

UNIDAD 8: INFERENCIA ESTADÍSTICA I

CONTENIDO

1.	ESTADÍSTICA INFERENCIAL. MUESTREO	2
2.	MUESTREOS ALEATORIOS.....	2
3.	DISTRIBUCIONES MUESTRALES.....	4
4.	ESTIMACIÓN DE PARÁMETROS. ESTIMACIÓN PUNTUAL	9
5.	ESTIMACIÓN POR INTERVALOS DE CONFIANZA	9
6.	TAMAÑO DE LAS MUESTRAS. ERROR MÁXIMO ADMISIBLE.....	15

1. ESTADÍSTICA INFERENCIAL. MUESTREO

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**

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.*

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*

La estadística inferencial se divide en:

- **Estadística inductiva:** Se encarga de estimar los parámetros de una población, y lo puede hacer de dos formas distintas:
 - Mediante un único valor o estimación puntual
 - Mediante un intervalo o estimación por intervalos
- **Estadística deductiva:** Su fin es comprobar si la información que proporciona la muestra concuerda o no con la hipótesis estadística formulada:
 - Mediante los contrastes de hipótesis

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

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

Los estadísticos que más vamos a usar son:

- $\bar{\mu}$ ó \bar{x} , la media muestral
- $\bar{\sigma}$ ó \bar{s} , la desviación típica muestral

2. MUESTREOS ALEATORIOS

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

Para poder obtener información de la población a través de la muestra es necesario que ambas tengan características parecidas, en tal caso se dice que la muestra es "representativa" de la población.

Ejemplo: Se hace un estudio sobre los pesos de los recién nacidos, para ello se considera la variable aleatoria:

$X =$ "peso de un recién nacido"

La población sobre la que se hace el estudio son todos los niños que nacen.

No se puede analizar toda la población (los niños no paran de nacer), por lo que es necesario tomar una muestra.

Es importante saber cuál debe ser el tamaño de la muestra para que sea representativa, parece evidente que 2 niños no son suficientes, pero ¿cuántos pesar?

Por otro lado, hay de determinar cómo elegir a los niños, ¿importa el sexo?, ¿pueden ser todos nacidos el mismo día?, ¿o en el mismo hospital?

Tipos de muestreo

- **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 no aleatorio o no probabilístico**: es el que se realiza de forma que todos los elementos de la población no tienen la misma probabilidad de ser incluidos en la muestra. Este tipo de muestreo suele ser de escasa representatividad y poco válidas las inferencias que puedan hacerse
- **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. Es el más utilizado ya que presenta un mayor grado de fiabilidad y es el que vamos a estudiar más en detalle

Muestreo aleatorio

Los tipos fundamentales de muestreos aleatorios son tres:

Muestreo aleatorio simple: Consiste en listar todos los elementos de la población y seleccionar aleatoriamente n elementos de la muestra. El muestreo aleatorio simple en poblaciones finitas se entenderá “con reemplazamiento” según las directrices dadas por la comisión de Selectividad.

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 $n^{\circ} k = \frac{N}{n}$ (llamado coeficiente de elevación), y

después elegimos, al azar, uno de los k 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 k en k hasta completar todos los elementos que componen 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. Supongamos que tenemos k estratos o colectivos que nos dividen a la población.

Llamaremos $N_1, N_2, N_3, \dots, N_k$ al n° de elementos de la población que tiene cada estrato, es decir su tamaño.

Se cumple que $N_1 + N_2 + N_3 + \dots + N_k = N$. Sean $n_1, n_2, n_3, \dots, n_k$ al número de individuos de los respectivos estratos que hay en la muestra (con $n_1 + n_2 + n_3 + \dots + n_k = n$).

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 $\frac{n}{5}$ individuos para la muestra, independientemente del peso que cada uno de ellos tuviera en la

población. Es decir, $n_1 = n_2 = n_3 = \dots = n_k = \frac{n}{k}$. No suele ser representativa y se usa poco.

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. Este tipo es el más usado.

En este caso han de ser proporcionales y por tanto se ha de cumplir que $\frac{n_1}{N_1} = \frac{n_2}{N_2} = \frac{n_3}{N_3} = \dots = \frac{n_k}{N_k} = \frac{n}{N}$

De estas proporciones podemos saber el nº de elementos de cada estrato que ha de tener la muestra.

