1. ¿QUÉ ES LA INFERENCIA ESTADÍSTICA?

La **inferencia estadística** busca obtener resultados estadísticos para una población objeto de estudia a partir de estudiar una muestra de dicha población. Es decir, está pensada cuando estudiar a toda la población es demasiado costoso o requiere de tiempo infinito.

2. DISTRIBUCIÓN DE LA MEDIA MUESTRAL

Dada una población en un estudio estadístico (pesos, alturas,…) cuyo tamaño es excesivamente grande, supongamos que tomamos una muestra de tamaño n. Se obtiene una media con los datos $\overline{x_1}$ de la variable que se estudia (pesos, alturas,…). SI tomamos otra muestra, se obtiene otra media $\overline{x_2}$ y así sucesivamente.
De esta forma tomando todas las posibles medias de cada uno de las posibles muestras de la población que se pueda realizar, la variable aleatoria que a cada muestra asocia una media se llama **media muestral ($\bar{X}$).** Esta variable tomará los valores:

$$\bar{X} = \{\overline{x_1}, \overline{x_2}, \dots., \overline{x_n}\}$$

De esta variable $\bar{X}$ se puede estudiar su **distribución de la media muestral** para ello es necesario saber:

- La media muestral que se calcula va a coincidir con la media de la población:
  $\mu_{\bar{x}} = \mu$

- La desviación típica muestral es igual a la desviación típica de la población dividida por la raíz cuadrada del tamaño de la muestra, es decir: $\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{n}}$

- Si la distribución de partida sigue una distribución $N(\mu, \sigma)$, la distribución de la muestra de tamaño n sigue una normal: $N(\mu, \frac{\sigma}{\sqrt{n}})$-

Si la distribución de partida no se distribuye normalmente, tenemos el siguiente teorema consecuencia del teorema central del límite:

Teorema:

Si X es una variable estadística y consideramos la distribución muestral de medias con n suficientemente grande $(n \geq 30)$, entonces la distribución muestral de medias se aproxima a una distribución normal y podemos considerar que $\bar{X} \in N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

OBSERVACIÓN: Nosotros en este curso solo trabajaremos con distribuciones normales o con tamaño muestral grande $(n \geq 30)$, por lo que siempre podremos considerar la distribución muestral de medias como una normal.

Ejemplo: El gasto total semanal de los jóvenes de una ciudad tiene una media de 25€ y una desviación típica de 3€ ¿Cuál es la probabilidad de que el gasto medio de 49 jóvenes, elegidos al azar, esté comprendido entre 24 y 26€?

Dado que el tamaño de la muestra es > 30 la distribución de las medias es $N(25,\frac{3}{\sqrt{49}})=$

$=N(25,3/7)$; $P(24<\overline{X}<26)=P(\frac{24-25}{3/7}<Z<\frac{26-25}{3/7})=P(-2'33<Z<2'33)=P(Z<2'33)-(1-P(Z<2'33))=$
$2.P(Z<2'33)-1=0'9802$

3. DISTRIBUCIÓN DE LA PROPORCIÓN MUESTRAL

De forma similar a lo que hemos visto con las medias, podemos considerar todas las muestras de tamaño n posibles de una población con una distribución binomial con probabilidad de éxito p, y considerar las proporciones muestrales $\hat{p}$ como una variable aleatoria. A esta variable $\hat{p}$ le llamamos distribución muestral de proporciones.

Esta variable $\hat{p}$ cumple:

- La media o esperanza de $\hat{p}$ coincide con la media poblacional, es decir, $\mu_{\hat{p}} = p$.
- La desviación típica de $\hat{p}$ coincide con $\sqrt{\frac{p\cdot q}{n}}$, es decir, $\sigma_{\hat{p}} = \sqrt{\frac{p\cdot q}{n}}$

En general tenemos que:

Si X es una variable binomial y consideramos la distribución muestral de proporciones con n suficientemente grande $(n \geq 30)$ y $n \cdot p \geq 5$ y $n \cdot q \geq 5$. Entonces la distribución muestral de proporciones se aproxima a una distribución normal: $\hat{p} \in N\left(p, \sqrt{\frac{p\cdot q}{n}}\right)$.

OBSERVACIÓN: Aunque es poco habitual si en algún ejercicio no tuviéramos el parámetro p, lo aproximaremos por la proporción muestral siempre que n sea suficientemente grande $(n \geq 100)$.

Ejemplo:
El 10% de las bolsas de pipas de una marca contiene menos peso del que anuncia. Se han seleccionado al azar 400 bolsas a) ¿Cuál es la distribución que sigue la proporción de envases no completos de la muestra?; b) Halla la probabilidad de que en la muestra haya más de 50 bolsas de pipas con menos peso del anunciado.

