Inferencia Estadística Paramétrica

a. Parámetros

*Un **Parámetro**, es una característica medible de los elementos de una población, también llamado **valor poblacional**, que la caracteriza e identifica con relación a otras poblaciones.*

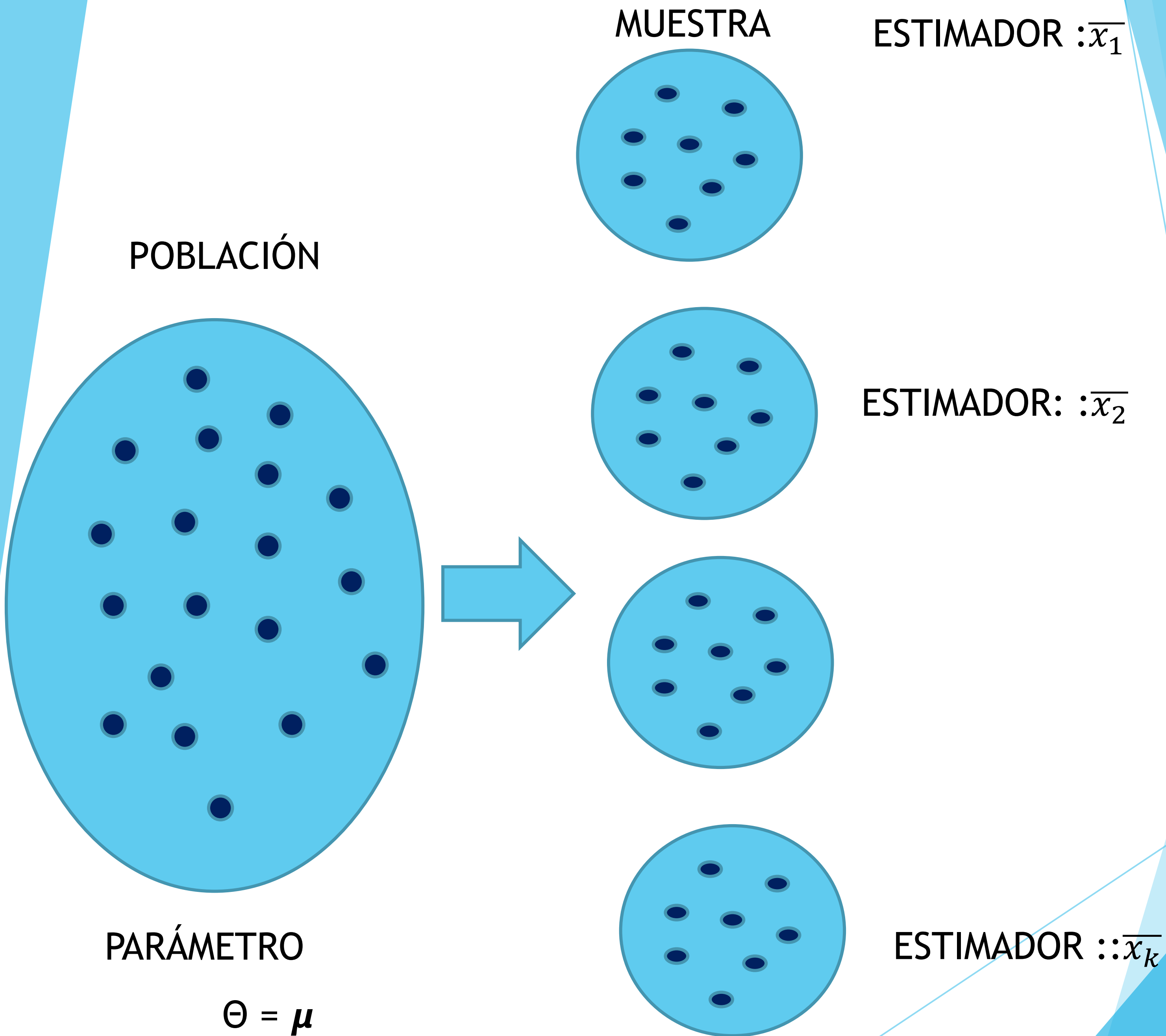
Ejemplos:

- Promedios
- Totales
- Proporciones
- Cantidad de Casos Favorables
- Razones (Proporciones, Promedios, Tasas, etc.).
- Variancias y Desvíos Estándar

b. Estimadores

En términos generales, se denomina ***Estimador*** *a una expresión matemática, que se construye con los valores correspondientes a los elementos que integran una muestra, y que como su nombre lo indica, sirven para obtener* ***estimaciones*** *del correspondiente valor poblacional o parámetro desconocido que se intenta estimar o verificar.*

El concepto anterior permite remarcar una diferencia sustancial dentro de la Inferencia Estadística: en el muestreo probabilístico los **Parámetros** deben ser considerados **constantes** usualmente desconocidas, en cambio los **Estimadores** son siempre **variables aleatorias**.


