Muestreo. Tipos de muestreo. Inferencia

Introducción

Nota.- Puede decirse que la **Estadística** es la ciencia que se preocupa de la recogida de datos, su organización y análisis, así como de las predicciones que, a partir de esos datos pueden hacerse Los aspectos anteriores hacen que pueda hablarse de dos tipos de **Estadística: Descriptiva** e **Inferencial**

Nota.- La **Estadística Descriptiva** se ocupa de tomar los datos de un conjunto dado, organizarlos en tablas o representaciones gráficas y del cálculo de unos números que nos informen de manera global del conjunto estudiado. *No utiliza la Probabilidad*

Nota.- La **Estadística Inferencial** trata sobre la elaboración de conclusiones para una población, partiendo de los resultados de una muestra y del grado de fiabilidad de las conclusiones. *Utiliza la Probabilidad*

Nota.- Nos dedicamos ya a la **Estadística Inferencial.**

Nota.- Cuando hay que hacer un estudio estadístico sobre una población, lo más habitual es que no se pueda acceder a todos los individuos que la componen; es necesario, entonces, *elegir una muestra, realizar el estudio sobre ella y después intentar extrapolar los datos a toda la población en general*. La muestra se tiene que elegir de manera que sea lo más representativa posible. El proceso que sigamos para la extracción de la muestra se *denomina muestreo*.

Estadística inferencial. Muestreo

Def.- La **Estadística Inferencial** se ocupa de inferir o deducir las características de la población a partir de las características de una muestra

Nota.- Existen dos formas de hacer Estadística Inferencial:

.La estimación de parámetros.

.Las pruebas de hipótesis.

En esta lección nos vamos a ocupar de la estimación de parámetros y en la siguiente de las pruebas de hipótesis.

Def.- Los **parámetros poblacionales o parámetros** son los índices centrales y de dispersión que definen a una población (media, varianza, proporción..).

Def.- Los **estadísticos muestrales o estadísticos** son los índices centrales y de dispersión que definen a una muestra (media, varianza, proporción.. muestrales).

Nota.- En la inferencia estadística es necesario utilizar muestras, que representen a la población. Esto se consigue mediante las **técnicas de muestreo**.

Tipos de muestreo

Nota.- En los muestreos hay que ver si hay o nó reemplazamiento, y si hay o nó aleato-

riedad. Según esto tenemos los siguientes conceptos.
En la Comunidad Andaluza, se consideran que los muestreos son aleatorios y con reemplazamiento. Desde el punto de vista teórico de la inferencia estadística, una población finita en la que los individuos son elegidos son elegidos con reemplazamiento puede ser considerada como infinita.

- **Muestreo con reemplazamiento** es el que se realiza cuando un elemento tomado de la población vuelve de nuevo a ella para poder volver a ser elegido. En esta situación, cada miembro de la población puede seleccionarse más de una vez.
Este tipo de muestreo hace que una población finita pueda ser considerada, al menos en su aspecto teórico, como una población infinita.
- Muestreo sin reemplazamiento es el que se efectúa sin devolver a la población los elementos que se van eligiendo para construir la muestra. En este caso, cada miembro de la población no puede seleccionarse más de una vez.
- Muestreo aleatorio es el que se efectúa teniendo en cuenta que cada miembro de la población tiene la misma probabilidad de ser elegido en la muestra. Con este tipo de muestreo, las muestras son representativas, es posible conocer los posibles errores cometidos y pueden hacerse inferencias estadísticas. (Bombo de lotería).

- En general, llamaremos **N** al tamaño de la población (número de individuos que la componen, en el caso que sea finita) y **n** al de la muestra.

- Muestreo aleatorio estratificado
Es el que se utiliza cuando en la población se pueden distinguir varios colectivos *(estratos)* cuya presencia queremos reflejar en la muestra. Llamaremos N_1, N_2, N_3,… al tamaño de los estratos (con $N_1 + N_2 + N_3 + ,.. = N$), y n_1 , n_2, n_3,.., al número de individuos de los respectivos estratos que hay en la muestra (con $n_1 + n_2 + n_3 + \ldots = n$).
Según el criterio que elijamos para reflejar los estratos en la muestra, tenemos dos subtipos en este muestreo: *con afijación igual* (también llamada constante o simple) y *con afijación proporcional*.
En el caso de *muestreo aleatorio estratificado con afijación igual,* no se toma en cuenta el número de individuos que componen cada estrato, sino que todos tienen la misma presencia en la muestra. Por ejemplo, si hay *5* estratos, de cada uno se elegirían n/5 individuos para la muestra, independientemente del peso que cada uno de ellos tuviera en la población. Es decir, $n_1 = n_2 = n_3 = \ldots = n/5$.

En el caso de *muestreo aleatorio estratificado con afijación proporcional*, sí se toma en cuenta el tamaño de cada estrato. Lo que se pretende es que la muestra mantenga, en su composición, la misma proporción de individuos que cada estrato tenga en la población.