A) n=400>30, p=10%=0'1 $\hat{P}$ sigue una distribución $N(0'1, \sqrt{\frac{0'1.0'9}{400}})=N(0'1, 0'015)$

B) $P(\text{más de 50 bolsas})=P(\hat{P}>50/400)=P(\hat{P}>0'125)=P(Z>\frac{0'125-0'1}{0'015})=P(Z>1'67)=$
$=1-P(Z<1'67)=1-0'9525=0'0475$. Es decir, hay una probabilidad de 4'75% de que en la muestra haya más de 50 bolsas con peso inadecuado.

4. ESTIMACIÓN DE PARÁMETROS

Denominamos **estimación de un parámetro** al valor que se calcula de los datos muestral y que nos dan información sobre ese parámetro de la población. Por ejemplo: la media muestral es un estimador del parámetro media de la población; la proporción de una muestra es un estimador de toda la población.
Existen dos formas distintas de realizar una estimación de un parámetro: estimación puntual o estimación por intervalos.

4.1. ESTIMACIÓN PUNTUAL

La **estimación puntual** consiste en asignar a la población el mismo resultado obtenido al calcular el parámetro en la muestra.

EJEMPLO 1:
Deseamos calcular la duración media de las baterías fabricadas por una empresa, para ello tomamos una muestra de 150 baterias y observamos su duración. Se tiene que la media muestral es de 8277 horas de uso en esta muestra.
Por estimación puntual, podemos concluir que 8277 horas es la duración media de todas las baterías de la fábrica.

EJEMPLO 2:
Deseamos conocer la proporción de conductores que superan la velocidad máxima permitida en un punto de la autopista. Tomamos una muestra de 300 conductores y se tiene que el 22% superan la velocidad.
De esta forma se estima que el 22% de todos los conductores que pasan por esa autopista superan el límite de velocidad en ese punto.

4.2. ESTIMACIÓN POR INTERVALOS DE CONFIANZA

Normalmente no se hace una estimación puntual de un parámetro, sino más bien una estimación de entre que dos valores se puede encontrar un cierto parámetro con una probabilidad dada, esto es lo que se conoce como estimación por intervalos de confianza.

- Llamamos Intervalo de confianza al intervalo que con un cierto nivel de confianza, contiene al parámetro que se está estimando.

- Nivel de confianza es la "probabilidad" de que el intervalo calculado contenga al verdadero valor del parámetro. Se indica por $1-\alpha$ y habitualmente se da en porcentaje $(1-\alpha)100\%$.

Hablamos de nivel de confianza y no de probabilidad ya que una vez extraída la muestra, el intervalo de confianza contendrá al verdadero valor del parámetro o no, lo que sabemos es que si repitiésemos el proceso con muchas muestras podríamos afirmar que el $(1-\alpha)\%$ de los intervalos así construidos contendría al verdadero valor del parámetro.

- Nivel de significación: es el valor de α representa la probabilidad de equivocarnos al estimar el parámetro.

INTERVALO DE CONFIANZA PARA LA MEDIA

Vamos a ver a continuación el intervalo de confianza para la media si la población sigue una distribución normal $N(\mu,\sigma)$, de la que conocemos la desviación típica σ.

Expresión del **intervalo de confianza para la media**: $\left(\bar{x}-Z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}},\bar{x}+Z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}\right)$

$$donde\begin{cases}\bar{x}\ es\ la\ media\ muestral\\ \sigma\ \ es\ la\ desviación\ típica\\ n\ es\ el\ tamaño\ de\ la\ muestra\\ z\alpha_{/2}\ es\ al\ valor\ crítico\ (se\ calcula\ en\ la\ tabla\ de\ la\ N(0,1))\end{cases}$$

Ejemplo:

Las estaturas de una muestra aleatoria de 50 estudiantes tienen una media de 174'5 cm, y se conoce que la desviación típica de la variable estatura es de 6'9 cm. Calcula un intervalo de confianza del 95% para la estatura media de todos los estudiantes.

