.UBA económicas
FACULTAD DE CIENCIAS ECONÓMICAS

Estadística I

Muestreo e Intervalos de Confianza

Natalia SALABERRY

Metodología

Conjunto de pasos que me permitirán obtener información a partir de los datos

Diseño → del plan y desarrollo del proceso de investigación

Descripción → de los datos, a través de la exploración

Inferencia → de un suceso, a partir de los datos (muestra)

Conclusión → toma de decisiones entorno al suceso

Población

Conjunto de todos los elementos del fenómeno a estudiar

Muestra

Sub conjunto de la población

Origen de los datos

Objetivo

Inferir (obtener medidas muestrales para inferir) características de la población

Técnicas de muestreos

Aleatorio Simple
Sistemático
Estratificado
Por conglomerados

Como la inferencia se realiza a partir de una muestra, entonces existirá un riesgo asociado: que la muestra no represente adecuadamente a la población

Medición del Riesgo

Intervalos de confianza

Muestra

Supongamos que se diseña un experimento y se selecciona un elemento para observar una propiedad X. En el primer ensayo se obtiene la observación X_1, en el segundo se obtiene X_2 y así sucesivamente hasta obtener n observaciones $\{X_1, X_2, \dots, X_n\}$. Entonces $\{X_1, X_2, \dots, X_n\}$ forma un conjunto de variables idénticamente distribuidas que constituyen una **muestra aleatoria** de la propiedad X.

El proceso de obtención de una muestra aleatoria puede realizarse mediante el empleo de distintas técnicas. A continuación, veremos la técnica muestreo aleatorio simple.

Muestreo Aleatorio Simple

Con este método, todos los elementos (N) de la población tienen la misma probabilidad $(1/N)$ de ser incluido en la muestra.

Cada extracción puede realizarse con reemplazo o sin reemplazo.

Sin reemplazo: se repite n veces el experimento para obtener n observaciones de la característica de interés. Cada observación es una VA cuya probabilidad de ocurrencia es igual a la de la población.

Con reemplazo: se extrae un elemento de la población , se observa la característica. Se devuelve el elemento a la población, se mezcla y se realiza una nueva extracción. Así sucesivamente hasta obtener n observaciones. Se obtienen VA independientes y con la misma probabilidad de ocurrencia que de la población.

Muestreo Estratificado

Cuando se trabaja con una población cuyos elementos poseen características diferentes (por ejemplo rangos de edad), se suele tomar muestras dentro de cada grupo que la conforma.

El método consiste en:

- Se toma una población U y se la divide en L subgrupos disjuntos de tamaño $N_1, N_2, \dots, N_L$ conformándose subpoblaciones que reciben el nombre de estratos. De esta forma:

$$\sum_{i=1}^{L} N_i = N$$

- La muestra final de la población será la conformada por todas las submuestras obtenidas en cada subpoblación o estrato.

Ejemplo

Se tiene una población de 1000 individuos repartida en 5 grupos:

El interés radica en seleccionar 100 individuos de los 1000.

Se calcula la submuestra en cada grupo (o estrato):

Muestra estratificada $1(N_1) = \frac{100}{1000} * 150 = 15$

Muestra estratificada $2(N_2) = \frac{100}{1000} * 250 = 25$

Muestra estratificada $3(N_3) = \frac{100}{1000} * 300 = 30$

Muestra estratificada $4(N_4) = \frac{100}{1000} * 200 = 20$

Muestra estratificada $5(N_5) = \frac{100}{1000} * 100 = 10$

Grupo	Individuos
1	150
2	250
3	300
4	200
5	100

Luego se verifica que:

$$\sum_{i=1}^{5} N_i = 15 + 25 + 30 + 20 + 10 = 100 = N$$

Muestreo Por Conglomerados

A diferencia del muestreo estratificado, el muestreo por conglomerados surge cuando no es posible dividir a los elementos de una población en grupos. De este modo, la muestra contendrá más de un elemento de la población. Por ejemplo, cuando en un país, no es posible enumerar a todos los habitantes del mismo para estudiar alguna característica en particular (el interés es la característica por ejemplo, presión alta sin importar a que provincia pertenece. Pero la población argentina ya se encuentra dividida en provincias, es decir, en conglomerados).

El método consiste en:

- Dividir a la población en conglomerados (por ejemplo, pensar en Argentina que se encuentra divido en provincias (estas son los conglomerados))
- En cada conglomerado tomar una muestra aleatoria simple con reemplazo
- Luego el total de la muestra es:

$\sum_{i=1}^{M} N_i = N$ donde i=1,…,M es cada conglomerado y N_i el tamaño de la muestra en cada conglomerado.

Al conformar a N_i con reemplazo, se puede incluir más de una vez al individuo del conglomerado.