En este caso $\frac{n_1}{N_1} = \frac{n_2}{N_2} = ... = \frac{n_k}{N_k} = \frac{n}{N}$

De $n_1/N_1 = n/N$, obtenemos $n_1 = (N_1.n)/N$

De $n_2/N_2 = n/N$, obtenemos $n = (N_2.n)/N$, y así sucesivamente.

♦♦ En cierta población habitan 1500 niños y jóvenes, 7500 adultos y 1000 ancianos. Se desea realizar un estudio para conocer el tipo de actividades de ocio que se desean incluir en el nuevo parque en construcción. Para ello, van a ser encuestados 200 individuos elegidos al azar.

a) Si se utiliza muestreo estratificado con afijación igual, ¿cuál será el tamaño muestral correspondiente a cada estrato?

b) Si se utiliza muestreo estratificado con afijación proporcional, ¿cuál será el tamaño muestral correspondiente a cada estrato?

sol

El muestreo evidentemente es sin reemplazamiento

(a) En el muestreo estratificado *con afijación igual* dividimos el total de la nuestra entre 3 (niños, adultos, ancianos) y tomamos esa cantidad de cada estrato.

En nuestro caso 200/ 3 = 66'66, como son personas elegimos 66 niños, 67 adultos y 67 viejos, porque 66 + 67 + 67,(la suma tiene que ser 200 y tenemos que aproximar los datos) .

b) En el muestreo estratificado *con afijación proporcional* deben considerarse los estratos formados por niños y jóvenes, adultos y ancianos. El tamaño de cada uno de los estratos debe ser proporcional a la cantidad de individuos de cada uno de ellos. Así, se tiene que:

$$\frac{x}{1500} = \frac{y}{7500} = \frac{z}{1000} = \frac{200}{10000} = \frac{1}{50}$$

$x = \frac{1500}{50} = 30$ niños y jóvenes; $y = \frac{7500}{50} = 150$ adultos; ; $z = \frac{1000}{50} = 20$ ancianos

La muestra debe estar formada por 30 niños y jóvenes, 150 adultos y 20 ancianos elegidos aleatoriamente entre sus respectivos colectivos.

Muestreo aleatorio sistemático

Se suele utilizar para ahorrar costes, y en este tipo de muestreo es necesario ordenar a

los individuos de la población asignándoles de este modo un número ordinal a cada uno. Dividimos **N** (tamaño de la población) entre **n** (tamaño de la muestra), nos da como resultado un nº **h** (llamado *coeficiente de elevación*), y después elegimos, al azar, uno de los h primeros individuos de la población, por ejemplo el que ocupa el lugar k, y a partir de ahí la muestra se iría obteniendo escogiendo individuos de h en h, es decir: k, k + h, k + 2h, k + 3h,..., k + (n-1)h.

Distribución de las medias, proporciones muestrales y diferencias de medias.

Nota.- Una vez obtenida la muestra de la población, y realizado el estudio sobre ella, llega la fase en que hay que obtener conclusiones sobre toda la población. Nosotros vamos a *estimar la media de la población*, o *la proporción de individuos de esa población que tienen una determinada ó la diferencia de medias* .

Distribución de las medias muestrales

Vamos a considerar ahora todas las muestras posibles de tamaño *n* que se puedan extraer de una población, y la variable aleatoria $\bar{X}$ formada por sus correspondientes medias muestrales. Si llamamos **"μ"** y **"σ"** a la media y la desviación típica de la población (respectivamente), y siendo $\bar{X}$ la variable aleatoria formada por las medias muestrales, entonces se verifica:

(1) La media de $\bar{X}$ es μ, es decir $\mu(\bar{X}) = \mu$.

(2) La desviación típica de $\bar{X}$ es $\sigma/\sqrt{(n)}$, es decir $\sigma(\bar{X}) = \sigma/\sqrt{(n)}$. *(Este resultado sólo es válido para poblaciones infinitas o para poblaciones finitas en las que el muestreo se ha hecho con reemplazamiento).*

(3) Si $X \rightarrow N(\mu,\sigma)$, entonces $\bar{X} \rightarrow N(\mu , \frac{\sigma}{\sqrt{n}})$. *Distribución muestral de medias*

(4) ***(Teorema Central del Límite)***.- Si X *no sigue una ley normal*, pero $n \geq 30$, entonces se puede considerar que $\bar{X} \rightarrow N(\mu , \frac{\sigma}{\sqrt{n}})$

♦♦ Una población está formada por sólo cinco elementos, con valores 3, 5, 7, 9 y 11. Consideramos todas las muestras posible de tamaño 2 con reemplazamiento que puedan extraerse de esta población. Se pide calcular:

a) La media de la población.

b) La desviación típica de la población

c) La media de la distribución muestral de medias.

d) La desviación típica de la distribución muestral de medias, es decir, el error típico de las

medias.

sol

a) La media de la población es $\mu = (3 + 5 + 7 + 9 + 11)/5 = 35/5 = 7$

b) La desviación típica de la población es:

$$\sigma = \sqrt{\frac{(3-7)^2 + (5-7)^2 + (7-7)^2 + (9-7)^2 + (11-7)^2}{5}} = \sqrt{8} = 2{,}8284$$