MUESTRA
ESTIMADOR :$\overline{x_1}$
POBLACIÓN
ESTIMADOR: :$\overline{x_2}$
PARÁMETRO
ESTIMADOR ::$\overline{x_k}$
$\Theta = \boldsymbol{\mu}$

c. Estimación de Parámetros

“La Estimación de ***Parámetros****, es la acción de aproximarse a los valores característicos y desconocidos de una Población, mediante la aplicación de los resultados de una muestra a un* ***Estimador****”.*

d. Estimación Puntual

*Como su nombre lo indica, la “****estimación puntual****”, consiste en aplicar la fórmula o definición de un* ***Estimador*** *a los resultados obtenidos en una muestra, y obtener una cifra o valor como Estimación del* ***Parámetro*** *en estudio.*

Los Parámetros y sus Estimadores

Estimación de la Media o Promedio

Es el Estimador Muestral que usualmente se toma como punto de partida para analizar el tema de Estimación de Parámetros.

Siendo $\boxed{\mu = \Sigma x_i \ / \ N}$ el **Parámetro** a estimar.

Será la **Media Muestral** su **Estimador**: $\boxed{\bar{x} = \Sigma x_i \ / \ n}$

Estimación del Total

Siendo $\boxed{X = \Sigma\, x_i}$ el **Parámetro** a estimar.

Será la Media muestral multiplicada o expandida por la población N su estimador:

$$\boxed{\hat{x} = N \,.\, \bar{x}}$$

Estimación de la Varianza

Siendo $\boxed{\sigma_x{}^2 = \Sigma\,(x_i - \mu)^2 \,/\, N}$ el **Parámetro** a estimar.

Uno de sus estimadores es: $\boxed{S_x{}^2 = \Sigma\,(x_i - \bar{x})^2 \,/\, n - 1}$

Estimación de la Proporción

Siendo $\boxed{P = N_A / N = \Sigma\, x_i \;/\; N}$

(donde $\mathbf{x_i}$ se suma a través de los valores de la Población **N** y vale 1 si corresponde a la categoría A y 0 si no corresponde a A)

el **Parámetro** a estimar, será su **Estimador**:

$$\boxed{p = \hat{P} = n_A / n = \Sigma\, x_i \;/\; n}$$

Estimación de la Cantidad de Casos

Siendo $\boxed{A = \Sigma\, x_i}$ $= N_A$ el **Parámetro** a estimar,

y será su **Estimador**: $\boxed{\hat{A} = N \,.\, p}$ $= \hat{N}_A$

RESUMEN

Parámetros		*Estimadores*	
Promedio Poblacional	$\mu = \Sigma\, x_i \;/\; N$	*Promedio Muestral*	$\bar{x} = \Sigma\, x_i \;/\; n$
Total Poblacional	$X = \Sigma\, x_i$	*Promedio Muestral Expandido*	$\hat{X} = N \,.\, \bar{x}$
Proporción Poblacional	$P = N_A / N = \Sigma\, x_i \;/\; N$	*Proporción Muestral*	$p = \hat{P} = n_A / n = \Sigma\, x_i \;/\; n$
Cantidad de Casos Favorables	$A = N_A = \Sigma\, x_i$	*Proporción Muestral Expandida*	$\hat{A} = \hat{N}_A = N \,.\, p$
Variancia Poblacional	$\sigma_x^{\,2} = \Sigma\, (x_i - \mu)^2 / N$	*Variancia Muestral*	$S_x^{\,2} = \Sigma\, (x_i - \bar{x})^2 / n - 1$

Ejemplo 1:

En la ciudad de Buenos Aires, una empresa lechera desea determinar el consumo promedio mensual de manteca de la población, para lo cual encarga realizar una encuesta a 300 personas, de la que se obtiene que entre todos los encuestados consumen por mes 225 kg. de manteca. Estimar el consumo medio de manteca de los cerca de 3.000.000 habitantes de la ciudad.

Se desea estimar *: el consumo mensual medio de manteca por habitante de la población de la ciudad de Buenos Aires.*

Datos disponibles*:*

- x_i *: consumo mensual de manteca del habitante i de la ciudad de Buenos Aires*
- *N = 3.000.000 total de habitantes de la ciudad (aproximadamente)*
- *n = 300 muestra de habitantes extraída para hacer el estudio*
- Suma *xi= 225 kg. consumo de manteca de las 300 personas de la muestra*
  La Estimación se realiza a partir del estimador

que representa la estimación del consumo mensual medio de manteca por habitante.

$$\bar{x} = \frac{225}{300} = 0{,}75\ kg\ por\ persona$$

Ejemplo:

En el ejemplo del punto anterior, se desea estimar el consumo total mensual de manteca de los cerca de 3.000.000 habitantes de la ciudad. Suponer que se extrae la misma muestra con el mismo resultado.

