,860	,918	91,965	95,755
3	114,700	,918	112,805	116,595

Comparaciones por pares

Variable dependiente: Crecimiento

(I) Nivel	(J) Nivel	Diferencia entre medias (I-J)	Error típ.	Significación ^a	Intervalo de confianza al 95 % para la diferencia ^a	
					Límite inferior	Límite superior
1	2	-22,050*	1,299	,000	-24,731	-19,369
	3	-42,890*	1,299	,000	-45,571	-40,209
2	1	22,050*	1,299	,000	19,369	24,731
	3	-20,840*	1,299	,000	-23,521	-18,159
3	1	42,890*	1,299	,000	40,209	45,571
	2	20,840*	1,299	,000	18,159	23,521

Basadas en las medias marginales estimadas.

*. La diferencia de las medias es significativa al nivel ,05.

a. Ajuste para comparaciones múltiples: Diferencia menos significativa (equivalente a la ausencia de ajuste).

Contrastes univariados

Variable dependiente: Crecimiento

	Suma de cuadrados	gl	Media cuadrática	F	Significación
Contraste	9200,201	2	4600,100	545,402	,000
Error	202,424	24	8,434		

Cada prueba F contrasta el efecto simple de Nivel en cada combinación de niveles del resto de los efectos mostrados. Estos contrastes se basan en las comparaciones por pares, linealmente independientes, entre las medias marginales estimadas.

Tipo de planta * Nivel de agua

Variable dependiente: Crecimiento

Tipo	Nivel	Media	Error típ.	Intervalo de confianza al 95%.	
				Límite inferior	Límite superior
1	1	72,600	1,299	69,919	75,281
	2	102,040	1,299	99,359	104,721
	3	123,580	1,299	120,899	126,261
2	1	71,020	1,299	68,339	73,701
	2	85,680	1,299	82,999	88,361
	3	105,820	1,299	103,139	108,501

Se observa que el crecimiento medio estimado del tallo de guisantes de las plantas de tipo 1 tratadas con el nivel de agua 1 es 72,60.

En el modelo factorial completo este crecimiento medio se obtiene mediante: $\mu + \alpha_1 + \beta_1 + (\alpha\beta)_{11}$

Su estimador se calcula a través de las estimaciones mínimo-cuadráticas de sus parámetros:

$$\hat{\mu} + \hat{\alpha}_1 + \hat{\beta}_1 + (\hat{\alpha}\hat{\beta})_{11} = \bar{Y}_{11} = 72,60$$

concluyendo que, para este modelo, la estimación coincide con la media muestral de los crecimientos alcanzados con las plantas de tipo 1 tratadas con nivel de agua 1.

Pruebas de los efectos inter-sujetos

Variable dependiente: Crecimiento

Fuente	Suma de cuadrados tipo III	gl	Media cuadrática	F	Significación	Eta al cuadrado parcial	Parámetro de no centralidad	Potencia observada ^a
Modelo corregido	10664,110 ^b	5	2132,822	252,874	,000	,981	1264,369	1,000
Intersección	262024,456	1	262024,456	31066,4	,000	,999	31066,410	1,000
Tipo	1062,075	1	1062,075	125,923	,000	,840	125,923	1,000
Nivel	9200,201	2	4600,100	545,402	,000	,978	1090,804	1,000
Tipo * Nivel	401,834	2	200,917	23,821	,000	,665	47,643	1,000
Error	202,424	24	8,434					
Total	272890,990	30						
Total corregida	10866,534	29						

a. Calculado con alfa = ,05

b. R cuadrado = ,981 (R cuadrado corregida = ,977)

Estimación del tamaño del efecto: En el cuadro ANOVA se añade el coeficiente η^2 para cada fuente de variación incluida en el modelo.

$$\eta_A^2 = 0,840 \quad \eta_B^2 = 0,978 \quad \eta_{AB}^2 = 0,665$$

es por tanto una estimación de la proporción de varianza explicada por diferencias entre los niveles de la fuente de variación una vez eliminado el efecto de las otras fuentes de variación incluidas en el modelo. Se observan coeficientes η^2 altos para todas las fuentes de variación, siendo el menor el correspondiente a la interacción (0,665).

