INTRODUCCIÓN A LA INFERENCIA ESTADÍSTICA

Ahora que conocemos las ideas básicas sobre probabilidad y distribuciones estadísticas, estamos en condiciones de acercarnos más al uso real de estas herramientas para conocer distintas situaciones objeto de estudio y para tomar decisiones fundamentadas.

Recordemos algunas de las definiciones básicas en estadística:

> **Población**: conjunto de elementos del que queremos conocer alguna característica.
> **Muestra**: parte de la población que estudiamos.
> **Individuo**: cada uno de los elementos de la población o de la muestra.

Lo ideal sería, como en estadística descriptiva, poder estudiar a toda la población (**censo**), y obtener los **parámetros** de forma exacta, pero no suele ser posible (es muy caro, o difícil, o directamente imposible porque se destruye el objeto estudiado, en un control de calidad de cristales, por ejemplo). Por eso **es necesario recurrir a muestras** y obtener medidas estadísticas para aproximarnos a los parámetros. En esto consiste la **inferencia estadística**: ***obtención de conclusiones para toda la población a través del estudio de una muestra***.

Tipos de muestreo

- **Aleatorio** vs **no aleatorio**, según si cada individuo tiene la misma probabilidad de ser seleccionado en una muestra, o no. **Con reposición** o **sin reposición**, dependiendo de si un individuo puede ser elegido más de una vez en la muestra o no.
  Dentro del aleatorio, distinguimos:

  - **Muestreo aleatorio simple**: se eligen aleatoriamente todos los individuos.
  - **Muestreo sistemático**: se elige al azar el primer individuo entre los k primeros, y a partir de él, con la población ordenada, se eligen los demás de k en k.
  - **Muestreo estratificado**: se divide la población en grupos **homogéneos entre sí**, y se hace un muestreo aleatorio simple en cada grupos (estrato). Puede ser con **afijación igual** si hay el mismo número de individuos de cada estrato, **o proporcional**, si el tamaño de la muestra de cada estrato conserva la proporción del estrato en la población.

    Ejemplo: para entenderlo mejor, supongamos que estudiamos precios de hoteles en una provincia, y hay 1000 de dos estrellas, 500 de 3 estrellas, 50 de 4 estrellas y 10 de 5 estrellas (estratos). Tomamos una muestra de tamaño 40. No parece lógico incluir los 10 hoteles de 5 estrellas que solo suponen un pequeño porcentaje del total, en el muestreo proporcional habría más hoteles de dos y menos o ninguno de 5 estrellas.

  - **Muestreo por conglomerados**: se divide a la población en grupos heterogéneos entre sí y se realiza un muestreo aleatorio entre esos grupos.

    Ejemplo: Hacemos un estudio sobre miopía en los centros educativos de Ferrol. No podemos visitar todos, así que los dividimos por barrios. Dentro de cada barrio, elegimos un centro o una muestra de centros. Así cubrimos todas las zonas de Ferrol sin necesidad de visitar todos los centros. Los conglomerados son heterogéneos, ya que la zona no debe afectar a si hay más o menos miopes entre el alumnado.

PROBLEMAS 1
CONOCIDA LA POBLACIÓN, RESPONDEMOS CUESTIONES SOBRE LAS MUESTRAS

POBLACIÓN: Tenemos una población de **media** μ y **desviación típica** σ**.**

CARACTERÍSTICA: Algo que estudiamos y que está en una proporción p en la población.

MEDIA MUESTRAL $\overline{X}$**:** variable aleatoria que vincula cada muestra a su media.

PROPORCIÓN MUESTRAL $\hat{P}$: "" "" "" "" que vincula cada muestra a su proporción. (referida la proporción a la característica estudiada)

DISTRIBUCIÓN DE LA MEDIA MUESTRAL

(condiciones: la población es normal, o $n > 30$)

La distribución de la media muestral es $\overline{X} \in N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right)$

Nota: consecuencia del llamado Teorema Central del Límite, en el que no profundizamos.