Recordemos que el intervalo de confianza para la media poblacional es: : $\left(\bar{x}-Z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}},\bar{x}+Z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}\right)$

$\bar{x}=174{,}5$

$\sigma=6{,}9$

$\sqrt{n}=\sqrt{50}$

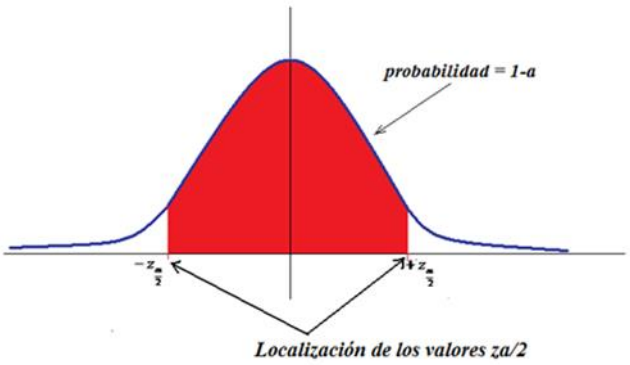
Tenemos que calcular $z\alpha_{/2}$ (*)

Nivel de confianza 95%$\Rightarrow 1-\alpha=0{,}95\Rightarrow\alpha=0{,}05\Rightarrow\frac{\alpha}{2}=0{,}025\Rightarrow z\alpha_{/2}=1{,}96.$

$$Para\ calcularlo\ busco\ en\ la\ tabla\ de\ la\ N(0,1): P\left(Z\leq z\alpha_{/2}\right)=1-0{,}025=0{,}975$$

Entonces el intervalo de confianza : $\left(174{,}5-1{,}96\cdot\frac{6{,}9}{\sqrt{50}},174{,}5+1{,}96\cdot\frac{6{,}9}{\sqrt{50}}\right)=(172{,}59,176{,}41)$

*



Por ejemplo; para un nivel de confianza del 95%, $1-\alpha = 0.95$
Tendríamos que observar el valor $z_{\frac{\alpha}{2}}$
que deja a su izquierda un barrido de $1-\left(\frac{0.05}{2}\right) = 0.975$ *en este caso*
correspondería a $z_{\frac{\alpha}{2}} = 1.96$

INTERVALO DE CONFIANZA PARA LA PROPORCIÓN

La expresión es la siguiente:

$$(\hat{p} - Z_{\frac{\alpha}{2}}.\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}}, \hat{p} + Z_{\frac{\alpha}{2}}.\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}})$$

$$donde \begin{cases} \hat{p}\ es\ la\ proporción\ muestral \\ \hat{q} = 1-\hat{p} \\ n\ es\ el\ tamaño\ de\ la\ muestra \\ z\alpha_{/2}\ es\ al\ valor\ crítico, cumple\ que\ P\left(-z\alpha_{/2} \le Z \le z\alpha_{/2}\right) = 1-\alpha \end{cases}$$

Ejemplos

1. Para saber qué proporción de alumnos de la ESO tienen teléfono móvil con conexión de datos se selecciona una muestra de 500 alumnos, de ellos contestan afirmativamente 225. ¿Cuál es el intervalo de confianza para la proporción de los alumnos que tienen móvil con conexión de datos, con un nivel de confianza del 95%?

Solución

En primer lugar hemos de hallar la proporción en la muestra $\hat{p}$= 225/500=0'45; por tanto $\hat{q}$=1-0'45=0'55; n=500

Ahora tenemos que calcular el valor crítico $z_{\alpha/2}$ para un nivel de confianza 1-α=0'95;

$P(Z<z_{\alpha/2}) = 1-\frac{\alpha}{2} = 1-\frac{0'05}{2} = 1-0'025 = 0'975$ Buscando en las tablas hallamos $z_{\alpha/2}$=1'96

Por tanto el intervalo de confianza para la proporción que nos piden será:

$I_c = (\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{p}.\hat{q}}{n}}, \hat{P} + z_{\alpha/2}.\sqrt{\frac{\hat{p}.\hat{q}}{n}}) = (0'45-1'96.\sqrt{\frac{0'45.0'55}{500}}, 0'45+1'96.\sqrt{\frac{0'45.0'55}{500}}) =$

=(0'4064, 0'4936)

ERROR Y TAMAÑO DE LA MUESTRA

Error del Intervalo de confianza para la media

La precisión del intervalo de confianza anterior $(\bar{x} - Z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\frac{\alpha}{2}}.\frac{\sigma}{\sqrt{n}})$ es $z\alpha_{/2}.\frac{\sigma}{\sqrt{n}}$

Esto significa que al utilizar $\bar{x}$ para estimar μ cometemos un error E que es menor o igual a $z\alpha_{/2}.\frac{\sigma}{\sqrt{n}}$

con una confianza del $(1-\alpha)\cdot 100$ por ciento: $E = z\alpha_{/2}.\frac{\sigma}{\sqrt{n}}$.

El valor de $E = z\alpha_{/2}.\frac{\sigma}{\sqrt{n}}$ depende de α y n, del siguiente modo:

- Cuanto mayor sea el tamaño de la muestra menos es el error.
- Cuanto mayor sea $(1-\alpha)$ (es decir, cuanto más seguros queramos estar de nuestra estimación), mayor es E.