Construyamos la distribución muestral de medias y, para ello, calculamos la media de todas las muestras posibles con reemplazamiento de tamaño 2 que son 25. Los resultados pueden verse en la tabla siguiente:

MUESTRAS																									
Elementos	3 3	3 5	3 7	3 9	3 11	5 3	5 5	5 7	5 9	5 11	7 3	7 5	7 7	7 9	7 11	9 3	9 5	9 7	9 9	9 11	11 3	11 5	11 7	11 9	11 11
Media de la muestra $\bar{x}_i$	3	4	5	6	7	4	5	6	7	8	5	6	7	8	9	6	7	8	9	10	7	8	9	10	11

La distribución muestral de medias puede verse en la tabla que sigue.

Media de la Muestra $\bar{x}_i$	Numero de muestras	Probabilidad $p(\bar{x}_i)$
$\bar{x}_1 = 3$	1	1/25
$\bar{x}_2 = 4$	2	2/25
$\bar{x}_3 = 5$	3	3/25
$\bar{x}_4 = 6$	4	4/25
$\bar{x}_5 = 7$	5	5/25
$\bar{x}_6 = 8$	4	4/25
$\bar{x}_7 = 9$	3	3/25
$\bar{x}_8 = 10$	2	2/25
$\bar{x}_9 = 11$	1	1/25

Podemos representarla poniendo en abscisas las medias muestrales y en ordenadas las probabilidades.

c) La media de la distribución muestral de medias (media de medias) es:

$$\mu = \sum_{i=1}^{11} \bar{x}_i \cdot p(\bar{x}_i) = 3\cdot(1/25) + 4\cdot(2/25) + \ldots\ldots + 10\cdot(2/25) + 11(1/25) = 175/25 = 7$$

d) La desviación típica de la distribución muestral de medias es:

$$\sigma = \sqrt{\sum_{i=1}^{11} \bar{x}_i \cdot p(\bar{x}_i) - \bar{x}^2} = \sqrt{\frac{1325}{25} - 7^2} = \sqrt{4} = 2$$

Cuando la población es infinita o las muestras se extraen con reemplazamiento, se verifica:

$$\mu_{\bar{x}} = \mu \qquad \text{y} \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

♦♦ Las estaturas de 1200 estudiantes de un centro d enseñanza superior se distribuyen normalmente con media 1'72 y desviación típica 0'9 m. Si se toman 100 muestras de 36 estudiantes cada una, se pide:

a) La media y la desviación típica esperada de la distribución muestral de medias.

b) ¿En cuantas muestras cabría esperar una media entre 1'68 y 1'73 m?

c) ¿En cuantas muestras es de esperar que la media sea menor que 1'69 m?

sol

a) La media y la desviación típica esperada de la distribución muestral de medias es:

$$\mu_{\bar{x}} = \mu = 1'72\text{m} \quad \text{y} \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0,9}{\sqrt{36}} = 0'15 \text{ m}$$

Por ser el tamaño muestral mayor que 30 aplicamos el teorema central del límite, que afirma que la distribución muestral de medias se aproxima a una distribución normal: $N(\mu, \frac{\sigma}{\sqrt{n}})$

b) Tipificamos los valores 1,68 y 1,73 según la distribución N(1'72, 0'15), obteniendo.

$z_{1,68}$ =(1'68 – 1'72)/0'15 = - 0'27 y $z_{1,73}$ =(1'73 – 1'72)/0'15 = 0'07

La probabilidad de muestras con medias entre 1'68 y 1'73 m es:

$p(1'68 \leq \bar{X} \leq 1'73) = p(-0'27 \leq Z \leq 0'07) = p(Z \leq 0'07) - [1 - p(Z \leq 027)] =$

= 0'5279 - (1 – 0'6064) = 0'1343

El número de muestras esperado es 100 .0'1343 = 13 muestras.

c) Tipificamos el valor 1'69 m que se distribuye según N(1'72, 0'15), obteniendo:

$$z_{1,69} = (1'69 - 1'72)/0'15 = - 0'2$$

La probabilidad de muestras con medias menores que 1'69 m es:

$$p(Z \leq - 0'2) = 1 - p(Z \leq 0'2) = 1 - 0'5793 = 0'4207$$

El número de muestras esperado es 100 .0'4207 = 42 muestras.

Distribución de las proporciones muestrales

Nota.- Vamos a estudiar ahora de todas las muestras posibles de tamaño *n,* la *proporción de sus individuos que tienen una determinada característica*. Llamaremos **p** al va-

lor de esa proporción en toda la población, y P a la variable aleatoria constituida por las proporciones muestrales. Entonces también se puede demostrar que:

(1) La media de P es **p**, es decir $\mu(P) = \mathbf{p}$.