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

La potencia de un contraste es la probabilidad de rechazar la hipótesis nula cuando es falsa.

El valor de la potencia se entiende como la capacidad de la prueba de hipótesis para detectar una diferencia real (poblacional) entre los niveles de la fuente de variación igual a la diferencia observada entre las muestras.

En todos los casos la potencia es igual a 1, lo que sugiere una alta capacidad de F-test para rechazar la hipótesis nula con diferencias reales iguales a las observadas.

Estimaciones de los parámetros

Variable dependiente: Crecimiento

Parámetro	B	Error típ.	t	Significación	Intervalo de confianza al 95%.	
					Límite inferior	Límite superior
Intersección	105,820	1,299	81,476	,000	103,139	108,501
[Tipo=1]	17,760	1,837	9,669	,000	13,969	21,551
[Tipo=2]	0 ^a	.	.	.	.	.
[Nivel=1]	-34,800	1,837	-18,946	,000	-38,591	-31,009
[Nivel=2]	-20,140	1,837	-10,965	,000	-23,931	-16,349
[Nivel=3]	0 ^a	.	.	.	.	.
[Tipo=1] * [Nivel=1]	-16,180	2,598	-6,229	,000	-21,541	-10,819
[Tipo=1] * [Nivel=2]	-1,400	2,598	-,539	,595	-6,761	3,961
[Tipo=1] * [Nivel=3]	0 ^a	.	.	.	.	.
[Tipo=2] * [Nivel=1]	0 ^a	.	.	.	.	.
[Tipo=2] * [Nivel=2]	0 ^a	.	.	.	.	.
[Tipo=2] * [Nivel=3]	0 ^a	.	.	.	.	.

a. Al parámetro se le ha asignado el valor cero porque es redundante.

El modelo estimado se resume:

$$Y = 105,82 + 17,76 \cdot X_1 - 34,80 \cdot Z_1 - 20,140 \cdot Z_2 - 16,180 \cdot X_1 Z_1 - 1,400 \cdot X_1 Z_2 + \varepsilon$$

$X_1 = 1$ si la planta es de tipo 1, en otro caso vale 0

Las variables dummy: $Z_1 = 1$ si es tratada con el nivel 1 de agua, en otro caso vale 0

$Z_2 = 1$ si es tratada con el nivel 2 de agua, en otro caso vale 0

Contraste de Levene sobre la igualdad de las varianzas error³

Variable dependiente: Crecimiento

F	gl1	gl2	Significación
1,130	5	24	,372

Contrasta la hipótesis nula de que la varianza error de la variable dependiente es igual a lo largo de todos los grupos.