Muestreo Sistemático

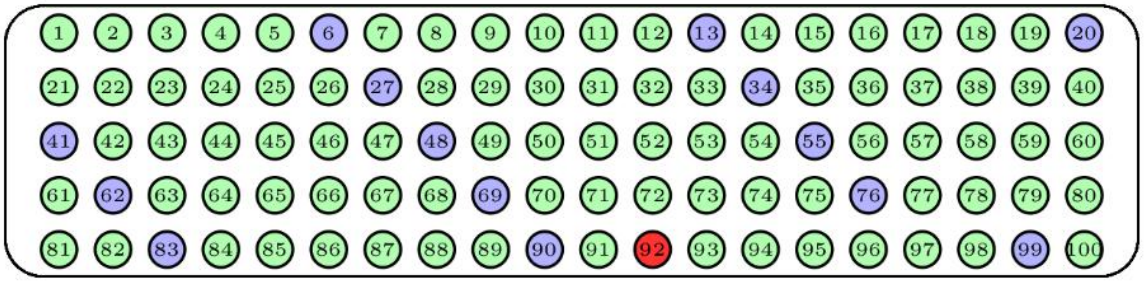
Para una población U de tamaño N, se elige un valor en el intervalo [1-k] con k aleatorio. Dicho intervalo se denomina intervalo de selección. Entonces, n será $\frac{N}{k}$ siendo un cociente que:

- Si el resto del cociente es 0, entonces se obtiene una muestra de tamaño exactamente n
- Si el resto del cociente es >0, entonces se obtendrá una muestra de tamaño n o n+1

El método consiste en:

- Elegir aleatoriamente un valor entre 1-k
- A partir del anterior, se suma la amplitud del intervalo (lo define quien realiza el muetsreo) para obtener el siguiente valor

Por ejemplo: de una urna que contiene 100 bolillas:



Se elige comenzar por el valor 92 (coloreado en rojo). A partir de ese valor, previamente definido un intervalo de amplitud 7, se le suma 7 y se extrae la bolilla 99. Se continúa con este procedimiento hasta agotar todas las posibilidades. En total se extraen 15 bolillas, confirmándose al muestra

Parámetros

Las características inferidas de la población se denominan *parámetros,* que son inferidos a partir de los estadísticos. Estos son desconocidos.

Están asociados a la distribución de probabilidad de una propiedad de la población de interés (por ejemplo, en la distribución normal tenemos el parámetro μ que será desconocido y puede ser inferido a partir de $\bar{X}$)

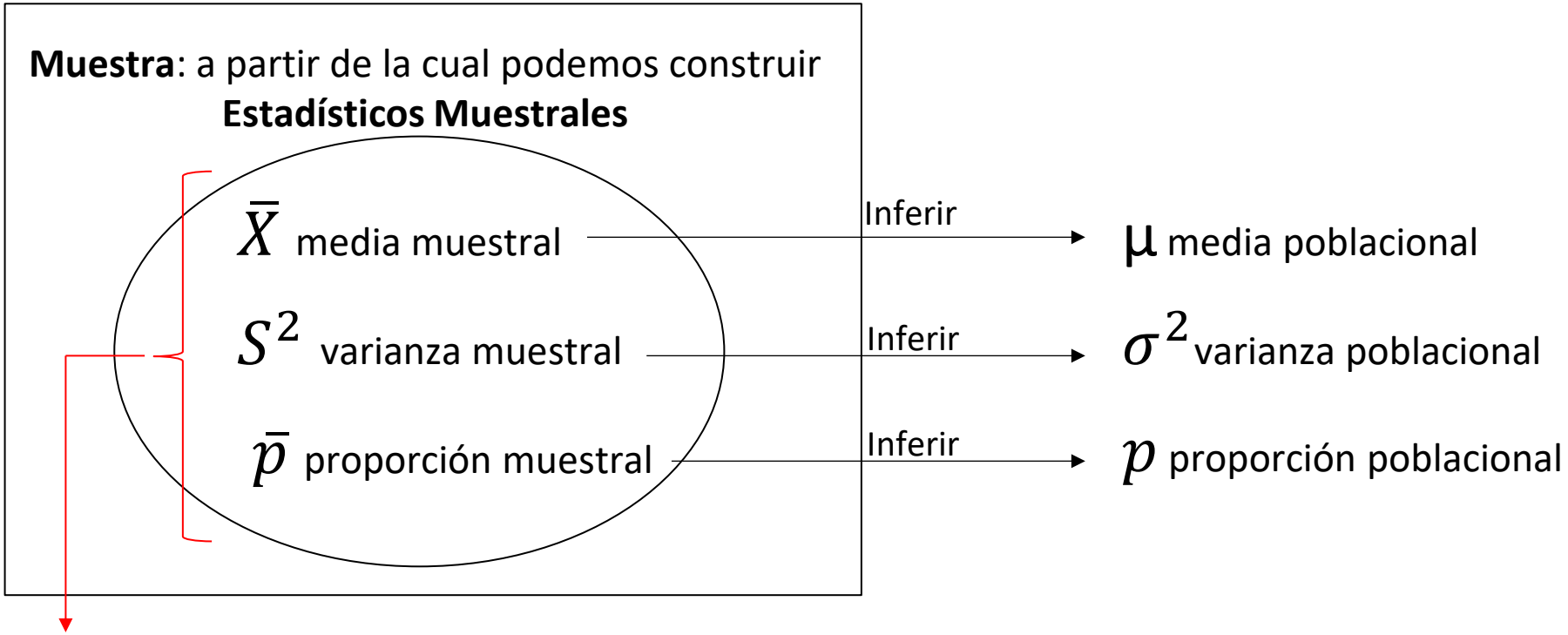
Estadísticos Muestrales

Dada una muestra $\{X_1, X_2, \dots, X_n\}$, se denomina *estadísticos muestrales* a las características observadas sobre esta, entonces son conocidos. Estos permiten realizar inferencias sobre características referidas a la población.