Se desea estimar X*: el consumo Total mensual de manteca de los habitantes de la ciudad de Buenos Aires.*

Datos disponibles*:*

- x_i : *consumo mensual de manteca del habitante i de la ciudad de Buenos Aires*
- *N = 3.000.000 total de habitantes de la ciudad (aproximadamente)*
- *n = 300 muestra de habitantes extraída para hacer el estudio*
- Suma *xi = 225 kg. consumo de manteca de las 300 personas de la muestra*

La Estimación se realiza a partir del estimador

$$X = N\bar{x} = 3.000.000 \ \frac{225}{300} = 3.000.000 \ 0{,}75 \ kg \ = 2.250.000 \ kg$$

que representa la estimación del consumo mensual de manteca por todos los habitantes de la ciudad de Buenos Aires.

Ejemplo:

Continuando con el ejemplo anterior, se desea estimar la proporción de habitantes que consume habitualmente manteca en la ciudad. En este caso la muestra indica que 240 personas refieren consumir habitualmente manteca.

***Se desea estimar P**: la proporción de personas que consume manteca entre los habitantes de la ciudad de Buenos Aires.*

***Datos disponibles**:*

- x_i *: consumo de manteca del habitante i de la ciudad de Buenos Aires: vale 1 si la consume y 0 si no la consume*
- *N = 3.000.000 total de habitantes de la ciudad (aproximadamente)*
- *n = 300 muestra de habitantes extraída para hacer el estudio*
- n_A = suma *xi*

$$\hat{P} = p = \frac{240}{300} = 0{,}80$$

La Estimación se realiza a partir del estimador:

*que representa la estimación de la proporción de personas que consumen manteca de los habitantes de la ciudad de Buenos Aires. Expresado en % sería que un **80 %** de los habitantes de la ciudad consumen manteca habitualmente.*

Si queremos estimar el total de personas que consumen Manteca

$Na = N \, . \, P = 3.000.000*0.8 = 2.400.000$

Distribución de los Estimadores

La diferencia básica dentro de la Inferencia Estadística entre los **Parámetros** y los **Estimadores** reside en responder:

¿ Porqué un Parámetro es una constante ?

- Los **Parámetros** no dependen de los elementos de una muestra **n** en particular, sino de los **N** valores de la población definida; es decir que no cambia a menos que se modifique la población, lo que por definición es imposible.
- En el muestreo, el valor de los parámetros es usualmente desconocido. Sobre esos parámetros desconocidos y constantes es que se realizan las "inferencias" (estimaciones, comprobaciones, proyecciones, etc.).

¿ Porqué un Estimador es una variable aleatoria ?

- Porque de una misma Población se pueden extraer diferentes Muestras (N tomadas de a n diferentes Muestras), y las diferentes selecciones aleatorias del mismo tamaño, extraídas de la misma Población y con la misma metodología, producen diferentes estimaciones de los valores poblacionales.
- En la realidad se extrae una sola Muestra, lo que equivale a considerar a un solo valor del conjunto total de posibles valores que podría tomar cada Estimador.

Concepto de Estimador como Variable Aleatoria

Siendo que cada muestra genera un resultado distinto en un Estimador, si se pudieran seleccionar ***todas las muestras posibles a extraer de una población*** *y con ellas calcular todos los resultados que daría el Estimador, se obtendría una distribución de los diversos resultados posibles del mismo, a la que se denomina "****Distribución del Estimador en el Muestreo****".*

Este es un concepto teórico (en la práctica es imposible disponer de todas las muestras que pueden llegar a obtenerse y calcular con ellas las estimaciones del Parámetro).

Pero dado un Estimador en particular (por ejemplo: la Media Aritmética), y dadas ciertas condiciones o conocimientos previos de la población, se dispone de desarrollos teóricos que demuestran **la forma que tienen esas Distribuciones Muestrales**, y cuales son sus parámetros.

La Distribución de la Media Muestral

Puede comprobarse sin demasiadas complicaciones algebraicas que:

- Si de una población de **N** unidades con una variable **x** de media μ *y variancia* σ^2;
- Se extrae una muestra de tamaño **n** unidades en la que se mide la misma variable con el fin de estimar la media μ ;
- El **Estimador Media Muestral** tendrá una distribución cuya media será también μ, es decir:

$$\mathbf{E}\,(\bar{\mathbf{x}}) = \mu\,,$$

y cuya variancia será:

$$\mathbf{Var}\,(\bar{\mathbf{x}}) = [\sigma^2 / \mathbf{n}]$$

La Forma de la Distribución de la Media Muestral

La conclusión de que:

- **E (x) = μ**] y que
- **Var (x) = [σ2 / n]**

se cumple sin limitaciones ni restricciones, y cualquiera sea la forma en que se distribuyan las medias muestrales.