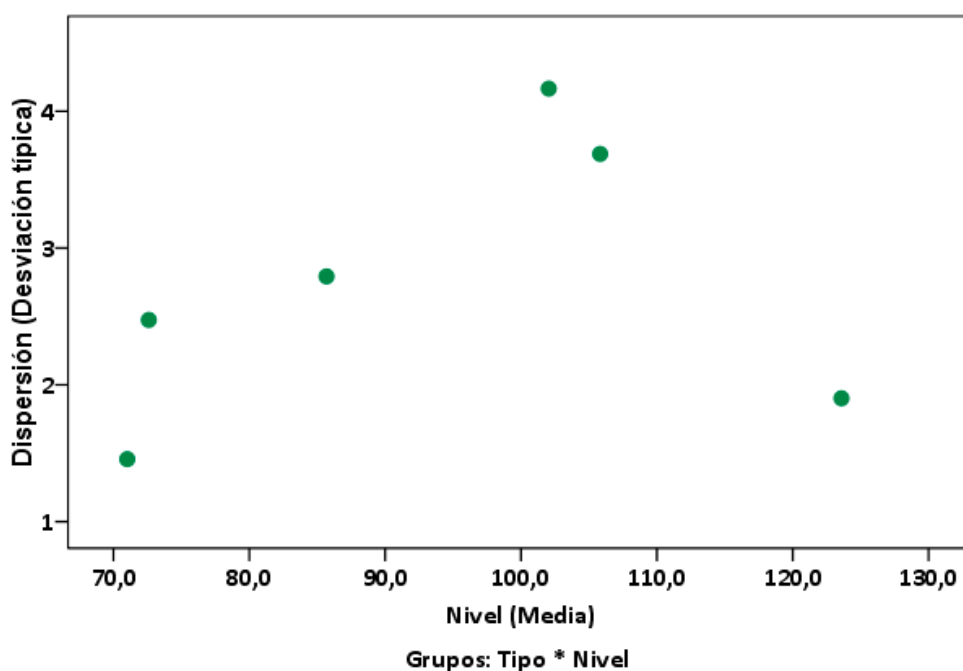
a. Diseño: Intersección + Tipo + Nivel + Tipo * Nivel

La prueba de Levene para testar la hipótesis nula de igualdad de varianzas de la variable crecimiento en todas las poblaciones definidas por combinaciones de los niveles de factores conduce a un estadístico igual a 1,130 que en una $F_{5, 24}$ deja a su derecha una cola de probabilidad 0,372.

Siendo $p_valor = 0,372 > 0,05$ no se rechaza la hipótesis nula de homogeneidad de varianzas de la variable crecimiento del tallo de guisantes en las 6 poblaciones definidas (combinación del tipo de planta y nivel de agua considerados), validando así una de las hipótesis estructurales del modelo propuesto (*homoscedasticidad*).

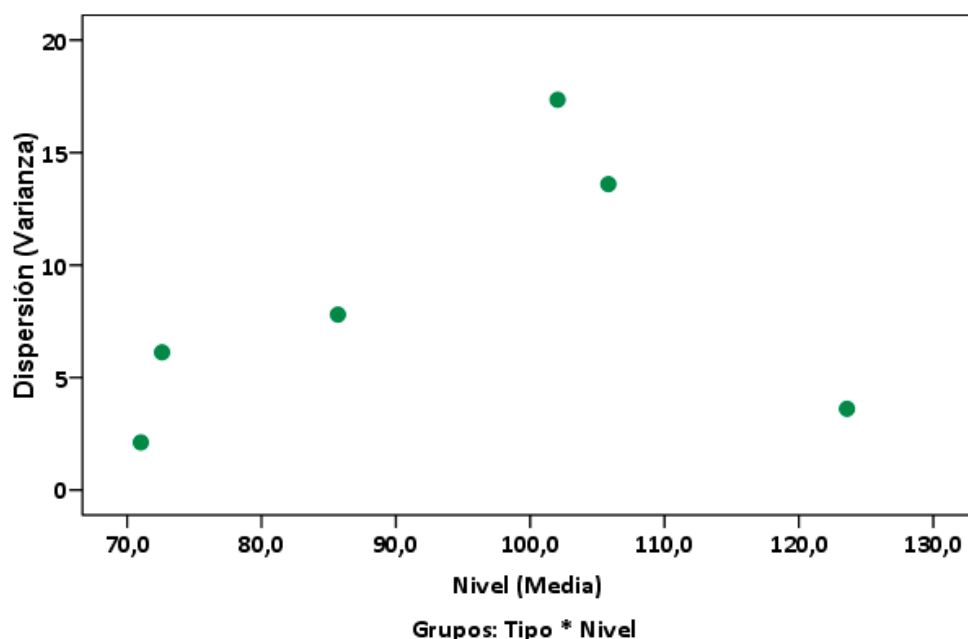
Diagrama de dispersión x nivel: Complementa las pruebas de homogeneidad (contraste de Levene). El diagrama consiste en un gráfico de puntos, cada uno representa una población (combinación de niveles) de modo que la abscisa es la media y la ordenada la desviación típica o la varianza. El objetivo es destacar si la varianza depende de la media (es frecuente comprobar que la respuesta crezca cuando lo hace la media).