Distribución de muestreo

si tomamos todas las muestras posibles de tamaño n de la población y se calcula un estadístico, con la frecuencia observada de cada valor del estadístico se obtiene su distribución de muestreo. Por ejemplo, tomo una muestra y calculo $\overline{X_1}$, tomo otra muestra y calculo $\overline{X_2}$ y así sucesivamente hasta n. Entonces puedo obtener la distribución de muestreo del estadístico $\bar{X}$.

Población: tiene **Parámetros** que son desconocidos



La **inferencia** de los parámetros poblacionales será posible de realizar debido a que cada **estadístico muestral** tiene una **distribución muestral** asociada:

$\bar{X}$ media muestral: se presentan dos situaciones, entonces habrá dos distribuciones que son la Normal y T-Student

S^2 varianza muestral: Chi-Cuadrado

$\bar{p}$ proporción muestral: Normal

Los **estadísticos muestrales** deben cumplir propiedades deseables:

- Ser **insesgado**: es decir, si el valor esperado del estadístico es igual al parámetro poblacional.

$$E(estadístico\ muestral) = parámetro$$

Luego, el sesgo de un estimador se define como la diferencia entre el valor esperado y el parámetro:

$$\text{Sesgo del estadístico} = E(estadístico\ muestral) - parámetro$$

De donde, si un estadístico es insesgado entonces su sesgo es nulo.
Para lograrlo, se deberían tomar infinitas muestras posibles, lo cual lo hace muy difícil de conseguir en la práctica.
No obstante, podemos saber que $\bar{X}$, S^2 y $\bar{p}$ son insesgados:

Sea $\{X_1, X_2, \dots, X_n\}$ una muestra de variables aleatorias idénticamente distribuidas tales que E(X_i)= μ y V(X_i)=σ^2

$$E(\bar{X}) = E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right] = aplicando\ E(a) = a\ para\ a = \frac{1}{n} \Rightarrow \frac{1}{n}E[X_1 + X_2 + \cdots + X_n] =$$

$$desarrollando\ al\ sumatoria\ y\ distribuyendo\ el\ E = \frac{1}{n}[E(X_1) + E(X_2) + \cdots + E(X_n)] =$$

$\frac{1}{n}[\mu + \mu + \cdots + \mu] = \frac{n*\mu}{n}$= μ entonces $\bar{X}$ es un estimador insesgado de μ

$$E(S^2) = E\left[\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}\right] = aplicando\ E(a) = a\ para\ a = \frac{1}{n-1} \Rightarrow \frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i - \bar{X})^2\right] =$$

Restando μ a cada variable $= \frac{1}{n-1}E[\sum_{i=1}^{n}((X_i - \mu) - (\bar{X} - \mu))^2] =$

Esto es posible ya que $\frac{1}{n-1}E\left[\sum_{i=1}^{n}\left((X_i - \mu) - \bar{X} + \mu\right)^2\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left((X_i) - \bar{X}\right)^2\right]$ *siendo lo que teniamos previamente*

Desarrollando el cuadrado de un binomio teniendo en cuanta que $a = (X_i - \mu)$ *y* $b = (\bar{X} - \mu)$

$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left((X_i - \mu)^2 - 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2\right)\right] = distribuyendo\ la\ sumatoria$

$= \frac{1}{n-1}E[\sum_{i=1}^{n}(X_i - \mu)^2 - \sum_{i=1}^{n}2(X_i - \mu)(\bar{X} - \mu) + \sum_{i=1}^{n}(\bar{X} - \mu)^2]$ *(1)*

Trabajando el segundo término

$\sum_{i=1}^{n}2(X_i - \mu)(\bar{X} - \mu) = por\ propiedad\ de\ la\ sumatoria = 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \mu) =$
$distribuyendo\ la\ sumatoria = 2(\bar{X} - \mu)[\sum_{i=1}^{n}X_i - \sum_{i=1}^{n}\mu]$

re expresando $\sum_{i=1}^{n}X_i$ como n$\bar{X}$ ya que $\bar{X} = \frac{\sum_{i=1}^{n}x_i}{n} \Rightarrow n * \frac{\sum_{i=1}^{n}x_i}{n} = \sum_{i=1}^{n}x_i$

$\Rightarrow 2(\bar{X} - \mu)[\mathrm{n}\bar{X} - n\mu] = 2(\bar{X} - \mu)n\ (\bar{X} - \mu) = 2n(\bar{X} - \mu)^2 = sustituyendo\ en\ (1), entonces$

$= \frac{1}{n-1}E[\sum_{i=1}^{n}(X_i - \mu)^2 - 2n(\bar{X} - \mu)^2 + \sum_{i=1}^{n}(\bar{X} - \mu)^2] =$