Ejemplo: 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?

a) En el muestreo estratificado con afijación igual dividimos el total de la muestra 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:

$$\frac{x}{1500} = \frac{y}{7500} = \frac{z}{1000} = \frac{200}{10000} = \frac{1}{50} \Rightarrow \begin{cases} x = \frac{1500}{50} = 30 \\ y = \frac{7500}{50} = 150 \\ z = \frac{1000}{50} = 20 \end{cases}$$

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

3. DISTRIBUCIONES MUESTRALES

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 característica o 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 \overline{X}_n es la que asigna a cada muestra su media. Esta distribución se llama distribución de las medias muestrales.

Si llamamos μ y σ a la media y la desviación típica de la variable aleatoria de la población (respectivamente), y siendo \overline{X}_n la variable aleatoria formada por las medias muestrales, entonces se verifica el siguiente teorema:

Teorema Central del Límite:

Para un n suficientemente grande (se considera grande para $n \geq 30$), se tiene que la distribución muestral de medias \overline{X}_n se aproxima a una distribución normal, es decir, $\overline{X}_n \rightarrow N(\mu, \frac{\sigma}{\sqrt{n}})$.

Resumiendo,

- La media de las medias muestrales, \overline{X}_n , es la de la población, $\overline{\mu} = \mu$
- La desviación típica de las medias muestrales, \overline{X}_n , es $\overline{\sigma} = \frac{\sigma}{\sqrt{n}}$

Corolario:

En el caso particular que sepamos que la población sigue una distribución normal, la distribución de medias muestrales también sigue una distribución normal, y en este caso independientemente del tamaño de la muestra, es decir,

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

En la práctica ocurre que la desviación típica de la población es desconocida. En estos casos se aproxima por la desviación típica de la muestra siempre que el tamaño de esta sea suficientemente grande, $n \geq 100$

Ejemplo: Una población está formada por sólo cinco elementos, con valores 3, 5, 7, 9 y 11.

Consideramos todas las muestras posibles 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.

a) La media de la población es $\mu = \frac{3 + 5 + 7 + 9 + 11}{5} = \frac{35}{5} = 7$

b) La desviación típica de la población es: $\sigma = \sqrt{\frac{1 \cdot 3^2 + 1 \cdot 5^2 + 1 \cdot 7^2 + 1 \cdot 9^2 + 1 \cdot 11^2}{5} - 7^2} = \sqrt{8} = 2,8284$

c) 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	3	3	5	5	5	5	5	7	7	7	7	7	9	9	9	9	9	11	11	11	11	11
	3	5	7	9	11	3	5	7	9	11	3	5	7	9	11	3	5	7	9	11	3	5	7	9	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 para calcular media, varianza y desviación típica:

\bar{x}_i	f_i	$f_i \cdot \bar{x}_i$	$f_i \cdot \bar{x}_i^2$
3	1	3	9
4	2	8	32
5	3	15	75
6	4	24	144
7	5	35	245
8	4	32	256
9	3	27	243
10	2	20	200
11	1	11	121
TOTALES:	25	175	1325

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

$$\bar{\mu} = \frac{\sum f_i \cdot \bar{x}_i}{N} \Rightarrow \bar{\mu} = \frac{175}{25} = 7$$

d) Primero calculamos la varianza:

$$\sigma^2 = \frac{\sum f_i \cdot \bar{x}_i^2}{N} - \bar{\mu}^2 \Rightarrow \sigma^2 = \frac{1325}{25} - 7^2 = 53 - 49 = 4$$

La desviación típica de la distribución muestral de medias es:

$\bar{\sigma} = \sqrt{4} = 2$, que como vemos coincide con la manera más fácil de calcularla, aplicando el teorema Central de Límite: $\bar{\sigma} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{8}}{\sqrt{2}} = \sqrt{\frac{8}{2}} = \sqrt{4} = 2$

Ejemplo: Las estaturas de 1200 estudiantes de un centro de enseñanza superior se distribuyen normalmente con media 1'72 y desviación típica 0'09 m. Si se toman 100 muestras de 36 estudiantes cada una, se pide:

- La media y la desviación típica esperada de la distribución muestral de medias.
- ¿En cuántas muestras cabría esperar una media entre 1'68 y 1'73 m?
- ¿En cuántas muestras es de esperar que la media sea menor que 1'69 m?