Diagramas de dispersión por nivel de Crecimiento



Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

Diagramas de dispersión por nivel de Crecimiento



En los gráficos de (media, desviación típica) y (media, varianza) se pone de manifiesto que las medidas de dispersión dependen de la media.

El gráfico 1 muestra que los niveles con mayor y menor crecimiento medio conducen a desviaciones típicas muy semejantes y, aparentemente, no parece claro que la dispersión de ambos gráficos pueda achacarse a una relación determinista, lo que concuerda con el hecho de que la Prueba de Levene no encuentra significación estadística.

Variable dependiente: Crecimiento

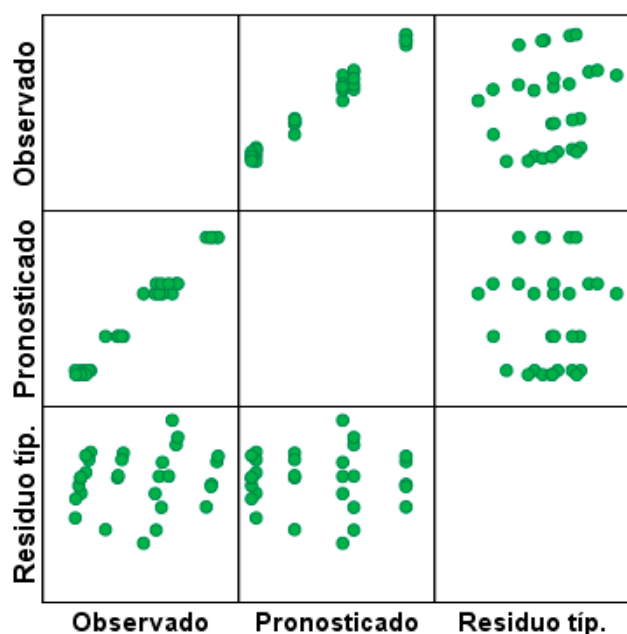


Gráfico de los residuos es un gráfico matricial de puntos enfrentando Valores Observados (variable respuesta) con Valores Pronosticados (por el método estimado) y los Residuos (valores observados menos valores pronosticados).

Los residuos se muestran tipificados (divididos por la raíz cuadrada del error cuadrático medio estimado).

El gráfico matricial da una idea de si se verifican los supuestos de independencia, homogeneidad de varianzas y falta de ajuste.



Asignatura..... Grupo.....
 Apellidos..... Nombre.....
 Ejercicio del día.....

Independencia: El gráfico de dispersión Residuos-Observados no debería mostrar una pauta de variación sistemática (línea, curva, ...)

Homoscedasticidad: El gráfico de dispersión Residuos-Pronosticados debería mostrar una dispersión de residuos similar para todos los valores pronosticados.

Ajuste de los datos: El gráfico de dispersión Pronosticados-Observados debería mostrar un perfil cercano a la linealidad, cuánto mas lineal mejor ajuste.

Residuos-Observados: Gráfico de coordenadas (1, 3) o (3, 1) en la matriz, se dispersa formando un círculo, redundando en su aleatoriedad, corroborando la hipótesis de independencia.

Residuos-Pronosticados: Gráfico de coordenadas (2, 3) o (3, 2) en la matriz, muestra una dispersión semejante para cada valor pronosticado. Algo mayor para las medias pronosticadas en lugar 4 y 5 en orden creciente, pero sin relevancia aparente.

Pronosticados-Observados: Gráfico de coordenadas (1, 2) o (2, 1) en la matriz, se ajusta a una relación lineal, ratificando el buen nivel de ajuste del modelo estimado.

Asignatura..... Grupo.....
Apellidos..... Nombre.....
Ejercicio del día.....

