ESTADISTICA INFERENCIAL

La **Estadística Inferencial** se ocupa de inferir o deducir las características de la población a partir de las características de las muestras.

Distinguiremos :

- **Parámetros poblacionales**. Son los índices centrales y de dispersión que definen a una población.
- **Estadísticos muéstrales**. Son los índices centrales y de dispersión que definen a una muestra.

MUESTREO

En la inferencia estadística es necesario utilizar muestras. La característica más importante que debe poseer una muestra es la representatividad, es decir que represente a la población. Para ello se utilizan las **técnicas de muestreo.**

- **Muestreo con remplazamiento:** Cada elemento de la población puede seleccionarse más de una vez.
- **Muestreo sin remplazamiento:** Cada elemento de la población no puede seleccionarse más de una vez.
- **Muestreo no aleatorio:** Los elementos de la población no tienen la misma probabilidad de ser incluidos en la muestra. La muestra elegida suele ser poco representativa.
- **Muestreo aleatorio:** cada miembro de la población tiene la misma probabilidad de ser incluido en la muestra. La muestra elegida es representativa. Se pueden conocer los errores cometidos y pueden hacerse inferencias válidas.

Simple: Es el más sencillo y sirve de base para todos los demás. Se parte de un listado de los elementos de la población y se seleccionan aleatoriamente n de ellos que constituyen la muestra. La elección se puede hacer asignándoles un número a cada elemento de la población y utilizar una urna o una tabla de números aleatorios.

Sistemático: Es una variante del simple. Conocidos N (tamaño de la población) y n (tamaño de la muestra), se divide $\frac{N}{n}$ y la parte entera del cociente k, nos indica que hemos de seleccionar los elementos de k en k, eligiendo al azar previamente el primero de ellos entre los k primeros elementos. La ventaja es que solo hay que determinar al azar un elemento.

Estratificado: Se divide la población en subgrupos o estratos homogéneos en los cuales se toman muestras aleatorias simples. La ventaja es que todas las partes en que la población se divide estarán representadas adecuadamente.

Si N_1,..., N_k es el nº de elementos en cada estrato (N_1+... +N_k = N) se elige el tamaño de la muestra n_i (n_1 +... + n_k = n) de forma que

$$\frac{n_1}{N_1} = \ldots = \frac{n_k}{N_k} = \frac{n}{N}$$

Este método recibe el nombre de muestreo estratificado proporcional. Dentro de cada estrato se puede aplicar el muestreo simple o sistemático para escoger los n_i elementos de la muestra.

ESTIMACIÓN POR PUNTOS

El estudio de determinadas características de una población se efectúa a través de diversas muestras que pueden extraerse de ella. Los estadísticos obtenidos de las muestras nos van a permitir decidir sobre los parámetros de la población. Para ello necesitamos conocer la distribución muestral de estos estadísticos

La idea de inferencia es la de deducción arriesgada. Estas inferencias se hacen a partir de los parámetros muestrales (estos parámetros suelen llamarse estimadores). El estimador más utilizado es la media muestral, o la proporción muestral.

Distribución muestral de medias

Si de una población de tamaño N se toman muestras de tamaño n, las medias de estas muestras forman una distribución de medias muestrales.

Si las medias muestrales de tamaño n han sido extraídas de una población normal N(μ , σ), la distribución de las medias muestrales se ajusta a una normal $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

La distribución de medias muestrales es normal incluso en el caso de que éstas procedan de poblaciones no normales, siempre que el tamaño dela muestra sea suficientemente grande (n≥30).

Teorema central del límite:

Al igual que en las N(μ , σ), la variable de partida se tipifica mediante el cambio

$$Z = \frac{X - \mu}{\sigma}$$

las $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$, de las medias muestrales de tamaño n se tipificará con

$$Z = \frac{\overline{x}_i - \mu}{\sigma / \sqrt{n}}$$

Distribución muestral de proporciones

En muchas situaciones se plantea estimar una proporción o porcentaje. Esto ocurre cuando la variable aleatoria puede tomar solamente dos valores diferentes: Si / no; Votantes de favor / votantes en contra, defectuoso / no defectuoso; etc...

En estos casos decimos que la población sigue una distribución binomial. Cuando el tamaño de la población es grande, la distribución binomial se aproxima a una normal.

Si llamamos p al parámetro poblacional que representa la proporción de uno de estos valores (éxito), entonces la proporción del otro valor (fracaso) es q = 1 – p

Si consideramos todas las muestras de tamaño n que pueden extraerse de la población, cada muestra determina un estadístico proporcional $\hat{P}$ de la variable.