Ejemplo: Se sabe que el gasto semanal de los universitarios en Santiago tiene una media de 25€ y una desviación típica de 3€. ¿Cuál es la probabilidade de que, elegidos al azar 49 universitarios, su gasto medio esté comprendido entre 24€ y 26€?

La muestra es aleatoria.Como la distribución no es normal, comprobamos $n = 49 > 30$. Por lo tanto, la media muestral en los 49 estudiantes es

$$\overline{X} \in N\left(25, \frac{3}{\sqrt{49}}\right) = N\left(25, \frac{3}{7}\right)$$

Resolvermos el problema tal y como hacemos en distribución normal:

$$P(24 \leq \overline{X} \leq 26) \overset{tipif.}{=} P(-2{,}33 \leq Z \leq 2{,}33) = P(Z \leq 2{,}33) - P(Z \leq -2{,}33) =$$
$$= 2P(Z \leq 2{,}33) - 1 = \mathbf{0{,}9802}$$

DISTRIBUCIÓN DE LA PROPORCIÓN MUESTRAL

(condiciones: $n > 30$ o $np > 5$ y $nq > 5$)

La distribución de la proporción muestral es $\hat{P} \in N\left(p, \sqrt{\dfrac{pq}{n}}\right)$

Ejemplo: Una de cada diez bolsas de gusanitos de una marca tiene menos peso del etiquetado. En una muestra al azar de 400 bolsas, ¿cuál es la probabilidad de que haya más de 50 bolsas con menos peso del etiquetado?

La muestra es aleatoria y $n = 49 > 30$. La proporción es $p = 0{,}1$ ($q = 0{,}9$).

La proporción muestral es $\hat{P} \in N\left(0{,}1; \sqrt{\dfrac{0{,}1 \cdot 0{,}9}{400}}\right) = N(0{,}1; 0{,}015)$

$$P(\hat{P} > 50/400) \overset{tipif.}{=} P(Z > 1{,}67) = 1 - P(Z < 1{,}67) = \mathbf{0{,}0475}$$

PROBLEMAS 2
CONOCIDA UNA MUESTRA, RESPONDEMOS CUESTIONES SOBRE LA POBLACIÓN

ESTIMACIÓN PUNTUAL: Llamamos **estimador** a un valor obtenido a través de una muestra para aproximarnos a un parámetro en la población (como la media, o la proporción).

ESTIMADOR PARA LA MEDIA EN UNA $N(\mu, \sigma)$: media muestral $\overline{x}$

ESTIMADOR PARA LA PROPORCIÓN EN UNA $B(n, p)$**:** proporción muestral $\hat{p}$

Ejemplo: Se realiza un estudio en el instituto sobre el número de miembros en la unidad familiar, a través de una muestra de 10 estudiantes. Los datos son: 3, 3, 5, 4, 6, 2, 3, 4, 7, 2.
Determina un estimador puntual para la media en la población, y para la proporción de hogares con más de 3 miembros en la unidad familiar.

La población serían las unidades familiares de todo el instituto. Para estimar la media del número de miembros de la población usaríamos la media muestral:

$\overline{x} = \dfrac{3+3+5+4+6+2+3+4+7+2}{10} = 3{,}9$ (consideramos 3,9, casi 4 la media)

La proporción de hogares con más de 3 miembros en la población la estimamos a través de la porporción muestral. En la muestra hay 4 familias de 10 con más de 3 miembros, luego:

$\hat{p} = 0{,}4$ suponemos que el 40% de las familias en el instituto tienen más de 3 miembros.

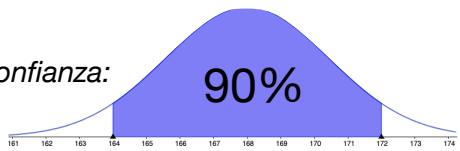
ESTIMACIÓN POR INTERVALOS DE CONFIANZA: La estimación puntual es muy limitada. Dos muestras distintas van a tener medias y proporciones distintas, en general. Una forma más fiable de poder tener una aproximación de la media o de otros parámetros de la población es a través de lo que llamamos **intervalos de confianza**.