En situaciones donde se puede controlar el tamaño de la muestra, es posible elegir n de forma que se tenga una confianza del $(1-\alpha)\cdot 100$ por ciento de que el error al estimar μ sea menor que el error especificado E. Despejando en la fórmula del error tenemos la fórmula para n:

$$n = \left(\frac{z\alpha_{/2}\cdot\sigma}{E}\right)^2$$

Si n no sale un número entero lo redondeamos al entero de sumar uno a su parte entera.

Error del Intervalo de confianza para la proporción

Error máximo para la proporción: $E = z\alpha_{/2}.\sqrt{\frac{\hat{p}\cdot\hat{q}}{n}}$

Tamaño de la muestra:

$$n = \left(\frac{z\alpha_{/2}}{E}\right)^2 \cdot \hat{p}\cdot\hat{q}$$

Ejemplos:

1.En una población, una variable aleatoria sigue una ley normal de media desconocida y desviación típica 9. ¿De qué tamaño, debe ser la muestra con la cual se estime la media poblacional con un nivel de confianza del 97% y un error máximo admisible igual a 3?

$$n = \left(\frac{z\alpha_{/2}\cdot\sigma}{E}\right)^2 \Rightarrow n = \left(\frac{2{,}17\cdot 9}{3}\right)^2 = 42{,}3081$$

El tamaño de la muestra debe ser como mínimo 43.

$$\left(1-\propto = 0{,}97 \Rightarrow \alpha = 0{,}03 \Rightarrow \frac{\alpha}{2} = 0{,}015 \Rightarrow P(Z \leq z\alpha_{/2}\right) = 0{,}985 \Rightarrow z\alpha_{/2} = 2{,}17$$

2. En una encuesta hecha por los alumnos y alumnas de un Instituto a un total de 100 votantes elegido al azar en su Ayuntamiento, se indica que el 55% volvería a votar por el alcalde actual.
   a) Calcular un intervalo de confianza al 99% y otro al 99,73% para la proporción de votantes favorables al alcalde actual.
   b)¿Cuáles deben ser los tamaños muestrales en el sondeo para tener, con los mismos niveles de confianza, la certeza de que el alcalde actual salga reelegido por mayoría absoluta, en el caso de arrojar la encuesta los mismos resultados?

Solución

N=100; $\hat{p} = 0'55$; $\hat{q} = 0'45$

a) Para calcular el intervalo de confianza para el 99% hallamos primero $z_{\alpha/2}$ para el nivel de confianza 1-α=0'99;

$P(Z<z_{\alpha/2}) = 1-\frac{\alpha}{2} = 1 - \frac{0'01}{2} = 1 - 0'005 = 0'995$ Buscando en las tablas hallamos $z_{\alpha/2}$=2'57

Por tanto el intervalo de confianza para la proporción que nos piden será:

$I_c = (\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{p}.\hat{q}}{n}}, \hat{P} + z_{\alpha/2}.\sqrt{\frac{\hat{p}.\hat{q}}{n}}) = (0'55-2'57.\sqrt{\frac{0'55.0'45}{100}}, 0'55+2'57.\sqrt{\frac{0'55.0'45}{100}}) =$

=(0'422, 0'677)

Para un nivel de confianza del 99'73%: $P(Z<z_{\alpha/2}) = 1-\frac{\alpha}{2} = 1 - \frac{0'0027}{2} = 1 - 0'00135 = 0'99865$ Buscando en las tablas hallamos $z_{\alpha/2}$=2'98

Por tanto el intervalo de confianza para la proporción que nos piden será:

$I_c = (\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{p}.\hat{q}}{n}}, \hat{P} + z_{\alpha/2}.\sqrt{\frac{\hat{p}.\hat{q}}{n}}) = (0'55-2'98.\sqrt{\frac{0'55.0'45}{100}}, 0'55+2'98.\sqrt{\frac{0'55.0'45}{100}}) =$

=(0'401, 0'698)

B) Nos piden el valor de n condicionado a que todos los valores del intervalo de confianza sean superiores a 0,5 (mayoría absoluta), es decir que, dado que la media muestral es 0,55, el radio del intervalo ha de ser necesariamente menor que 0,55-0,5 = 0,05.

Por tanto, en el caso del nivel de confianza del 99% $E = z_{\alpha/2}.\sqrt{\frac{\hat{p}.\hat{q}}{n}} = 2'57.\sqrt{\frac{0'55.0'45}{n}} < 0'05$, despejando n resulta n>653'8, es decir hemos de tomar una muestra de al menos 654 personas.

En el caso del 99'73% se hace igual resultando n>891 personas.