Esta distribución se aproxima a una normal para valores grandes de n (n>30) por lo que puede estudiarse como una normal

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

teniendo en cuenta $\begin{cases} \hat{P} = p \\ \sigma_{\hat{P}} = \sqrt{\frac{p \cdot q}{n}} \end{cases}$

los valores de esta distribución se tipificaran con

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

Distribución muestral de diferencia de medias

Cuando estudiamos conjuntamente dos colectivos, se consideran los siguientes estadísticos

- Medias: $\begin{cases} 1^a \text{ Colectivo} : \mu_1 \\ 2^a \text{ Colectivo} : \mu_2 \end{cases}$
- Desviaciones típicas: $\begin{cases} 1^a \text{ Colectivo} : \sigma_1 \\ 2^a \text{ Colectivo} : \sigma_2 \end{cases}$
- Tamaño de las muestras: $\begin{cases} 1^a \text{ Colectivo} : n_1 \\ 2^a \text{ Colectivo} : n_2 \end{cases}$

Si las dos poblaciones siguen distribuciones normales $N(\mu_1;\sigma_1)$ y $N(\mu_2;\sigma_2)$ o bien, si ambas poblaciones tienen distribuciones cualesquiera y las respectivas muestras son de tamaño n_1, $n_2>30$, entonces la distribución muestral de diferencia de medias sigue una distribución normal

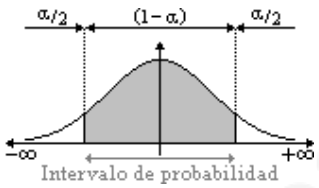
$$N\left(\mu_1-\mu_2,\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$$

Estimación por intervalos de confianza

La estimación por puntos se utiliza poco, es más útil utilizar la estimación por intervalos, que consiste en calcular dos valores que definen el intervalo en el cual estimamos se encontrará el parámetro poblacional con una probabilidad fijada de antemano.

Cuanto más amplio sea el intervalo, más probable será que incluya el valor estimado y mayor será el grado de confianza en que así sea.

A los intervalos simétricos respecto de la media o proporción poblacional se los denomina **intervalos de probabilidad**.

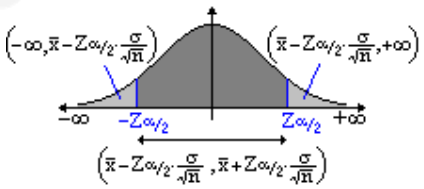


Se denomina **intervalos de confianza** al intervalo que, con una cierta probabilidad, contenga el parámetro que se está estimando.

Se denomina nivel de confianza $1-\alpha$, a la probabilidad de que el intervalo de confianza contenga al verdadero valor del parámetro, siendo α el riesgo o significación.

A cada nivel de confianza le corresponde un Z_c llamado **valor crítico** correspondiente a la N(0;1) y que cumple $P(|Z|\le Z_c)=1-\alpha$.

El $100\cdot(1-\alpha)\%$ de las muestras de tamaño n tendrá una media comprendida entre $\left(\bar{x}-Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\ ,\ \bar{x}+Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right)$, con un nivel de confianza $(1-\alpha)\%$, quedando el $100\cdot\alpha\%$ restante fuera del intervalo, repartidos al 50% entre exceso y defecto, él $100\cdot\alpha/2\,\%$ tendrá su media en el intervalo $\left(\bar{x}+Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\ ,\ +\infty\right)$(exceso) y el otro $100\cdot\alpha/2\,\%$ pertenecerá a $\left(-\infty\ ,\ \bar{x}-Z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right)$(defecto



Los niveles de confianza y sus valores críticos, más utilizados son:

1−α %	99'73	99	98	98	95'45	95	90	80	68'27	50
Z_c	3	2'58	2'53	2'05	2	1'96	1'645	1'28	1	0'6745

Si el intervalo de confianza es un intervalo de probabilidad, el intervalo del parámetro poblacional que estimemos vendrá dado por la siguiente tabla:

Parámetros	Intervalos 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	$\left(p - Z_{\alpha/2} \cdot \sqrt{\frac{P \cdot Q}{n}}, p + Z_{\alpha/2} \cdot \sqrt{\frac{P \cdot Q}{n}}\right)$
Diferencias de medias	$\left(\bar{x}_1 - \bar{x}_2 - Z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + Z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$

TAMAÑO DE LAS MUESTRAS

La determinación del tamaño de las muestras para que sean representativas depende del error máximo que queramos admitir:

- En el caso de estimación de una media, el intervalo de confianza es de la forma $\left(\bar{x} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$ por tanto el **error máximo** que se puede cometer, al nivel de confianza $1-\alpha$ es $E > Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ y el **tamaño de la muestra** es $n > \frac{Z_{\alpha/2}^2 \cdot \sigma^2}{E^2}$
- En el caso de estimación de una proporción, el intervalo de confianza es de la forma $\left(p - Z_{\alpha/2} \cdot \sqrt{\frac{P \cdot Q}{n}}, p + Z_{\alpha/2} \cdot \sqrt{\frac{P \cdot Q}{n}}\right)$ por tanto el **error máximo** que se puede cometer, al nivel de confianza $1-\alpha$ es $E > Z_{\alpha/2} \cdot \sqrt{\frac{P \cdot Q}{n}}$ y el **tamaño de la muestra** es $n > \frac{Z_{\alpha/2}^2 \cdot P \cdot Q}{E^2}$