Un **intervalo de confianza** es un intervalo que contiene al parámetro desconocido con una probabilidad (confianza) determinada previamente. El **error máximo admisible** es el radio de dicho intervalo, y nos indica el error máximo que cometemos con esa confianza si consideramos el parámetro como el centro del intervalo.

Ejemplo: En una muestra de 50 estudiantes, obtenemos el siguiente intervalo, con un nivel del confianza del 90%, para la altura media: (164,172). Interpreta el resultado e indica el error máximo admisible.

La probabilidad de que la media de la población esté entre 164 cm y 172 cm es del 90 %.

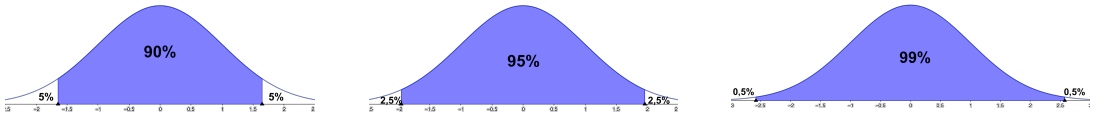
Zona probable para la media poblacional con un 90% de confianza:



El centro del intervalo es 168 cm. Si consideramos 168 cm la altura media de la población, estaremos cometiendo un error máximo de 4 cm, con una confianza del 90%.

INTERVALOS CARACTERÍSTICOS

Nos interesa conocer, por tanto, los **intervalos característicos**: aquellos que encierran un tanto por ciento de la población determinado para una distribución normal $N(0,1)$:



Definimos algunos conceptos necesarios:

α: **nivel de significación** (o riesgo). Es la probabilidad excluída

$1-\alpha$: **nivel de confianza**. Es la probabilidad incluída en el intervalo.

$z_{\frac{\alpha}{2}}$: **valor crítico**. Representa los extremos del intervalo que encierra el porcentaje deseado de la población. $[-z_{\frac{\alpha}{2}}, z_{\frac{\alpha}{2}}]$. Nuestro objetivo es conocer este valor crítico.

Ejemplo: cálculo del intervalo que encierra al 90% de la población en $N(0,1)$.

Nivel de confianza: $1-\alpha = 0{,}9$ Nivel de significación: $\alpha = 0{,}1 \implies \alpha/2 = 0{,}05$

Resolvermos el ejercicio tal y como hacemos siempre con la distribución normal:

$$P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 0{,}9 \ :$$

$$P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = P(Z \leq z_{\frac{\alpha}{2}}) - P(Z \leq -z_{\frac{\alpha}{2}}) \overset{sim.}{=} P(Z \leq z_{\frac{\alpha}{2}}) - P(Z \geq z_{\frac{\alpha}{2}}) \overset{cont.}{=}$$

$$= P(Z \leq z_{\frac{\alpha}{2}}) - (1 - P(Z \leq z_{\frac{\alpha}{2}})) = 2P(Z \leq z_{\frac{\alpha}{2}}) - 1 = 0{,}9 \implies$$

$$\implies P(Z \leq z_{\frac{\alpha}{2}}) = \frac{1{,}9}{2} = 0{,}95.$$

Buscamos 0,95 en la tabla $N(0,1)$:

Y así obtenemos el valor $z_{\frac{\alpha}{2}} = z_{0{,}05} = \mathbf{1{,}645}$

Los demás los obtendríamos igual:

%	$1-\alpha$	$\alpha/2$	$z_{\alpha/2}$
90 %	0,90	0,05	**1,645**
95 %	0,95	0,025	**1,96**
98 %	0,98	0,01	**2,325**
99 %	0,99	0,005	**2,575**

x	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767

INTERVALOS DE CONFIANZA PARA LA MEDIA

Dada una población con distribución $N(\mu, \sigma)$ con σ conocida, el **intervalo de confianza para la media**, con confianza $1-\alpha$, para una muestra de tamaño n, con media muestral $\overline{x}$, es:

$$\left(\overline{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

Error máximo admisible: $E = z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$

NOTA: Si la muestra es más grande, el error máximo admisible es más pequeño y mejora la estimación.Si la confianza es mayor (el riesgo admisible menor) el error máximo admisible es más grande y la estimación es menos precisa (si queremos tener más seguridad de que no nos equivocamos estimando una media, tenemos que sacrificar precisión y admitir un rango de valores más amplio)

Ejemplo: estudiamos el gasto mensual en café en una muestra de 150 adultos, y observamos un gasto medio de 22€. El consumo sigue una distribución normal de desviación típica de 6€. Determina un intervalo de confianza para la media con un nivel de confianza del 95%.