$\frac{1}{n-1}E[\sum_{i=1}^{n}(X_i - \mu)^2 - 2n(\bar{X} - \mu)^2 + n(\bar{X} - \mu)^2] = \frac{1}{n-1}E[\sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2] =$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n}(X_i-\mu)^2 - n(\bar{X}-\mu)^2\right] = teniendo\ en\ cuenta\ que\ V(X) = E(X_i-\mu)^2 = \sigma^2 \text{ y que}$$

$$V(\bar{X}-\mu) = \left(\frac{\sum_{i=1}^{n}(X_i-\mu)}{n}\right)^2 = \sum_{i=1}^{n}\frac{1}{n^2}\sigma^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

$$= \frac{1}{n-1} E\left[\sum_{i=1}^{n}\sigma^2 - n\frac{\sigma^2}{n}\right] = \frac{1}{n-1} E\left[n\sigma^2 - n\frac{\sigma^2}{n}\right] == \frac{1}{n-1} E[n\sigma^2 - \sigma^2]$$

$$= \frac{1}{n-1}[E[n\sigma^2] - E[\sigma^2]] = aplicando\ E(a) = a\ en\ cada\ caso = \frac{1}{n-1}[n\sigma^2 - \sigma^2] =$$

$$= \frac{1}{n-1}[n-1]\sigma^2 = \sigma^2 \text{ entonces } S^2 \text{ es un estimador insesgado de } \sigma^2$$

$$E(\bar{p}) = E\left[\frac{\sum_{i=1}^{n} x_i}{n}\right] = aplicando\ E(a) = a\ para\ a = \frac{1}{n} \Rightarrow \frac{1}{n}E[X]$$

$$como\ p = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{X}{n} \Rightarrow X = np \Rightarrow = \frac{1}{n}E[np] = aplicando\ E(a) = a\ para\ a = np \Rightarrow \frac{np}{n} = p$$

entonces $\bar{p}$ es un estimador insesgado de p

- Ser **eficiente**: es decir, ser lo más estable posible entre diferentes muestras. Por lo tanto, se busca que el desvío sea lo más chico posible.

  Sean estadístico muestral 1 y estadístico muestral 2 insesgados.

  Si Var(estadístico 1) < Var(estadístico 2), entonces estadístico 1 es más eficiente

Se demuestra que si un estimador (estadístico) es insesgado, entonces es eficiente. Como $\bar{X}$, S^2 y $\bar{p}$ ya demostramos que son insesgados, entonces son eficientes

- Ser **consistente**: es decir, a medida que se aumenta el tamaño de la muestra se converge en probabilidad al valor del parámetro poblacional.

$$\lim_{n\to\infty} P\,(Estadístico\ -parámetro < \varepsilon) = 1$$

Esto es, la probabilidad de que la diferencia entre el estimador y el parámetro sea muy pequeña tiende al 100% cuando el tamaño de la muestra (n) crece.

Surge entonces que un estimador es **asintóticamente insesgado**, si su sesgo tiende a cero cuando el tamaño de la muestra tiende infinito.

Lo anterior se justifica con:

Ley de grandes números: Si $\{X_1, X_2, \dots, X_n\}$ es una muestra de variables aleatorias, cada una con E(X)= μ y V(X)= σ^2, entonces $\bar{X}$ es un estimador consistente de μ

Teorema de Chebyshevff: sea X una variable aleatoria con E(X)= μ y V(X)= σ^2 entonces

$$\lim_{n\to\infty} P(|X-\mu| \le k\sigma) \ge 1 - \frac{1}{k^2} \text{ o } \lim_{n\to\infty} P(|X-\mu| > k\sigma) \le \frac{1}{k^2}$$

Este teorema es muy importante ya que asegura que, independientemente de la distribución que tenga la variable aleatoria, la probabilidad de que una variable aleatoria se aleje no más de k desviaciones de su media, es menor o igual a $\frac{1}{k^2}$ para $k \ge 1$

Por ejemplo: $\lim_{n\to\infty} P(|X-\mu| \le 2\sigma) \ge 1 - \frac{1}{2^2} = \frac{3}{4}$ y $\lim_{n\to\infty} P(|X-\mu| > 2\sigma) \ge \frac{1}{2^2} = \frac{1}{4}$

Entonces P[$\frac{1}{4} < X - \mu < \frac{3}{4}$]=100% para $n \to \infty$

De aquí que en la consistencia, $\varepsilon = k\sigma$

- Ser **suficiente**: es decir, el estadístico muestral resulta suficiente cuando se utiliza una gran cantidad de información y ningún otro resulta mejor. $\bar{X}$, S^2 y $\bar{p}$ resultan ser suficientes.

De este modo, para poder utilizar los estadísticos muestrales $\bar{X}$, S^2 y $\bar{p}$ para estimar los parámetros poblacionales garantizando que sean insesgados y consistentes, surge el **Teorema Central del Límite (TCL).** Este teorema demuestra que, para cada muestra tomada, a medida que crece el tamaño de la muestra se tiende a una distribución normal estándar con media poblacional μ=0 y σ=1 donde la variable aleatoria Z= $\frac{\sum_{i=1}^{n} x_i - \mu}{\sigma\sqrt{n}}$ ahora vendrá dada por: $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0;1)$, con $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ que es insesgado por Ley de los grandes números y de donde es posible observar que a medida que n crece, $\frac{\sigma}{\sqrt{n}}$ será más chico; y a su vez, $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ también será más chica por lo que se aproximará a μ, es decir, es consistente.