a) La media y la desviación típica esperada de la distribución muestral de medias es, por el Teorema Central del Límite:

$$\bar{\mu} = \mu = 1,72 \quad \text{y} \quad \bar{\sigma} = \frac{\sigma}{\sqrt{n}} = \frac{0,09}{\sqrt{36}} = 0,015$$

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: $\bar{X} = N(\mu, \frac{\sigma}{\sqrt{n}}) = N(1,72, 0,015)$

b) Nos están pidiendo $P(1,68 < \bar{X} \leq 1,73)$, y para ello hemos de tipificar la variable haciendo $Z = \frac{\bar{X} - 1,72}{0,015}$, con lo

$$\text{cual } P(1,68 < \bar{X} \leq 1,73) = P\left(\frac{1,68 - 1,72}{0,015} < \frac{\bar{X} - 1,72}{0,015} \leq \frac{1,73 - 1,72}{0,015}\right) = P(-2,67 < Z \leq 0,67) =$$

$$P(Z \leq 0,67) - P(Z \leq -2,67) = P(Z \leq 0,67) + P(Z \leq 2,67) - 1 = 0,7486 + 0,9962 - 1 = 0,7448$$

El nº de muestras esperado es aproximadamente $0,7448 \cdot 100 = 74,48 \approx 74$

c) Nos están pidiendo $P(\bar{X} \leq 1,69)$, y tipificando igual que en el apartado anterior tenemos que:

$$P(\bar{X} \leq 1,69) = P\left(Z \leq \frac{1,69 - 1,72}{0,015}\right) = P(Z \leq -2) = 1 - P(Z \leq 2) = 1 - 0,9772 = 0,0228$$

$$\text{nº muestras} = 0,0228 \cdot 100 = 2,28$$

Es decir, aproximadamente 2 muestras tienen la media menor que 1,69

Distribución de las proporciones muestrales

Cuando estudiamos en una población una determinada característica que sólo puede tomar dos valores: éxito y fracaso, la población objeto del estudio sigue una distribución binomial

Cada una de las muestras que extraigamos de esa población tendrá un porcentaje de individuos con esa misma característica.

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 valor de esa proporción en toda la población y llamaremos \hat{p} a la proporción de individuos con esa característica en cada una de las muestras.

La distribución asociada a la variable aleatoria que asocia a cada muestra su proporción es la **distribución muestral de proporciones**.

Teorema:

a) La media de \hat{p} es p , es decir, $\mu_{\hat{p}} = p$

b) La desviación típica de \hat{p} es $\sqrt{\frac{p \cdot q}{n}}$, es decir, $\sigma_{\hat{p}} = \sqrt{\frac{p \cdot q}{n}}$ donde $q = 1 - p$

c) Si $n \geq 30, n \cdot p \geq 5$ y $n \cdot q \geq 5$, la distribución muestral de proporciones \hat{p} se aproxima a una distribución

$$\text{normal } \hat{p} \approx N\left(p, \sqrt{\frac{p \cdot q}{n}}\right)$$

En la práctica ocurre que las proporciones p y q de la población son desconocidas. En estos casos, se aproximan por las respectivas de una muestra, $\hat{p} = \mu_{\hat{p}}$, por ser \hat{p} un estimador insesgado.

Ejemplo: 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 de cifras impares.

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

a) La proporción de cifras impares es: $p = \frac{1}{4} = 0,25$

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

Muestra	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 \hat{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:

$$\mu_{\hat{p}} = \frac{1+0,5+0,5+0,5+0,5+0,5+0,5}{16} = 0,25 = p$$

Y la desviación típica es

$$\sigma_{\hat{p}} = \sqrt{\frac{1^2 + 0,5^2 + 0,5^2 + 0,5^2 + 0,5^2 + 0,5^2 + 0,5^2}{16} - (0,25)^2} = 0,3062, \text{ que verifica que es igual a}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{(0,25) \cdot (0,75)}{2}} = 0,3062, \text{ pues la población es finita y las muestras se extraen con reemplazamiento}$$

Ejemplo: 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?