Los elementos que influyen para determinar la **forma de la distribución de las medias muestrales** son tres:

- La forma de la distribución de la variable o población original.
- El conocimiento previo (o no) de la dispersión de la población.
- El tamaño de la muestra (n).

La conjugación de esos tres factores, determina la forma de la distribución de la media de las muestras.

De los tres, el más importante es **el tamaño de la muestra**, ya que el ***Teorema Central del Límite*** ha demostrado que **si una muestra elegida aleatoriamente es "grande"** (teóricamente se la considera **tendiendo a infinito**), la distribución de cualquier función lineal de valores muestrales tiende a distribuirse normalmente.

En este teorema se relaciona la teoría de las probabilidades con la teoría del muestreo.

El Teorema Central del Límite

El enunciado del Teorema, para el caso particular del **promedio muestral**, es el siguiente:

Si se tiene una variable aleatoria de una población cualquiera con promedio $\boldsymbol{\mu}$ *y variancia* $\boldsymbol{\sigma}^2$*, de la cual se extraen muestras de* $\boldsymbol{n}$ *unidades con las cuales se construye otra variable aleatoria que es el promedio muestral, el* ***Teorema Central del Límite*** *expresa y demuestra que si la muestra es suficientemente grande (teóricamente tendiendo a infinito), el promedio muestral tiende a distribuirse normalmente con una media que es el promedio poblacional* $\boldsymbol{\mu}$ *original y una variancia* $\boldsymbol{\sigma}^2/\boldsymbol{n}$ *multiplicada por el factor de corrección para poblaciones finitas.*

El Teorema Central del Límite (Cont.)

Es decir, cuando el tamaño de la muestra es suficientemente grande, no importa cual es la distribución de la variable original ni otros aspectos de la población de la cual se extrae la muestra.

El Teorema Central del Límite demuestra que la media muestral tiene una Distribución Normal de probabilidad.

La importancia de este Teorema reside en el hecho de que nada se dice acerca de la distribución de la variable original en la población ni de otros aspectos de la población de la cual se extrae la muestra, ya que con tal que la media y la variancia poblacionales sean finitas (situación usual), **cuando el tamaño de la muestra es grande la media muestral presenta una distribución Normal**.

El Teorema Central del Límite (Cont.)

En símbolos la expresión del **Teorema Central del Límite** es:

$x \sim$ ¿? $(\mu ; \sigma^2)$ con $E(x) = \mu$ y $V(x) = \sigma^2$

Se extrae una muestra de n variables aleatorias independientes $(x_1, x_2, \ldots, x_n)$, con la que se calcula la media muestral:

$$\boxed{\bar{x} = \Sigma\, x_i \;/\; n}$$

Si la muestra es "grande" $(n \sim \infty)$, la media aritmética presenta una distribución Normal con media μ y variancia σ^2 / n (por el factor de corrección para poblaciones finitas).

Es decir, expresándolo en forma más resumida:

$\bar{x} \sim N\,[\mu\, ;\, (\sigma^2 / n)]$

Intervalos de Confianza para la Media Poblacional μ

El **estimador puntual** de la media poblacional **μ** es la media muestral:

$$\bar{x} = \Sigma\, x_i \;/\; n$$

Sabemos que dadas ciertas condiciones el **estimador media muestral** tiene una **distribución Normal**, es decir:

$$\bar{x} \approx N\,[\mu\ ;\sigma^2_{\bar{x}} = \frac{\sigma^2}{n})$$

A partir de allí, estandarizando el estimador:

$$\mathbf{z} = (\bar{x} - \mu) / [(\sigma / \sqrt{n}) \,.\sim \, \mathbf{N}\,[0\,;\,1]$$

Y operando se llega : $\bar{x} \sim \mathbf{N}\,[\mu\ ;\sigma^2_{\bar{x}} = \sigma^2 / \mathbf{n})$