(2) La desviación típica de P es $\sqrt{\frac{p \cdot q}{n}}$, es decir $\sigma(P) = \sqrt{\frac{p \cdot q}{n}}$, donde q = 1 - p

(3) *Si n* ≥ 30, entonces se puede considerar que P $\rightarrow$ N(p, $\sqrt{\frac{p \cdot q}{n}}$), que es *la distribución muestral de proporciones.*

♦♦ Una población está formada por los elementos 1, 2, 4 y 6.

a) Calcula la proporción **p** de cifras impares.

b) Para cada una de las muestras con reemplazamiento de tamaño dos, calcula la proporción **P** de cifras impares.

c) Calcula la media y la desviación típica de la distribución muestral de proporciones.

sol

a) La proporción de cifras impares es p = 1/4 = 0,25

b) La proporción de cifras impares de cada una de las muestras puede verse en la tabla.

Muestras	1 1	1 2	1 4	1 6	2 1	2 2	2 4	2 6	4 1	4 2	4 4	4 6	6 1	6 2	6 4	6 6
Proporción (P)	1	0,5	0,5	0,5	0,5	0	0	0	0,5	0	0	0	0,5	0	0	0

c) La media de las proporciones anteriores es:

$\mu(P)$ = (1 + 0,5 + 0,5 + 0,5 + 0,5 + 0,5+ 0,5 + 0 + ...+ 0)/16 = 0’25

La desviación típica de la distribución de proporciones es:

$$\sigma(P) = \sqrt{\frac{1^2 + 0{,}5^2 + 0{,}5^2 + 0{,}5^2 + 0{,}5^2 + 0{,}5^2 + 0{,}5^2 + 0^2 + ... + 0^2}{16} - (0{,}25)^2} = 0'3062$$

Cuando la población es finita o las muestras se extraen con reemplazamiento en una población finita con proporciones p y q, se verifican las relaciones siguientes:

$$\mu(P) = p = 0'25 \quad \text{y} \quad \sigma(P) = \sqrt{\frac{p \cdot q}{n}} = 0'3062$$

♦♦ Una máquina fabrica piezas de precisión. En su producción habitual fabrica un 3% de piezas defectuosas. Un cliente recibe una caja de 500 piezas procedentes de la fábrica.

a) ¿Cuál es la probabilidad de que encuentre más del 5% de piezas defectuosas en la caja?

b) ¿Cuál es la probabilidad de que encuentre menos de un 1% de piezas defectuosas?

sol

La distribución muestral de proporciones admite como media y desviación típica:

$$\mu(P) = p = 0{,}03 \quad \text{y} \quad \sigma(P) = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{(0{,}03)(0{,}97)}{500}} = 0'0076$$

La distribución muestral se distribuye según la normal N(0'03; 0'0076), dado que el tamaño de las muestras es superior a 30. Las probabilidades pedidas son:

a) $p(P>0'05)=1- p(P\leq 0'05)=1- p(Z\leq \frac{0{,}05-0{,}03}{0{,}0076})=1-p(Z\leq 2'63)=1-0'9957=0'0043$

b) $p(P < 0'01) = p(Z< \frac{0{,}01-0{,}03}{0{,}0076}) = p(Z < -2'63) = 1 - p(Z < 2'63) = 1-0'9957 = 0'0043$

Distribución muestral de diferencia de medias

Nota.- Cuando estudiarnos dos colectivos conjunta y comparativamente se consideran: μ_1 la media del primer colectivo, σ_1 su desviación típica y n_1 el número de elementos de una muestra; así como μ_2 , σ_2 y n_2 las del segundo colectivo.

Nota.- Las relaciones existentes entre los estadísticos de la distribución muestral y los parámetros de las poblaciones, así como la relación entre las distribuciones de las poblaciones y la distribución muestral de diferencia de medias se muestran a continuación.

Nota.- Si dos poblaciones siguen sendas distribuciones normales $N(\mu_1,\sigma_1)$ y $N(\mu_2,\sigma_2)$, o bien, si ambas poblaciones tienen distribuciones cualesquiera con medias μ_1 y μ_2, desviaciones típicas σ_1 y σ_2 , y las respectivas muestras son de tamaños n_1 y n_2 mayor que 30, entonces la distribución muestral de diferencias de medias sigue una distribución normal $N(\mu_1 - \mu_2 ; \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}})$, y la variable tipificada viene dada por la expresión

$$Z = \frac{\left(\overline{x}_1 - \overline{x}_2\right) - (\mu_1 - \mu_2)}{\sigma_{\overline{x}_1 - \overline{x}_2}}$$

Nota.- Si σ_1 y σ_2 no son conocidas, se aproximan estas por las desviaciones típicas de sendas muestras siempre que el tamaño de ambas sea superior a 100.

♦♦ Los tubos de imagen de televisión fabricados por la empresa A tienen una duración media de vida de 2500 horas, con una desviación típica de 500 horas, mientras que los fabricados por la empresa B tienen una duración media de vida de 2300 horas con una desviación típica de 800 horas. Se toman 300 tubos de imagen de la empresa A y 200 de la empresa B. Calcula la probabilidad de que la duración media de vida de la muestra de A no sea superior en más de 100 horas a la duración media de vida de la muestra de B.

sol

La distribución muestral de medias de las poblaciones A y B, $\overline{X}_A$ y $\overline{X}_B$ está caracterizada

por $\mu_{\overline{X}_A}$ = 2500; $\mu_{\overline{X}_B}$ = 2500; $\sigma_{\overline{X}_A}$ = $\frac{500}{\sqrt{300}}$; $\sigma_{\overline{X}_A}$ = $\frac{800}{\sqrt{200}}$.