La distribución muestral de proporciones admite como media y desviación típica con los datos que tenemos:

$$\mu_{\hat{p}} = p = 0,03 \text{ y } \sigma_{\hat{p}} = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{0,03 \cdot 0,97}{500}} = 0,0076$$

La distribución muestral de proporciones se distribuye según una $N(0,03; 0,0076)$ dado que el tamaño de las muestras es mayor que 30. Con esto respondemos a los apartados:

- a) $P(\hat{p} > 0,05) = (\text{tipificamos}) P\left(Z > \frac{0,05 - 0,03}{0,0076}\right) = 1 - P(Z \leq 2,63) = 1 - 0,9957 = 0,0043$
 b) $P(\hat{p} \leq 0,01) = (\text{tipificamos}) P\left(Z \leq \frac{0,01 - 0,03}{0,0076}\right) = P(Z \leq -2,63) = P(Z > 2,63) =$
 $= 1 - P(Z \leq 2,63) = 1 - 0,9957 = 0,0043$

4. ESTIMACIÓN DE PARÁMETROS. ESTIMACIÓN PUNTUAL

Dentro de la estadística inferencial está la estadística inductiva, que se basa en la estimación de parámetros poblacionales a partir de los correspondientes estadísticos muestrales.

Esta estimación puede hacerse de dos formas: **estimación puntual** y **estimación por intervalos**

Por ejemplo, cuando decimos que la altura media de los adolescentes es de 1,75 m estamos haciendo una estimación puntual; en cambio, si decimos que la altura media de los adolescentes está entre 1,73 y 1,77 m estamos haciendo una estimación por intervalos.

Por tanto, la **estimación puntual** consiste en estimar mediante un único valor el parámetro poblacional desconocido.

En la estimación puntual, el estadístico que utilizamos para la estimación se llama estimador puntual.

Los estimadores puntuales pueden ser:

- **Estimador puntual insesgado:** Si la media de la distribución muestral de un estadístico es igual a su correspondiente parámetro poblacional. Son estimadores puntuales insesgado la media muestral o la proporción muestral.
- **Estimador puntual sesgado:** Si la media de la distribución muestral de un estadístico no coincide con su correspondiente parámetro poblacional.

5. ESTIMACIÓN POR INTERVALOS DE CONFIANZA

Consideraremos una variable aleatoria X con una media μ desconocida y una desviación típica σ conocida (parámetros poblacionales).

Lo que deseamos es estimar el valor de μ a partir de la información de una muestra de tamaño n , con parámetros \bar{x} y $\bar{\sigma}$ (parámetros de la muestra).

¿Qué es un intervalo de confianza?

Para estimar la media poblacional μ , vamos a dar un intervalo alrededor de la media muestral \bar{x} , que llamaremos *intervalo de confianza* $[\bar{x} - E, \bar{x} + E]$.

Observaciones:

- Cada posible muestra generará un intervalo de confianza.
- Los diferentes intervalos de confianza puede que contengan a la media poblacional o no.
- El valor E es el *error máximo* cometido si el intervalo de confianza contiene a la media poblacional (μ).

Llamaremos *nivel de confianza* $N_c = (1 - \alpha)$, si el $(1 - \alpha) \cdot 100\%$ de los intervalos de confianza contienen a la media poblacional o al parámetro poblacional que se está estimando. Al valor α se le denomina *nivel de riesgo o nivel de significación* y representa el porcentaje de intervalos de confianza que no contendrán a la media.

Por ejemplo, si el nivel de confianza es del 95% ($1 - \alpha = 0,95$), quiere decir que el 95% de las muestras generarán intervalos de confianza que contendrán a la media. Y en este caso, el nivel de riesgo o de significación es del 0,05% ($\alpha = 0,05$) y es el porcentaje de intervalos de confianza que no contendrán a la media

Así pues:

Dada una muestra de tamaño n , con una media \bar{x} , llamaremos *intervalo de confianza* con un nivel de confianza $1 - \alpha$ a un intervalo alrededor de \bar{x} , $[\bar{x} - E, \bar{x} + E]$, tal que la media poblacional, μ , estará en ese intervalo en el $(1 - \alpha) \cdot 100\%$ de las posibles muestras.

$$P(\bar{X} - E \leq \mu \leq \bar{X} + E) = 1 - \alpha$$

Nota: es un error decir que, dado un intervalo de confianza, éste contendrá a la media con una cierta probabilidad, porque una vez dado el intervalo, éste contiene la media o no la contiene.

Ejemplo:

Si tenemos una población de 60 individuos, con una media desconocida.

Si tomamos muestras de tamaño 5, tendríamos $60^5 = 777.600.000$ posibles muestras distintas.