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

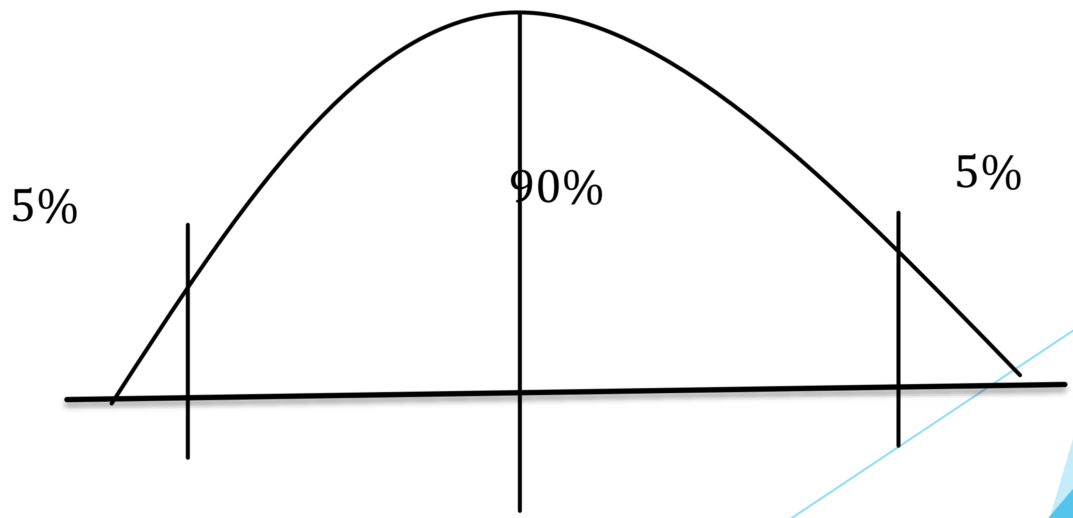
$$\boxed{P\left[\underbrace{\bar{x} - z\,.\frac{\sigma}{\sqrt{n}}}_{L_I} < \mu < \underbrace{\bar{x} + z\,.\frac{\sigma}{\sqrt{n}}}_{L_S}\right] = 1 - \alpha}$$

Supongamos que tenemos una distribución normal N(0,1).

Entonces para encontrar el intervalo de confianza debemos encontrar 2 valores simétricos respecto a la media que encierren un porcentaje que yo quiero conocer.

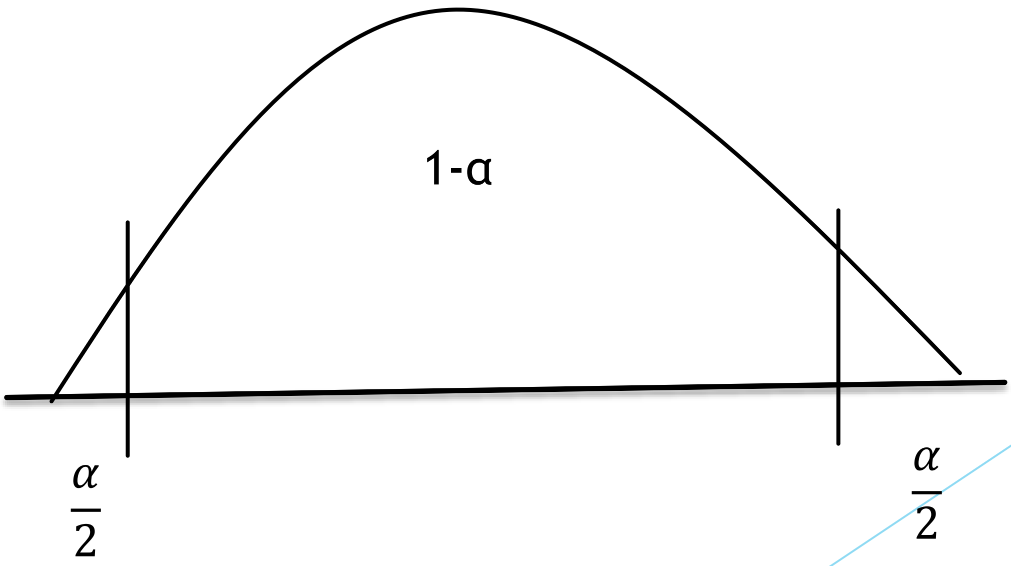
Por ejemplo: entre que valores está el 90% de los pesos de una pobalción?

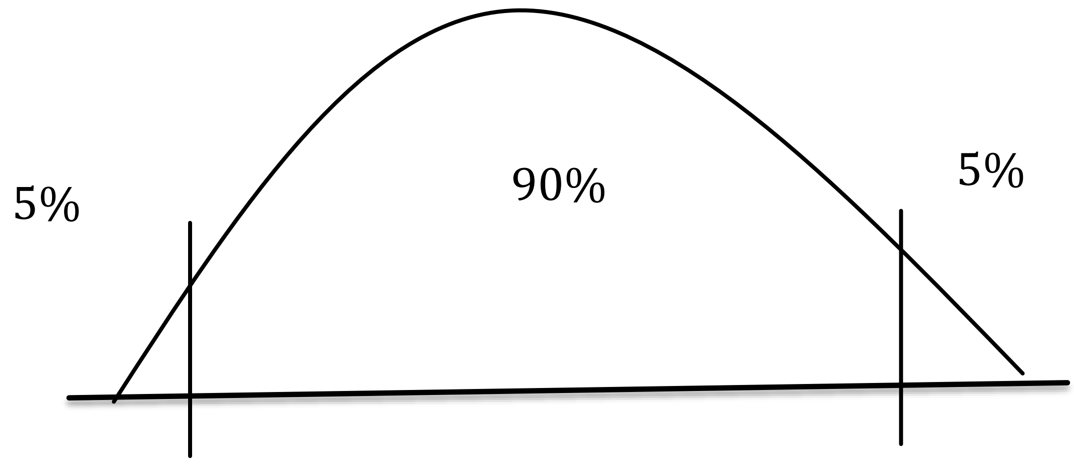
A este porcentaje se lo denomina nivel de confianza.