La distribución muestral de diferencia de medias, $\overline{X}_A - \overline{X}_B$; admite como media y desviación típica:

$\mu_{\overline{X}_A-\overline{X}_B}$ = $\mu_{\overline{X}_A}$ - $\mu_{\overline{X}_B}$ = 2500 – 2300 = 200 y $\sigma_{\overline{x}_1-\overline{x}_2}$ = $\sqrt{\frac{(500)^2}{300}+\frac{(800)^2}{200}}$ = 63’5

La probabilidad de que $\overline{X}_A - \overline{X}_B \leq$ 100 es:

P($\overline{X}_A - \overline{X}_B \leq$ 100) = p (Z $\leq \frac{100-200}{63{,}5}$) = P(Z ≤ -1,’57) = 0,0582

Inferencia

Introducción

Nota.- La inferencia estadística trata de obtener conclusiones sobre la población a partir de la información proporcionada por una muestra aleatoria; es decir, obtener de las propiedades de las muestras una aproximación fiable a las del colectivo o población en estudio.

Nota.- Las inferencias sobre el valor de un parámetro poblacional, como es la media **μ,** la proporción **p** ó la diferencia de medias, se pueden hacer mediante **estimaciones** (puntuales o por intervalos de confianza) y mediante **contrastes de hipótesis** (lo veremos en otra lección).

Def.- Un **parámetro** es un valor numérico que describe una característica de la población (**μ**, **p**, σ^2 , etc.)

Def.- Un **estadístico** es toda función de los datos muestrales, que asigna a cada muestra de tamaño **n** elegida de la población (por muestreo aleatorio simple), un valor numérico. Tenemos una variable aleatoria que tendrá una distribución de probabilidad **llamada Distribución en el muestreo del estadístico.**

Def.- Un **estimador para un parámetro poblacional** desconocido es un estadístico que nos da un valor que pertenece al conjunto de valores que puede tomar el parámetro que se estima. Los que usaremos son:

- Para la media poblacional **μ** utilizaremos el **estimador** MEDIA MUESTRAL $\overline{X}$, que sabemos sigue una N(μ , $\frac{\sigma}{\sqrt{n}}$), es decir:

$$\overline{X} \approx N(\mu , \frac{\sigma}{\sqrt{n}})$$

(Se considerarán las muestras de tamaño n ≥ 30 para poder aplicar el Teorema Central del Límite y asegurar la distribución anterior).
- Para la proporción muestral **p** utilizaremos el **estimador** PROPORCIÓN MUESTRAL P, que sabemos sigue una N(p, $\sqrt{\frac{p \cdot q}{n}}$), es decir:

$$P \approx N(p, \sqrt{\frac{p \cdot q}{n}}), \text{ donde } \mathbf{q} = 1 - \mathbf{p}$$

(Se considerarán las muestras de tamaño n ≥ 30 para poder aplicar el Teorema Central del Límite y asegurar la distribución anterior).

- Para la diferencia de medias $\mu_1 - \mu_2$ utilizaremos el **estimador** DIFERENCIA DE MEDIAS $\bar{X}_2 - \bar{X}_1$, que sabemos sigue una $N(\mu_1 - \mu_2; \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$, es decir:

$$\bar{X}_2 - \bar{X}_1 \approx N(\mu_1 - \mu_2; \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

(Se considerarán las muestras de tamaño n ≥ 30 para poder aplicar el Teorema Central

Estimación Puntual

Nota.- Consiste en tomar como valor del parámetro poblacional desconocido (μ, **p**...), el de un estadístico ($\bar{x}$, $\hat{p}$, ...), obtenido en una muestra aleatoria elegida de la población objeto de estudio; es decir, al ofrecido por el estimador sobre una muestra.

Se utilizarán los estimadores definidos en el apartado anterior para hacer estimaciones de la media y la proporción poblacional.

Estimación por intervalos de confianza

Nota.- Consiste en encontrar un intervalo (a, b) de manera que tengamos una cierta confianza (nivel de confianza $1 - \alpha$) de que el parámetro poblacional desconocido μ, **p**...), se encuentre en dicho intervalo.
Se considera que la población de partida sigue una distribución Normal con desviación típica conocida (σ) para la estimación de μ, o una distribución Binomial para la estimación de **p.**