Teorema Central del Límite v1. Sean $\{X_1, X_2, \dots, X_n\}$ variables aleatorias independientes e idénticamente distribuidas, siendo μ la media y σ^2 la varianza de cada una. Entonces, cuando $n \to \infty$, la distribución de la variable:

$Z_n = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n} * \sigma}$ converge a una distribución Normal Estándar. Es decir: $\lim_{n\to\infty} Z_n = Z \sim N(0;1)$

Teorema Central del Límite v2. Cuando se muestrea una población cualquiera con media μ y varianza σ^2, la variable:

$Z_n = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ tiene una distribución que tiende a Normal Estándar cuando $n \to \infty$.

Dado $Z = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\frac{\sum_{i=1}^{n} X_i}{n} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\frac{\sum_{i=1}^{n} X_i - n\mu}{n}}{\frac{\sigma}{\sqrt{n}}} = \frac{\sum_{i=1}^{n} X_i - n\mu}{n} \frac{\sqrt{n}}{\sigma} = \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}} \frac{1}{\sigma} = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sqrt{n} * \sigma}$ siendo la variable aleatoria Z definida por el TCL.

Es decir, que la distribución de $\lim_{n\to\infty} \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = Z \sim N(0;1)$

A partir de lo anterior, entonces el tamaño de la muestra resulta de mayor importancia y es viable de conseguir en la práctica. Además, si se trabaja con los estadísticos muestrales $\bar{X}$, S^2 y $\bar{p}$ sabemos que son insesgados y consistentes, y que en el límite tienen distribución normal.

Por ello,

- **Si cada $X_1, X_2 \dots X_n$ tiene distribución normal** y n<30 entonces $\sum_{i=1}^{n} X_i \sim$ N(n*μ; $\sqrt{n*\sigma^2}$) donde

$E(X_1, X_2 \dots X_n)$=n*μ V($X_1, X_2 \dots X_n$)=n*σ^2 σ=$\sqrt{n*\sigma^2}$

- **Si $X_1, X_2 \dots X_n$ son variables aleatorias independientes e idénticamente distribuidas (sin importar cual es la distribución)** y n (tamaño de la muestra) es lo suficientemente grande [n>30] entonces aplica el **Teorema Central del Límite** por el cual $\sum_{i=1}^{n} X_i$ tiene aproximadamente una distribución normal con $\mu_{\sum_{i=1}^{n} X_i}$=n*μ y $\sigma_{\sum_{i=1}^{n} X_i}$ =$\sqrt{n*\sigma^2}$

Por lo tanto, mientras cada $X_1, X_2 \dots X_n$ tiende a una distribución normal entonces

$$\sum_{i=1}^{n} X_i \sim \text{N(n*μ; } \sqrt{n*\sigma^2} \text{)}$$

Media Muestral $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

- **Situación 1:** Si $X_1, X_2 \ldots X_n$ son variables aleatorias independientes e idénticamente distribuidas y n (tamaño de la muestra) es lo suficientemente grande **[n>30]** entonces por **el Teorema Central del Límite**

$$\bar{X} \sim N(\mu; \frac{\sigma}{\sqrt{n}}) \Rightarrow Z_{obs} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

- **Situación 2:** Si la varianza poblacional es **conocida** y X_i tiene distribución normal con media conocida (y n <30) entonces:

$$\bar{X} \sim N(\mu; \frac{\sigma}{\sqrt{n}}) \Rightarrow Z_{obs} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Por lo tanto, tanto en situación 1 como en 2,

$$\bar{X} \sim N(\mu; \frac{\sigma}{\sqrt{n}}) \Rightarrow Z_{obs} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Ejemplo situación 1: Sea X una variable aleatoria normal. Hallar el mínimo tamaño de muestra que se debe tomar si se quiere que **la probabilidad de que la media muestral** sea menor que la media poblacional más 1/3 del desvío estándar, sea mayor que 0,9505.

Solución:

$X_i \sim N(\mu; \sqrt{\sigma^2}) \Rightarrow \bar{X} \sim N(\mu; \frac{\sigma}{\sqrt{n}})$

Entonces queremos el mínimo n tal que $P(\bar{X} < \mu + \frac{1}{3}\sigma) \geq 0{,}9505 \Rightarrow P(\bar{X} - \mu < \frac{1}{3}\sigma) \geq 0{,}9505 \Rightarrow$ estandarizamos:

$P(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < \frac{\frac{1}{3}\sigma}{\frac{\sigma}{\sqrt{n}}}) \geq 0{,}9505 \Rightarrow P(Z < \frac{\frac{1}{3}\sigma * \sqrt{n}}{\sigma}) \geq 0{,}9505 \Rightarrow P(Z < \frac{1}{3} * \sqrt{n}) \geq 0{,}9505 \Rightarrow$ busco Z para 0,9505

en la tabla => Z= 1,65 => $1{,}65 > \frac{1}{3} * \sqrt{n} \Rightarrow (1{,}65 * 3)^2 = n \Rightarrow n = 24{,}5025 == 25$

