INFERENCIA ESTADÍSTICA

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:

- $\overline{\mu}$ ó $\bar{x}$, la media muestral
- $\overline{\sigma}$ 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 los n elementos de la muestra. El muestreo aleatorio simple 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 nº $k = \frac{N}{n}$ (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 $N_1, N_2, N_3, ...N_k$ al nº de elementos de la población que tiene cada estrato, es decir su tamaño. Se cumple que $N_1 + N_2 + N_3 + ... + N_k = N$. Sean $n_1, n_2, n_3, ..., n_k$ al número de individuos de los respectivos estratos que hay en la muestra (con $n_1 + n_2 + n_3 + ... + n_k = n$).

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, $n_1 = n_2 = n_3 = ... = n_k = \frac{n}{k}$. 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 $\frac{n_1}{N_1} = \frac{n_2}{N_2} = \frac{n_3}{N_3} = ... = \frac{n_k}{N_k} = \frac{n}{N}$

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:

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

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 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 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, $\overline{X_n} \to N(\mu, \frac{\sigma}{\sqrt{n}})$.

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, es decir, el error típico de las

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

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 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

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:

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

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:

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

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

c) ¿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:

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

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

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

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

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

El nº de muestras esperado es aproximadamente 74

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

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

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

<u>Distribución de las proporciones muestrales</u>

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**.

<u>Teorema</u>:

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 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, $p = \mu_{\hat{p}}$, por ser $\hat{p}$ un estimador insesgado.

<u>Ejemplo</u>: 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

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

$\sigma_{\hat{p}} = \sqrt{\frac{p \cdot q}{n}} = \sqrt{\frac{(0,25)\cdot(0,75)}{2}} = 0,3062$, 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) $P(\hat{p} > 0,05) = \text{(tipificamos) } P\left(Z > \frac{0,05-0,03}{0,0076}\right) = 1 - P(Z \le 2,63) = 1 - 0,9957 = 0,0043$

b) $P(\hat{p} \le 0,01) = \text{(tipificamos) } P\left(Z \le \frac{0,01-0,03}{0,0076}\right) = P(Z \le -2,63) = P(Z > 2,63) =$

$= 1 - P(Z \le 2,63) = 1 - 0,9957 = 0,0043$

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 se:

- *Estimador puntual insesgado*: Si la media de la distribución muestral de un estadístico es igual a su correspondiente parámetro poblacional. Son estimadores puntuales insesgados 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.