Cada una de las muestras tendrá su media, \bar{x}_i , si fijamos un nivel de confianza del 95% y construimos los 777.600.000 intervalos de confianza (cada uno alrededor de su media muestral), lo que sabremos es que el 95% de esos intervalos, es decir, 738.720.000 contendrán a la media de la población.

El paso siguiente es construir los intervalos de confianza, para ello nos va a hacer falta un valor numérico que se conoce como VALOR CRÍTICO que se representa por $z_{\alpha/2}$. Veamos cómo se calcula:

CÁLCULO DEL VALOR CRÍTICO $z_{\alpha/2}$

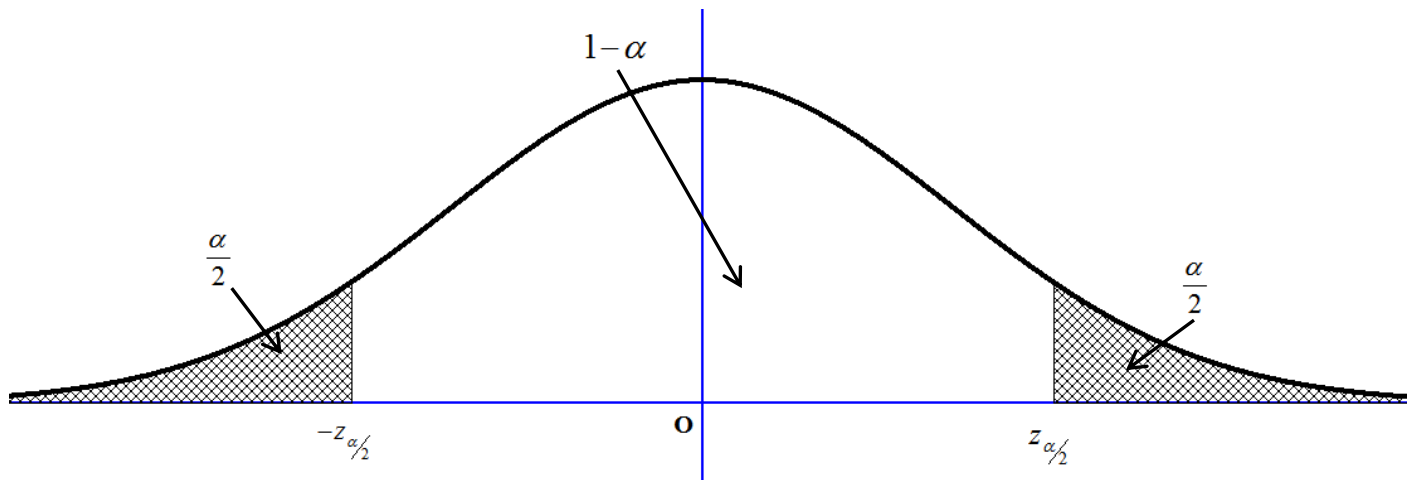
Dado un nivel de confianza $1 - \alpha$ (o lo que es lo mismo, nivel de significación o riesgo, α), el número $z_{\alpha/2}$ es aquel nº real que, en la distribución normal estándar, $N(0;1)$, deja a su derecha un área igual a $\frac{\alpha}{2}$. Es decir,

$$P(Z \geq z_{\alpha/2}) = \frac{\alpha}{2}$$

Como, además, $P(Z \geq z_{\alpha/2}) = 1 - P(Z \leq z_{\alpha/2}) \Rightarrow 1 - P(Z \leq z_{\alpha/2}) = \frac{\alpha}{2} \Rightarrow P(Z \leq z_{\alpha/2}) = 1 - \frac{\alpha}{2}$ que es la

usada para calcular $z_{\alpha/2}$. También se puede usar $P(Z \leq z_{\alpha/2}) = \frac{1 + N_c}{2}$, que resulta creo más fácil, y son iguales pues $N_c = 1 - \alpha$.