El tamaño mínimo de muestra que se requiere es 25

Ejemplo situación 2: Supongamos que la cantidad de cereal que se coloca en cada caja es una variable aleatoria normalmente distribuida con media 500 gramos y desvío estándar igual a 20 gramos. Para verificar que el peso promedio de cada caja se mantiene en 500 gramos se toma una **muestra** aleatoria de **25** cajas y se pesa el contenido de cada una. El gerente de la planta ha decidido detener el proceso y encontrar la falla cada vez que el **valor promedio de la muestra** sea mayor a 510 gramos o menor a 490 gramos. Obtener la probabilidad de detener el proceso.

Solución:

X_i="peso de la caja i" donde 1≤ i ≤ 25 y $X_i \sim$ N(μ; $\sqrt{\sigma^2}$) = N(500; 20)

Entonces $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\sum_{i=1}^{n} X_i}{25}$

nos preguntamos: ¿n es mayor a 30? No, entonces no aplica el TCL. Pero ¿conocemos la varianza poblacional? SI (20 al cuadrado) , entonces

$$\bar{X} \sim N(\mu; \frac{\sigma}{\sqrt{n}}) \Rightarrow \bar{X} \sim N(500; \frac{20}{\sqrt{25}})$$

Luego $P(de\ deterner\ el\ proceso) = P(\bar{X} > 510) + P(\bar{X} < 490) = [1 - P(\bar{X} < 510)] + P(\bar{X} < 490) = [1 - P(Z < \frac{510-500}{\frac{20}{\sqrt{25}}})] + P(Z < \frac{490-500}{\frac{20}{\sqrt{25}}}) = [1 - P(Z < 2{,}5)] + P(Z < -2{,}5) = [1 - P(Z < 2{,}5)] + [1 - P(Z < 2{,}5)] \Rightarrow buscamos\ en\ la\ tabla\ para\ Z = 2{,}5 \Rightarrow 0{,}9938$

=> (1-0,9938) + (1- 0,9938) = 0,0062 + 0,0062= 0,0124

La probabilidad de detener el proceso es 1,24%

- **Situación 3:** Si X_i tiene distribución normal con media conocida y la varianza poblacional es **desconocida**, entonces se presentan nuevas situaciones que veremos a continuación

- **Situación 3.1:** Si X_i tiene distribución normal con media conocida y la varianza poblacional es **desconocida (y n<30),** X sigue una distribución t-student con n grados de libertad

$$t_{obs} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_n$$

- **Situación 3.2:** Si X_i tiene distribución normal con media conocida y la varianza poblacional es **desconocida, y además n es grande [n>30]** entonces X sigue una distribución t-student con n-1 grados de libertad

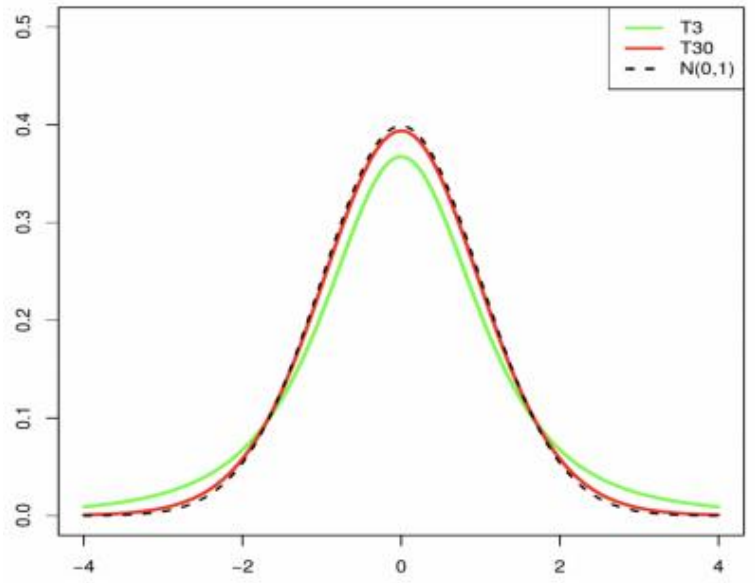
$$t_{obs} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

De ahora en más, en los enunciados hay parámetros poblacionales y estadísticos muestrales. Entonces me voy a hacer dos preguntas para identificar la distribución muestral asociada al estadístico muestral:

- ¿Conozco la varianza (o desvío) POBLACIONAL?

- ¿El tamaño de la muestra es mayor o menor que 30?