Pasos para construir el intervalo de confianza

(a) Se elige un **estimador** del parámetro que se desea estimar ($\bar{X}$ para μ, $\hat{P}$ para **p** y $\bar{X}_1 - \bar{X}_2$ para $\mu_1 - \mu_2$).
(b) Se elige un **nivel de confianza 1 – α** con el que se desea construir el intervalo, eso quiere decir que, antes de elegir la muestra, se tendrá un probabilidad $1 - \alpha$ de que el intervalo construido a partir de esa muestra contenga al parámetro de la población.
(c) **Se toma una muestra aleatoria de la población de tamaño n** y en ella se obtiene el valor del estadístico correspondiente.
(d) Se construye el intervalo centrado en el estadístico ($\bar{x}, \hat{p}, \bar{x}_2 - \bar{x}_1$), teniendo en cuenta que al ser intervalos simétricos, se tiene que cumplir $p(|Z| < z_{1-\alpha/2}) = 1 - \alpha$. Desarrollando esta expresión obtenemos, según la distribución muestral correspondiente, obtendremos las probabilidades:

$$p\left(\bar{x} - z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

$$p\left(\hat{p} - z_{1-\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) = 1 - \alpha$$

$$p\left((\bar{x}_1-\bar{x}_2) - z_{1-\alpha/2}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1-\bar{x}_2) + z_{1-\alpha/2}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right) = 1 - \alpha$$

con lo cual los respectivos intervalos de confianza serán:

$$I(\mu) = \left(\bar{x} - z_{1-\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) \text{ para estimar } \mu$$

$$I(\mathbf{p}) = \left(\hat{p} - z_{1-\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\cdot\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \text{ para estimar } \mathbf{p}$$

$$I(\mu_1 - \mu_2) = \left((\bar{x}_1-\bar{x}_2) - z_{1-\alpha/2}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}, (\bar{x}_1-\bar{x}_2) + z_{1-\alpha/2}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right) \text{ para estimar } \mu_1 - \mu_2$$

Donde $z_{1-\alpha/2}$ **es el punto crítico** de la variable aleatoria Normal tipificada $Z \approx N(0,1)$ tal que $p(-z_{1-\alpha/2} \le Z \le z_{1-\alpha/2}) = 1 - \alpha$, siendo $1 - \alpha$ el nivel de confianza elegido.

De la igualdad $p(-z_{1-\alpha/2} \le Z \le z_{1-\alpha/2}) = 1 - \alpha$, se deduce que $p(Z \le z_{1-\alpha/2}) = 1 - \alpha/2$, que se mira en la tabla de la distribución Normal, y nos dará el correspondiente valor crítico $z_{1-\alpha/2}$.

Def.- Se llama **amplitud del intervalo** a la diferencia = Extremo superior - Extremo inferior del intervalo de confianza.

♦♦ Se ha extraído una muestra de 145 alumnos de una escuela de artes, a los que se les ha propuesto un test de habilidad. La media y la desviación típica obtenida de la muestra son 82 y 14, respectivamente. A partir de estos datos, calcula el intervalo en el cual se hallará la media de población al nivel de confianza del 95%. Calcula el intervalo de confianza para los mismos datos correspondientes al nivel de confianza del 99%.

Sol

Los valores que proporciona la muestra de tamaño n = 145 son: $\bar{x}$ = 82 y σ = 14. La distribución muestral de medias sigue una distribución normal $N(\mu, \sigma_{\bar{x}})$. Como el tamaño muestral es superior a 100, podemos aproximar la desviación típica de la muestra por la de la población:

El valor crítico $z_{1-\alpha/2}$, correspondiente al nivel de confianza $1 - \alpha$ = 95% es $z_{1-\alpha/2}$ = 1'96; porque $p(Z \le z_{1-\alpha/2}) = 1 - \alpha/2 = 1 - 0'05/2 = 0'975$, y mirando en la N(0;1) obtenemos $z_{1-\alpha/2}$= 1'96

$$\left(\bar{x} - z_{1-\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right)$$

Sustituyendo y operando, se obtiene: (82 – 1'96.1'162; 82 + 1'96.1'162) = (79'72; 84'28).

Por tanto, el intervalo (79'72; 84'28) contendrá la media μ de la población con una probabilidad de 95%.

En el caso del nivel de confianza del 99% se tiene que el valor crítico $z_{1-\alpha/2}$ correspondiente a este nivel de confianza es $z_{1-\alpha/2} = 2{,}58$; pues de $p(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2 = 1 + 0'01/2 = 0'995$, y mirando en la N(0;1) obtenemos $z_{1-\alpha/2} = 2'58$

Sustituyendo y operando, se tiene: $(82 - 2'58.1'162, 82 + 2'58.1'162) = (79'00, 85'00)$.

Luego, el intervalo (79'00, 85'00) contendrá a la media de la población con una probabilidad del 99%.

Se observa que, *al aumentar el nivel de confianza, se amplía el intervalo* y tenemos más seguridad de encontrar la media de la población en el último intervalo calculado.

♦♦ Para estimar la proporción de estudiantes de una universidad que está a favor de la reinserción social del delincuente, se entrevistó aleatoriamente a 500 estudiantes. El 58% estaba a favor. Calcula el intervalo de confianza, al nivel de confianza del 95%, en el cual se hallará la población universitaria que se encuentra a favor.

Sol

Como el tamaño muestral es superior a 100, podemos aproximar P y Q de la población por las proporciones p y q de la muestra.