Para $1-\alpha = 0{,}95$, vimos que $z_{\frac{\alpha}{2}} = z_{0{,}025} = 1{,}96$. Sabemos que $\overline{x} = 22$ y $\sigma = 6$.

Intervalo de confianza: $\left(22 - 1{,}96\frac{6}{\sqrt{150}},\ 22 + 1{,}96\frac{6}{\sqrt{150}}\right) = (\mathbf{21,04}; \mathbf{22,96})$

El error máximo admisible sería $E = 1{,}96\frac{6}{\sqrt{150}} = 0{,}96$

Ejemplo: seguimos con el mismo estudio. ¿Qué tamaño de la muestra sería necesario para estudiar el gasto medio con un error de menos de 1€ y un nivel de confianza del 90%?

Para $1-\alpha = 0{,}90$, vimos que $z_{\frac{\alpha}{2}} = z_{0{,}05} = 1{,}645$.

El error máximo admisible sería $1 = 1{,}645\frac{6}{\sqrt{n}} \implies n = \left(\frac{1{,}645 \cdot 6}{1}\right)^2 = 97{,}42$

Por lo tanto la muestra tendría que tener al menos 98 individuos para que el error con ese nivel de confianza sea menor de 1€.

INTERVALOS DE CONFIANZA PARA LA PROPORCIÓN

Dada una población en la que estudiamos si cumplen, o no, una característica, el **intervalo de confianza para la proporción**, con confianza $1-\alpha$, para una muestra de tamaño n, con proporción muestral $\hat{p}$ ($\hat{q}=1-\hat{p}$), es:

$$\left(\hat{p}-z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}},\hat{p}+z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

Error máximo admisible: $E=z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}\hat{q}}{n}}$

NOTA: Si la muestra es más grande, el error máximo admisible es más pequeño y mejora la estimación.Si la confianza es mayor (el riesgo admisible menor) el error máximo admisible es más grande y la estimación es menos precisa (si queremos tener más seguridad de que no nos equivocamos estimando una media, tenemos que sacrificar precisión y admitir un rango de valores más amplio)

Ejemplo: para saber la proporción de hipermétropes en una ciudad, con una muestra de 600 habitantes, se observa que 36 lo son. Determina el intervalo de confianza correspondiente con un nivel de confianza del 99%.

Para $1-\alpha=0{,}99$, vimos que $z_{\frac{\alpha}{2}}=z_{0{,}005}=2{,}575$. Sabemos que $\hat{p}=36/600=0{,}06$.

Intervalo de confianza: $\left(0{,}06-2{,}575\sqrt{\frac{0{,}06\cdot 0{,}94}{600}},\ 0{,}06+2{,}575\sqrt{\frac{0{,}06\cdot 0{,}94}{600}}\right)=(\mathbf{0},\mathbf{034};\mathbf{0},\mathbf{085})$

El error máximo admisible sería $E=2{,}575\sqrt{\frac{0{,}06\cdot 0{,}94}{600}}=0{,}03$

Ejemplo: seguimos con el mismo estudio. ¿Qué tamaño de la muestra sería necesario para conocer la proporción media con un error máximo del 1% y un nivel de significación del 0,05?

Para $1-\alpha=0{,}95$, vimos que $z_{\frac{\alpha}{2}}=z_{0{,}025}=1{,}96$.

Error máximo admisible: $0{,}01=1{,}96\sqrt{\frac{0{,}06\cdot 0{,}94}{n}}\implies n=\frac{1{,}96^2\cdot 0{,}06\cdot 0{,}94}{0{,}01^2}=2166{,}66$

Por lo tanto la muestra tendría que tener al menos 2167 individuos para que el error con ese nivel de confianza (95%) sea menor del 1%.