T de Student: $X \sim t_n$ E(X) =0 V(X)= $\frac{n}{n-2}$



Donde n son los grados de libertad (son todos los elementos (cantidad de valores que toma la variable aleatoria) de la muestra, es decir, el tamaño de la muestra)
Característica: es una distribución simétrica. Requiere uso de tabla

A medida que n crece, tiende a una distribución normal

Si $X \sim x_n^2$ y $X = \{Z_1, \ldots, Z_n\}$ son una muestra de variables aleatorias normales con media 0 y desvío 1 independientes, entonces:

t= $\frac{Z}{\sqrt{\frac{X}{n}}} \sim t_n$ donde X es una variable aleatoria Chi-cuadrado con n grados de libertad $\frac{nS^2}{\sigma^2}$

Es decir, $t = \frac{Z}{\sqrt{\frac{X}{n}}} = \frac{Z}{\sqrt{\frac{\frac{nS^2}{\sigma^2}}{n}}} = \frac{Z}{\sqrt{\frac{nS^2}{\sigma^2 n}}} = \frac{Z}{\sqrt{\frac{S^2}{\sigma^2}}} = \frac{Z}{\frac{S}{\sigma}} = \frac{\frac{X_i - \mu}{\sigma}}{\frac{S}{\sigma}} = \frac{X_i - \mu}{\sigma}\frac{\sigma S}{S} = \frac{X_i - \mu}{S}$ siendo una variable aleatoria normal estandarizadas como establece el TCL

Ejemplos Distribuciones Muestrales uso de tablas

Ejemplo 2: Supongamos $X \sim t_n$ con 8 grados de libertad. Hallar la probabilidad de que X sea menor 1,86.

La P(X < 1,86) = 0,95

Tablas. Cátedra: BIANCO, María José.

Distribución T de Student con "GL" grados de libertad

(Áreas acumuladas a izquierda)

GL	0,7	0,75	0,8	0,9	0,95	0,975	0,98	0,99	0,995
1	0,727	1	1,376	3,078	[illegible]	12,71	15,89	31,82	63,66
2	0,617	0,816	1,061	1,886	2,92	4,303	4,849	6,965	9,925
3	0,584	0,765	0,978	1,638	2,353	3,182	3,482	4,541	5,841
4	0,569	0,741	0,941	1,533	2,132	2,776	2,999	3,747	4,604
5	0,559	0,727	0,92	1,476	2,015	2,571	2,757	3,365	4,032
6	0,553	0,718	0,906	1,44	1,943	2,447	2,612	3,143	3,707
7	0,549	0,711	0,896	1,415	1,895	2,365	2,517	2,998	3,499
8	0,546	0,706	0,889	1,397	1,86	2,306	2,449	2,896	3,355
9	0,543	0,703	0,883	1,383	1,833	2,262	2,398	2,821	3,25
10	0,542	0,7	0,879	1,372	1,812	2,228	2,359	2,764	3,169

GL	0,7	0,75	0,8	0,9	0,95	0,975	0,98	0,99	0,995
29	0,53	0,683	0,854	1,311	1,699	2,045	2,15	2,462	2,756
30	0,53	0,683	0,854	1,31	1,697	2,042	2,147	2,457	2,75
31	0,53	0,682	0,853	1,309	1,696	2,04	2,144	2,453	2,744
34	0,529	0,682	0,852	1,307	1,691	2,032	2,136	2,441	2,728
35	0,529	0,682	0,852	1,306	1,69	2,03	2,133	2,438	2,724
39	0,529	0,681	0,851	1,304	1,685	2,023	2,125	2,426	2,708
41	0,529	0,681	0,85	1,303	1,683	2,02	2,121	2,421	2,701
44	0,528	0,68	0,85	1,301	1,68	2,015	2,116	2,414	2,692
48	0,528	0,68	0,849	1,299	1,677	2,011	2,111	2,407	2,682
49	0,528	0,68	0,849	1,299	1,677	2,01	2,11	2,405	2,68

Ejemplo situación 3.1 : Supongamos que el gobierno de un país está interesado en controlar la emisión de CO2 de los automóviles para proteger el medio ambiente. Para ello, ha intimado a las empresas a que la emisión debe ser de como máximo de 140 gramos por kilómetro al finalizar el corriente año.

Al finalizar el año, el gobierno empieza a realizar los controles correspondientes, para lo cual tomará muestras de 20 coches en cada fábrica. Supongamos que en la muestra correspondiente a la fábrica A se observó una media de 143 g/km. y un desvío estándar de 5 g/km. Si la emisión de CO2 de los coches de la fábrica A sigue una distribución Normal ¿Cuál es la probabilidad de que la empresa esté violando el requisito del gobierno?

Solución:

Nos preguntamos: ¿conocemos el σ poblacional? NO. Además, ¿n es mayor o menor que 30? Es menor. Entonces: $t_{obs} = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_n$

Entonces P($\bar{X} \geq 143$)=1-P($\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} < \frac{143-140}{\frac{5}{\sqrt{20}}}$) =1- P($t_{20}$<2,6833) => buscamos en la tabla t-Student para 20 grados de libertad y un valor t= 2,6833 (si no lo encuentro exacto tomo el más cercano. En este caso entre 2,528 y 2,845 el más cercano que es 2,528 ya que 2,6833 – 2,528= 0,1553 mientras que 2,845 – 2,6833= 0,1617)

=> 1-P($\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} < 2{,}6833$)=1-0,99=0,01 **La probabilidad de que la empresa este violando el requisito del gobierno es 1%**

Ejemplo situación 3.2 : Sea X- N(μ, σ). Si se toman muestras aleatorias de tamaño 49, cada muestra con distribución normal. Calcular la probabilidad de que la media muestral no difiera en más de 2,011 de la media poblacional.