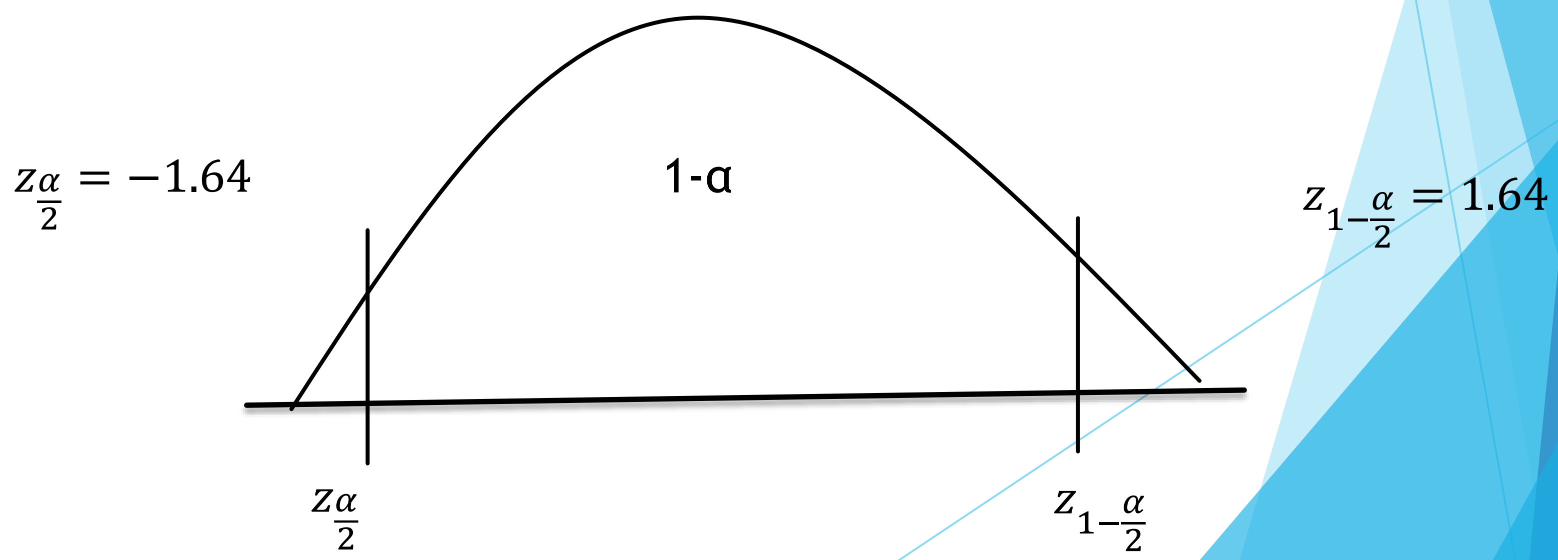
Cómo usar la tabla Normal N(0,1)?

La tabla nos da el área por debajo de un determinado valor.




5%
90%
5%

$$90\% + 5\% = 95\%$$


$z_{\frac{\alpha}{2}} = -1.64$
1-α
$z_{1-\frac{\alpha}{2}} = 1.64$
$z_{\frac{\alpha}{2}}$
$z_{1-\frac{\alpha}{2}}$

$$-1.64 \leq \mathrm{z} \leq 1.64$$

$$Z = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$$

$$-1.64 \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.64$$

$$-1.64.\frac{\sigma}{\sqrt{n}} \leq \bar{x}-\mu \leq 1.64.\frac{\sigma}{\sqrt{n}}$$

$$-\bar{x}-1.64.\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{x}+1.64.\frac{\sigma}{\sqrt{n}}$$

$$\bar{x}+1.64.\frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{x}-1.64.\frac{\sigma}{\sqrt{n}}$$

$$\bar{x}-z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq .\,\mu \leq \bar{x}+z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$$

$$\bar{x}-1.64\frac{\sigma}{\sqrt{n}} \leq .\,\mu \leq \bar{x}+1.64.\frac{\sigma}{\sqrt{n}}$$

$$P\left[\bar{x} - k\,.\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + k\,.\frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha$$

Donde, aparte de la Media Muestral y la poblacional se tiene:

1) K es el coeficiente de precisión que depende de la distribución del Estimador $\bar{x}$. En la Distribución de la Media Muestral se tenía que:

- si la variable se distribuye Normalmente con Variancia conocida aunque la muestra sea pequeña, o - si el tamaño de muestra es grande ($n \sim \infty$):	$\mathbf{k = z \sim N(0, 1)}$
- si la variable se distribuye Normalmente con Variancia desconocida y n chica:	$\mathbf{k = t \sim t_{S\,(n-1)}}$