Gráficamente podemos observar a que corresponde cada expresión



Donde se observa fácilmente que:

$$P(Z \geq z_{\alpha/2}) = \frac{\alpha}{2} \qquad P(Z \leq -z_{\alpha/2}) = \frac{\alpha}{2} \qquad P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

Esta última expresión, $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$, se usa para calcular los intervalos de confianza

Ejemplos:

- Para un nivel de confianza del 90%, $z_{\alpha/2} = 1,645$.

ya que: $1 - \alpha = 0,90$ y por tanto $\alpha = 0,10$, $\frac{\alpha}{2} = 0,05$ y $1 - \frac{\alpha}{2} = 0,95$.

$$P(Z \leq z_{\alpha/2}) = 1 - \frac{\alpha}{2} \Rightarrow P(Z \leq z_{\alpha/2}) = 1 - \frac{0,10}{2} \Rightarrow P(Z \leq z_{\alpha/2}) = 0,95 \Rightarrow z_{\alpha/2} = 1,645$$

O bien

$$P(Z \leq z_{\alpha/2}) = \frac{1 + N_c}{2} \Rightarrow P(Z \leq z_{\alpha/2}) = \frac{1 + 0,90}{2} \Rightarrow P(Z \leq z_{\alpha/2}) = 0,95 \Rightarrow z_{\alpha/2} = 1,645$$

- Para un nivel de confianza del 95%, $z_{\alpha/2} = 1,96$.

ya que: $1 - \alpha = 0,95$ y por tanto $\alpha = 0,05$, $\frac{\alpha}{2} = 0,025$ y $1 - \frac{\alpha}{2} = 0,975$.

$$P(Z \leq z_{\alpha/2}) = 1 - \frac{\alpha}{2} \Rightarrow P(Z \leq z_{\alpha/2}) = 0,975 \Rightarrow z_{\alpha/2} = 1,96$$

O bien

$$P(Z \leq z_{\alpha/2}) = \frac{1 + N_c}{2} \Rightarrow P(Z \leq z_{\alpha/2}) = \frac{1 + 0,95}{2} \Rightarrow P(Z \leq z_{\alpha/2}) = 0,975 \Rightarrow z_{\alpha/2} = 1,96$$

- Para un nivel de confianza del 99%, $z_{\alpha/2} = 2,575$

ya que: $1 - \alpha = 0,99$ y por tanto $\alpha = 0,01$, $\frac{\alpha}{2} = 0,005$ y $1 - \frac{\alpha}{2} = 0,995$

$$P(Z \leq z_{\alpha/2}) = \frac{1 + 0,99}{2} \Rightarrow P(Z \leq z_{\alpha/2}) = 0,995 \Rightarrow z_{\alpha/2} = 2,575$$

O bien

$$P(Z \leq z_{\alpha/2}) = 0,995 \Rightarrow z_{\alpha/2} = 2,575$$

CÁLCULO DEL INTERVALO DE CONFIANZA

La *distribución muestral de medias*, \bar{X} , sigue una distribución normal $\bar{X} = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

Por tanto, $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ sigue una distribución $N(0,1)$.

Según se ha definido el valor $z_{\alpha/2}$ se tiene que: $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$

Sustituyendo:

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

Multiplicando todos los miembros de las desigualdades por $\frac{\sigma}{\sqrt{n}}$

$$P\left(-z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Restando a todos los miembros de las desigualdades \bar{X}

$$P\left(-\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Cambiando el signo de los miembros de las desigualdades (cambia el sentido de la desigualdad).

$$P\left(\bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Ó lo que es lo mismo

$$P\left(\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

Si \bar{x} es una media muestral, diremos que el intervalo:

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

es un intervalo de confianza para la media con un nivel de confianza del $(1 - \alpha) \cdot 100\%$

Esto quiere decir, que el $(1 - \alpha) \cdot 100\%$ de los intervalos contruidos de esta manera contendrán a la media, es decir, el $(1 - \alpha) \cdot 100\%$ de las muestras posibles generarán intervalos de confianza que contendrán a la media.

Nota importante: No es correcto decir “el intervalo de confianza dado contiene a la media con una probabilidad $(1 - \alpha)$ ”, porque una vez dado un intervalo, éste o contiene a la media o no la contiene.