$p = 0{,}58$; $q = 0{,}42$; $\sigma(p) = \sqrt{[(PQ)/n]} = \sqrt{[(pq)/n]} = \sqrt{[(0'58 \cdot 0'42)/500]} = 0'02$.

Hemos visto en un problema anterior que a un nivel de confianza $1 - \alpha = 95\% = 0'95$, le corresponde el valor $z_{1-\alpha/2} = 1{,}96$.

El intervalo de confianza para una proporción p es $\left(P - z_{1-\alpha/2} \cdot \sqrt{\frac{PQ}{n}}, P + z_{1-\alpha/2} \cdot \sqrt{\frac{PQ}{n}} \right)$, sustituyendo y operando con los datos, se obtiene $(0'58 - 1'96.0'02, 0'58 + 1'96.0'02)$ es decir el intervalo es (0'5408; 0'6192) al nivel de confianza del 95%

El verdadero porcentaje poblacional P se encontrará en el intervalo (0'5408; 0'6192) con una probabilidad del 95%.

- El **error de la estimación** es la diferencia, en valor absoluto, entre el parámetro poblacional y el estadístico muestral, por lo tanto el error máximo de estimación será el radio del intervalo (lo que sumamos o restamos al punto medio del intervalo):

-**Error máximo** = $E = z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$, para el intervalo de la media (radio del intervalo) y

-**Error máximo** = $E = z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{p}.\hat{q}}{n}}$, para el intervalo de la proporción (radio

del intervalo) .

-**Error máximo** = E = $z_{1-\alpha/2}\cdot\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$, para el intervalo diferencia de medias (radio del intervalo) .

Nota.- A veces si me dan el intervalo (a,b) tenemos en cuenta que el error es E = $\frac{b-a}{2}$, y que el punto medio del intervalos $\frac{a+b}{2}$, será $\overline{x}$, $\hat{p}$ ó $(\overline{x}_1-\overline{x}_2)$, dependiendo del tipo de intervalo.

Tamaño de la muestra n

-Si aumentamos el tamaño de la muestra elegida, manteniendo la misma confianza, o sea, el mismo valor crítico $z_{1-\alpha/2}$, menor error cometemos al inferir el valor del parámetro.

-Esto quiere decir que podemos calcular el tamaño de muestra necesario para tener un error máximo concreto con un nivel de confianza previamente fijado.

-Si se mantiene fijo el tamaño de la muestra y se desea aumentar el nivel de confianza $1-\alpha$ (con lo que aumentaría el valor crítico $z_{1-\alpha/2}$), aumentaría también el error de la estimación.

Nota.- De la fórmula del error (lo que sumamos o restamos al punto medio del intervalo de confianza) se suele sacar el tamaño de la muestra, despejando la incógnita "n".

De E = $z_{1-\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}$, obtenemos n = $\left(\frac{z_{1-\alpha/2}\ .\ \sigma}{E}\right)^2$ en $N(\mu;\ \frac{\sigma}{\sqrt{n}})$

De E = $z_{1-\alpha/2}\cdot\sqrt{\frac{\hat{p}.\hat{q}}{n}}$, obtenemos n = $\frac{(z_{1-\alpha/2})^2.\hat{p}.\hat{q}}{E^2}$ en $N(\hat{p}\ ;\ \sqrt{\frac{\hat{p}.\hat{q}}{n}})$

♦♦Se desea hacer una estimación sobre la edad media de una determinada población. Calcula el tamaño de la muestra necesario para poder realizar dicha estimación con un error medio de medio año a un nivel de confianza del 99,73%. Se conoce de estudios previos que la edad media de dicha población tiene una desviación típica de σ = 3.

Sol

A un nivel de confianza $1-\alpha$ = 99'73% le corresponde un valor crítico $z_{1-\alpha/2}$ = 3, porque de $p(Z \le z_{1-\alpha/2}) = 1-\alpha/2 = 1-0'0027/2 = 0'99865$, y mirando en la N(0;1) obtenemos $z_{1-\alpha/2}$= 3

Además, σ = 3 y el error E = 0,5.

Con estos datos si lo ponemos en la fórmula n = $\left(\frac{z_{1-\alpha/2}\ .\ \sigma}{E}\right)^2$, se obtiene n = $(3^2.3^2)/(0'5)^2$= =324

De 324 personas, al menos, debe estar compuesta la muestra.

♦♦ Deseamos conocer el número de personas mayores de edad que sería necesario incluir en una muestra nacional para estimar la clase de actividad en España con un error absoluto de E = 0,04 y un nivel de confianza del 99'73%. Se dispone de una valor P = 0,45 del último censo.

Sol

A un nivel de confianza 1 - α = 99,73% le corresponde un valor crítico $z_{1-\alpha/2}$ = 3, porque de $p(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2 = 1 - 0'0027/2 = 0'99865$, y mirando en la N(0;1) obtenemos $z_{1-\alpha/2} = 3$. Además P = 0,45, entonces Q = 0,55.