Ejemplo 1)

En un estudio se pretende estimar la edad media a la que se diagnostica la Diabetes Mellitus en una ciudad de la Provincia de Buenos Aires. Para ello se dispone de una muestra de 100 pacientes a los que se les ha preguntado la edad de diagnóstico de la enfermedad. A partir de estos 100 pacientes se ha obtenido una edad media (muestral) de 48,78 años. Si es conocido, a raíz de otros estudios, que la desviación típica poblacional de esta variable (Edad de diagnóstico de la enfermedad) es σ = 16,32, calcula un intervalo de confianza al 95% para la edad media de diagnóstico de esta enfermedad en la región de estudio.

- Datos para realizar la estimación:

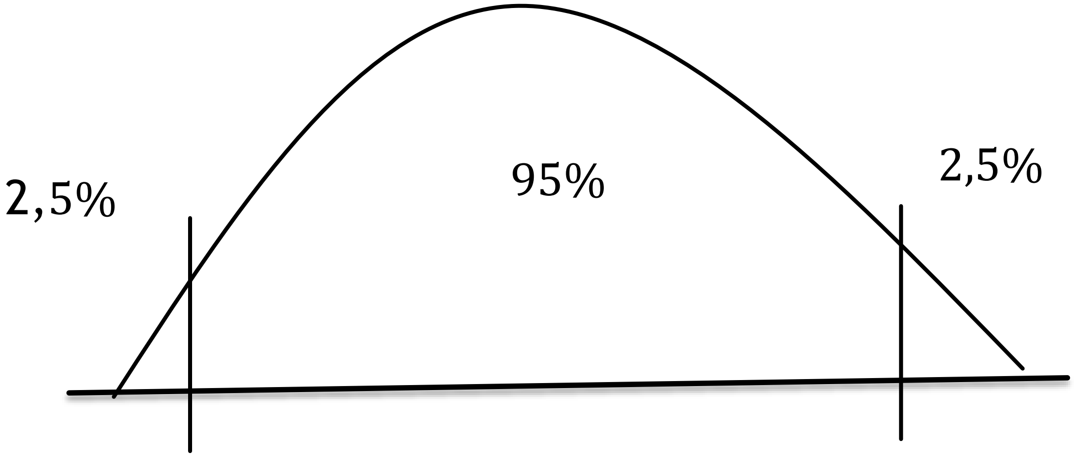
  $n = 100, \ \bar{x} = 48{,}78 \ \text{ y } \sigma = 16{,}32$

  Como queremos obtener un intervalo con un 95% de confianza, tenemos:

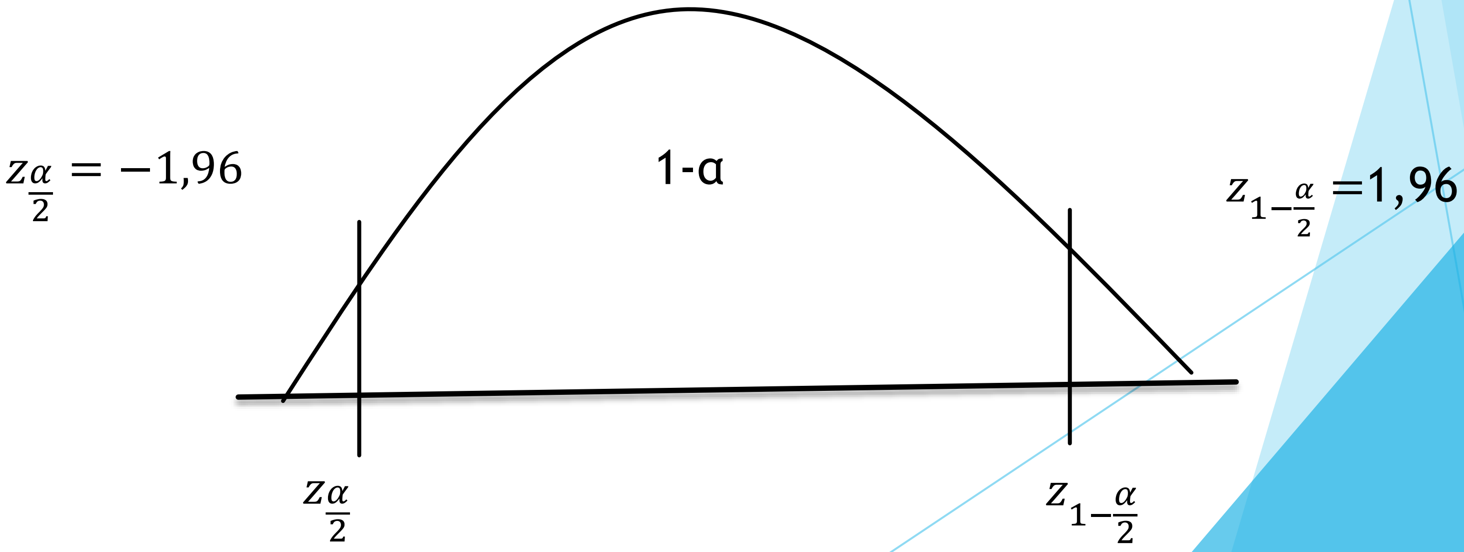
  $1 - \alpha = 0{,}95$, y por tanto $\alpha = 0{,}05$ y $\alpha / 2 = 0{,}025$, así, $(1 - \alpha / 2) = 0{,}975$.