Para la distribución de proporciones muestrales el proceso es el mismo (en lugar de \bar{x} usaremos \hat{p}) y el intervalo de confianza que queda es:

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

Resumiendo, para encontrar el intervalo de confianza en el que se encuentra un parámetro poblacional (media o proporción) con un nivel de confianza $(1 - \alpha) \cdot 100\%$, hemos de seguir los siguientes pasos:

- Determinar el estimador muestral (medias muestrales o proporciones muestrales)
- Determinar la desviación típica correspondiente al estimador
- Determinar el valor crítico $z_{\alpha/2}$ correspondiente a ese nivel de confianza $N_c = (1 - \alpha) \cdot 100\%$
- El intervalo de confianza buscado es:

$$\left(\text{estimador} - z_{\alpha/2} \cdot \frac{\text{desviación típica}}{\sqrt{n}} , \text{estimador} + z_{\alpha/2} \cdot \frac{\text{desviación típica}}{\sqrt{n}} \right)$$

Tabla resumen:

Parámetro	Intervalo de confianza
Media μ	$\left(\bar{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} , \bar{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \right)$

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

NOTA: Recordemos que hemos supuesto que la media de la población (μ) era desconocida, la desviación típica (σ) era conocida y queríamos estimar la media, pero es razonable pensar que también la desviación típica sea desconocida, ¿qué hacer en esos casos?, pues para n suficientemente grande ($n \geq 30$), se puede sustituir σ por la desviación típica de la muestra $\hat{\sigma}$ en los cálculos de los intervalos de confianza y la estimación sigue siendo satisfactoria.

Análogamente, podemos decir de la proporción p en una población, pues ésta no será conocida. En su lugar usaremos la proporción de la muestra \hat{p} , con un tamaño de la muestra conveniente

Ejemplo: Tenemos una población con una desviación típica $\sigma = 2$. Tomamos una muestra de 100 individuos y obtenemos una media $\bar{x} = 8$. Queremos dar el intervalo de confianza para la media con un nivel de confianza del 95%.

Así: $1 - \alpha = 0,95$; por tanto $\alpha = 0,05$ y $\frac{\alpha}{2} = 0,025$

$$P(Z \leq z_{\alpha/2}) = 1 - \frac{\alpha}{2} = 0,975 \Rightarrow z_{\alpha/2} = 1,96$$

Por lo que el intervalo de confianza es: $\left(8 - 1,96 \cdot \frac{2}{\sqrt{100}}, 8 + 1,96 \cdot \frac{2}{\sqrt{100}} \right) = (7,608, 8,392)$

No sé si este intervalo de confianza contiene a la media, pero sé que el 95% de los intervalos de confianza creados de esta manera contienen a la media.

NOTA: La tabla siguiente recoge los valores críticos $Z_{\frac{\alpha}{2}}$ más utilizados:

%	$1 - \alpha$	α	$\frac{\alpha}{2}$	$Z_{\frac{\alpha}{2}}$
90 %	0,9	0,1	0,05	1,645
95 %	0,95	0,05	0,025	1,96
99 %	0,99	0,01	0,005	2,575

Ejemplo: Si conocemos que la desviación típica para la altura de las chicas de 18 años en Aragón es 10 cm. Supongamos que hemos tomado una muestra de 100 chicas sale que la altura media de la muestra: $\bar{X} = 170$. Hallar los intervalos de confianza para la altura media de las chicas de 18 años en Aragón para un nivel de confianza del 90%, 95% y 99%

Nivel de confianza del 90%: $\left(170 - \frac{10}{\sqrt{100}} \cdot 1,645, 170 + \frac{10}{\sqrt{100}} \cdot 1,645 \right) = (168,355, 171,645)$

Nivel de confianza del 95%: $\left(170 - \frac{10}{\sqrt{100}} \cdot 1,96, 170 + \frac{10}{\sqrt{100}} \cdot 1,96 \right) = (168,04, 171,96)$

Nivel de confianza del 99%: $(170 - \frac{10}{\sqrt{100}} \cdot 2,575, 170 + \frac{10}{\sqrt{100}} \cdot 2,575) = (167,425, 172,575)$

Ejemplo: Tomada, al azar, una muestra de 120 estudiantes de una Universidad, se encontró que 54 de ellos hablaban inglés. Halle, con un nivel de confianza del 90%, un intervalo de confianza para estimar la proporción de estudiantes que hablan el idioma inglés entre los estudiantes de esa Universidad.