Con estos datos llevados a la fórmula $n = \frac{(z_{1-\alpha/2})^2 . \hat{p} . \hat{q}}{E^2}$, se obtiene $n=(3^2.0'45.0'55)/(0'44)^2$ = 1392

Se necesitan para la muestra, al menos 1392 personas.

♦♦ Al medir el tiempo de reacción, un psicólogo estima que la desviación típica es de 0,05 segundos. ¿De qué tamaño ha de tomarse una muestra de medidas para tener una confianza de 99% de que el error de estimación no supera 0,01 segundos?

Sol

La variable tiempo de reacción tiene media μ y desviación típica σ = 0,05.

La distribución de medias muestrales $\bar{x}$ sigue una ley de media μ y desviación típica $\sigma/\sqrt{(n)} = 0,05/\sqrt{(n)}$

Por tanto, la variable $Z = \frac{\bar{x} - \mu}{0'05/\sqrt{n}}$ se distribuye según la normal N(0, 1).

El valor crítico, $z_{1-\alpha/2}$, para el intervalo en al nivel de confianza 1 - α = 0,99 es $z_{1-\alpha/2}$ = 2,58; porque de $p(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2 = 1 - 0'01/2 = 0'995$, y mirando en la N(0;1) obtenemos $z_{1-\alpha/2}$ = 2,58.

Sabemos que en las medias muestrales el error es $E = z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$, de donde despejando se obtiene $n = \left(\frac{z_{1-\alpha/2} \cdot \sigma}{E} \right)^2 = \left(\frac{2'58.0'05}{0'01} \right)^2$ = 166'41, de donde el tamaño de la muestra de medidas tiene que ser 167 o mayor.

Algunos Ejercicios de Selectividad

♦♦ En cierto barrio se quiere hacer un estudio para conocer mejor el tipo de actividades de ocio que gustan más a sus habitantes. Para ello, van a ser encuestados 100 individuos elegidos al azar.

a) Explica qué procedimiento de selección sería más adecuado utilizar: muestreo con o sin reposición. ¿Por qué?

b) Como los gustos cambian con la edad y se sabe que en el barrio viven 2500 niños, 7000 adultos y 500

ancianos, posteriormente se decide elegir la muestra anterior utilizando muestreo estratificado.

b_1) Define los estratos.

b_2) Determina el tamaño muestra¡ correspondiente a cada estrato.

Sol ((a) 25 niños, 70 adultos, 5 ancianos)

♦♦ Se sabe que el cociente intelectual de los alumnos de una universidad se distribuye según la ley normal de media 100 y varianza 729.

a) Halla la probabilidad de que la muestra de 81 alumnos tenga un cociente intelectual medio inferior a 109.

b) Halla la probabilidad de que la muestra de 36 alumnos tenga un cociente intelectual medio superior a 109.

Sol ((a) $P(x < 109) = 0{,}9987$; (b) $P(x > 109) = 0{,}0228$)

♦♦ Los 6000 huevos de una gran partida tienen masas **que están distribuidas normalmente.** Se escogen al azar 10 huevos y se halla que sus masas son: 40, 36, 44, 42, 48, 49, 38, 50 y 38 gramos, respectivamente.

a) Halla la media y la desviación de la muestra.

b) Suponiendo que la masa media de los huevos de la partida es la misma que la calculada en a), pero que la desviación típica de la masa es de 5,5 gramos, demuestra que el número de huevos de la partida con masa superior a 50 gramos es aproximadamente 440.

c) Sabiendo que 5 000 de los 6 000 huevos tienen masas superiores a x gramos, estima el valor de x.

Sol ((a) $\overline{X} = 42$ g; $\sigma = 5{,}23\ 9$; (b) 441 huevos; (c) x = 36,72 gramos)

♦♦ Una muestra aleatoria de 100 alumnos que se presenta a las pruebas de selectividad revela que la media de edad es de 18,1 años. Halla un intervalo de confianza de 90% para la edad media de todos los estudiantes que se presentan a las pruebas, sabiendo que la desviación típica de la población es de 0,4.

Sol (18,034; 18,166)

♦♦ Se quiere conocer la permanencia media de pacientes en un hospital. Se tienen datos referidos a la estancia, expresada en días, de 800 pacientes, de donde se han sacado los resultados siguientes: $\overline{x} = 8{,}1$ días, s = 9 días

Se pide obtener un intervalo de confianza del 95% para la estancia media.

Sol (7,57; 8,73)

♦♦ En una muestra aleatoria de 1 000 personas, están a favor de que el ministerio de Economía mantenga la presión fiscal el 65%. Halla el intervalo de confianza del 99%. En una encuesta realizada un año antes había resultado un 68% favorable al mantenimiento de la presión. ¿Cae este valor dentro del margen de confianza de la nueva encuesta?

Sol (p = 68% = 0,68; este valor cae dentro del intervalo)

♦♦ La duración de bombillas sigue una distribución normal de media desconocida y desviación típica de 50 horas. Para estimar la duración media, se experimenta con una muestra de tamaño n Calcula el valor de n para que, con un nivel de confianza del 95%, se haya conseguido un error en la estimación inferior a 5 horas.

Sol (385 bombillas)