  Por tanto debemos buscar el valor de la Normal estándar que cumple que el 97,5% de los valores son inferiores a el.

  Este valor de la N(0,1) es $Z_{1-\alpha/2} = 1{,}96$, y por tanto, el intervalo es:


2,5%
95%
2,5%

$$95\% + 2,5\% = 97,5\%$$


$z_{\frac{\alpha}{2}} = -1,96$
1-α
$z_{1-\frac{\alpha}{2}} = 1,96$
$z_{\frac{\alpha}{2}}$
$z_{1-\frac{\alpha}{2}}$

$$\left[48{,}78 - \mathbf{1,96}\,\frac{\mathbf{16{,}32}}{\sqrt{\mathbf{100}}} \leq .\, \boldsymbol{\mu} \leq 48{,}78 + \mathbf{1,96}\,\frac{\mathbf{16{,}32}}{\sqrt{\mathbf{100}}}\right] = [\mathbf{45,59}; \mathbf{51,98}]$$

$$\boldsymbol{P}(45{,}59 \leq \boldsymbol{\mu} \leq \mathbf{51,98}) = \mathbf{0,95}$$

Con un 95% de confianza, la edad media a la que se diagnostica la Diabetes será un valor contenido en el intervalo [45,59 , 51,98].

Ejemplo 2)

Siguiendo con el ejemplo anterior. En un estudio se pretende estimar la edad media a la que se diagnostica la Diabetes Mellitus. Para ello se dispone de una muestra de 21 pacientes a los que se les ha preguntado la edad de diagnóstico de la enfermedad. A partir de estos 21 pacientes se ha obtenido una edad media (muestral) de 48,78 años y una desviación típica muestral de 16,32. Calcula un intervalo de confianza al 95% para la edad media de diagnóstico de esta enfermedad en la región de estudio, suponiendo que tiene una distribución Normal.

- Tenemos como datos para realizar la estimación:

n = 21,

$\bar{x}$ = 48,78

S = 16,32.

$$P\left[\bar{x} - k\,.\frac{S}{\sqrt{n}} \leq \mu \leq \bar{x} + k\,.\frac{S}{\sqrt{n}}\right] = 1 - \alpha$$

- Como queremos obtener un intervalo con un 95% de confianza, tenemos

$1 - \alpha = 0{,}95$, y por tanto $\alpha = 0{,}05$ y $\alpha / 2 = 0{,}025$, así, $(1- \alpha / 2) = 0{,}975$

Debemos buscar el valor de la distribución t de Student con $n-1$ = 21-1= 20 grados de libertad que cumple que el 97,5% de los valores son inferiores a él.

Este valor de la t de Student es t(20,0,975) = 2,086, y por tanto, el intervalo que queríamos calcular tomaría la siguiente expresión:

$$P\left[48{,}78 - 2{,}086\,.\frac{16{,}32}{\sqrt{21}} \leq \mu \leq 48{,}78 + 2{,}086\,.\frac{16{,}32}{\sqrt{21}}\right] = P[\,41{,}35 \leq \mu \leq 56{,}21] = 0{,}95$$

Con un 95% de confianza, la edad media a la que se diagnostica la Diabetes Mellitus estará contenida en el intervalo [41,35 , 56,21], es decir, entre 41 y 56 años aproximadamente.

Intervalos de Confianza para la Proporción Poblacional P

El **estimador puntual** de la Proporción poblacional **P** es la Proporción muestral:

$$\boxed{\hat{P} = p = n_A / n}$$

Y sabemos que dadas ciertas condiciones ese **estimador** tiene una **distribución Normal**, es decir:

$$\sigma_p^2 = \frac{p.q}{n} \qquad \boxed{p \sim N(P\ ;\ \sigma_p^2)} \qquad \boxed{p \approx N(P\ ;\ \sigma_p^2)}$$

A partir de allí, estandarizando el estimador:

$$(p - P) / \sqrt{\frac{p.q}{n}} \sim N[0\ ;\ 1]$$; de donde se obtiene:

$$\boxed{P\left(p - z\,.\,\sqrt{\frac{p.q}{n}} < P < p + z\,.\,\sqrt{\frac{p.q}{n}}\right) = 1 - \alpha}$$