Como vemos se trata de una muestra donde estudiamos la característica hablar inglés y no conocemos la proporción de la población. Por ello usaremos la proporción de la muestra $\hat{p} = \frac{54}{120} = 0,45$ y por tanto $\hat{q} = 1 - 0,45 = 0,55$. Tenemos

que $1 - \alpha = 0,90 \Rightarrow \alpha = 0,10 \Rightarrow \frac{\alpha}{2} = 0,05 \Rightarrow 1 - \frac{\alpha}{2} = 0,95 \Rightarrow P(Z \leq z_{\alpha/2}) = 0,95 \Rightarrow z_{\alpha/2} = 1,645$

Aplicando la fórmula $\left(\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} \right) = \left(0,45 - 1,645 \cdot \sqrt{\frac{0,45 \cdot 0,55}{120}}, 0,45 + 1,645 \cdot \sqrt{\frac{0,45 \cdot 0,55}{120}} \right) = (0,45 - 0,0747, 0,45 + 0,0747) = (0,3753, 0,5247)$

6. TAMAÑO DE LAS MUESTRAS. ERROR MÁXIMO ADMISIBLE

Al estudiar los intervalos de confianza hemos visto que su amplitud depende del siguiente factor:

$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ para la media ó $z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$ para la proporción.

Es decir, cuanto mayor sea el tamaño de la muestra más fiable es el intervalo elegido, pero también es mucho más caro el estudio. Las expresiones anteriores nos dan el error máximo que cometemos.

Definición: El **error máximo admisible** para la estimación de medias o proporciones viene dado, respectivamente, por:

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \text{ para la media o } E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}} \text{ para la proporción}$$

De estas expresiones deducimos que:

- A mayor nivel de confianza, mayor $z_{\alpha/2}$, y por tanto mayor es el error cometido
- A mayor tamaño muestral, menor es el error cometido

TAMAÑO DE LAS MUESTRAS

El valor $E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ es el error máximo cometido al hacer una estimación de la media.

Si fijamos de antemano un error máximo E, es decir, conocemos E, debe ocurrir que

$$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq E$$

Despejando:

$$z_{\alpha/2} \cdot \sigma \leq E \cdot \sqrt{n} \Rightarrow \frac{z_{\alpha/2} \cdot \sigma}{E} \leq \sqrt{n} \Rightarrow \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2 \leq n$$

Luego:

$$n \geq \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$$

De forma análoga, para la proporción si despejamos en $E = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$, nos queda que

$$n \geq \frac{z_{\alpha/2}^2 \cdot \hat{p} \cdot \hat{q}}{E^2}$$

No hace falta aprenderse estas fórmulas, simplemente sustituyendo cada factor por su valor resulta más fácil despejar

Ejemplo: Tenemos una población con una desviación típica $\sigma = 2$. Queremos tomar una muestra para estimar la media con un nivel de confianza del 95%, ¿cuál debe ser el tamaño mínimo de la muestra para cometer un error menor de 0,4?

$$\text{Así: } 1 - \alpha = 0,95 ; \text{ por tanto } \alpha = 0,05 \text{ y } \frac{\alpha}{2} = 0,025 \Rightarrow P(Z \leq z_{\alpha/2}) = 0,975 \Rightarrow z_{\alpha/2} = 1,96$$

$$\text{Como se ha de cumplir que } z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq E \Rightarrow 1,96 \cdot \frac{2}{\sqrt{n}} \leq 0,4 \Rightarrow \frac{3,92}{\sqrt{n}} \leq 0,4 \Rightarrow \frac{3,92}{0,4} \leq \sqrt{n} \Rightarrow \sqrt{n} \geq 9,8 \Rightarrow n \geq 96,04$$

Luego la muestra debe ser al menos de 97 individuos.

Ejemplo: Deseamos conocer el nº de personas mayores de edad que sería necesario incluir en una muestra nacional con un error absoluto de 0,04 y un nivel de confianza del 99,73%. Se dispone de un valor de $p = 0,45$.

Para ese nivel de confianza,

$$1 - \alpha = 0,9973 \Rightarrow \alpha = 0,0027 \Rightarrow \frac{\alpha}{2} = 0,00135 \Rightarrow 1 - \frac{\alpha}{2} = 0,99865 \Rightarrow P\left(Z \leq z_{\alpha/2}\right) = 0,99865 \Rightarrow z_{\alpha/2} = 3$$

$$\text{Aquí aplicamos la fórmula } n \geq \frac{z_{\alpha/2}^2 \cdot \hat{p} \cdot \hat{q}}{E^2} \Rightarrow n \geq \frac{3^2 \cdot 0,45 \cdot 0,55}{0,04^2} \Rightarrow n \geq 1392,19$$

Se necesita para la muestra al menos 1393 personas