Solución:

Sabemos que $X \sim$ N(μ;σ)

Nos preguntamos: ¿conocemos el σ poblacional? NO. Además, ¿n es mayor o menor que 30? Es mayor.

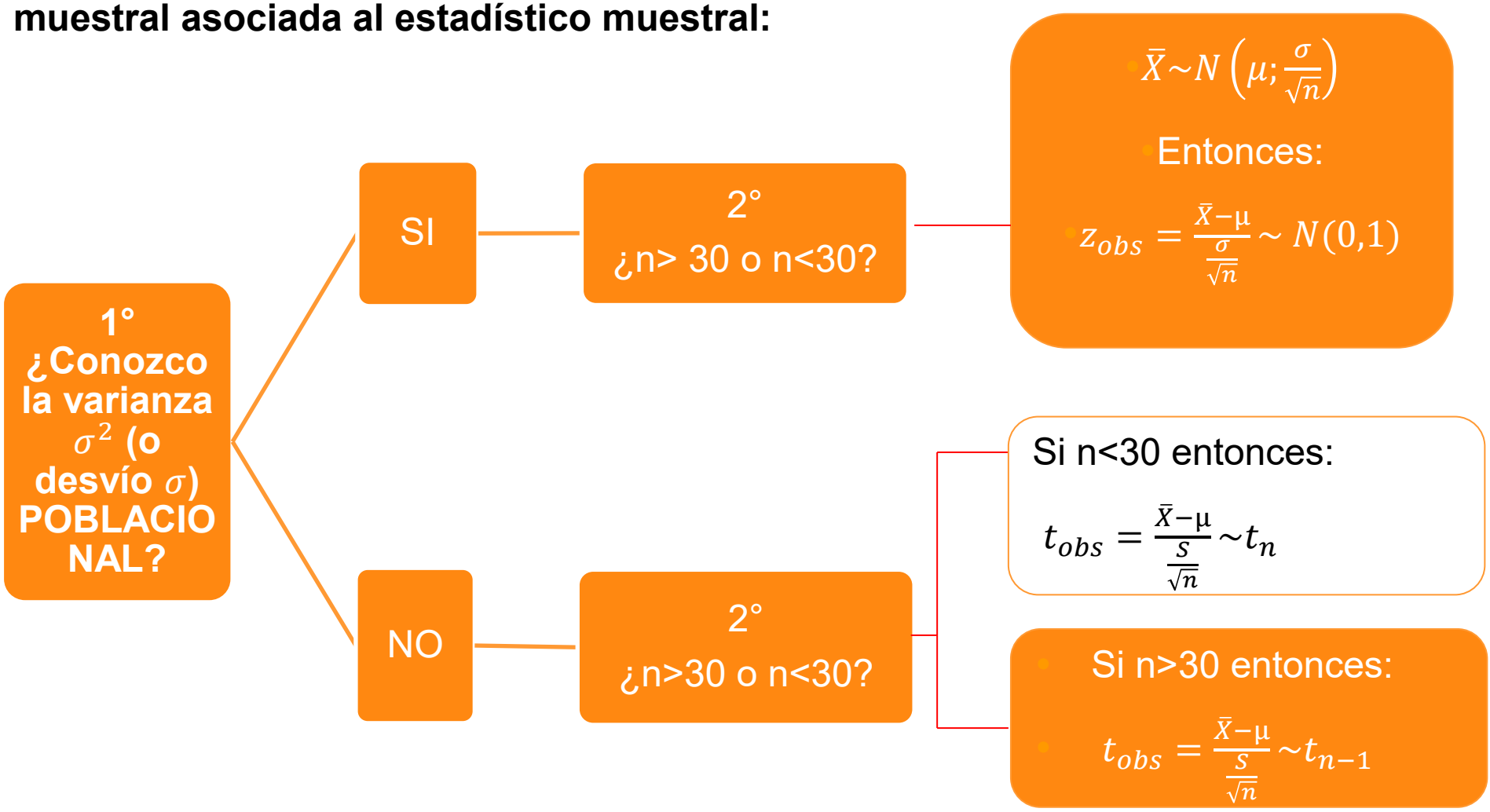
Entonces: $t_{obs} = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$

Entonces P($\bar{X} - \mu$<2,011)=P($\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$<2,011) => buscamos en la tabla t-Student para 48 grados de

libertad y un valor t=2,011 => P($\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}}$<2,011)=0,975

La probabilidad de que la media muestral no difiera es mas de 2,011 de la media poblacional es 97,5%

De ahora en más, en los enunciados de los ejercicios hay parámetros poblacionales y estadísticos muestrales. Entonces, cuando me preguntan sobre la media muestral, me voy a hacer dos preguntas para identificar la distribución muestral asociada al estadístico muestral:



Varianza Muestral $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$

- La varianza muestral sigue una distribución Chi-cuadrado con n grados de libertad

$$X \sim x_n^2$$

- **Si la muestra tiene distribución normal con media y desvío conocidos,** entonces la distribución de muestreo es Chi-cuadrado con n-1 grados de libertad

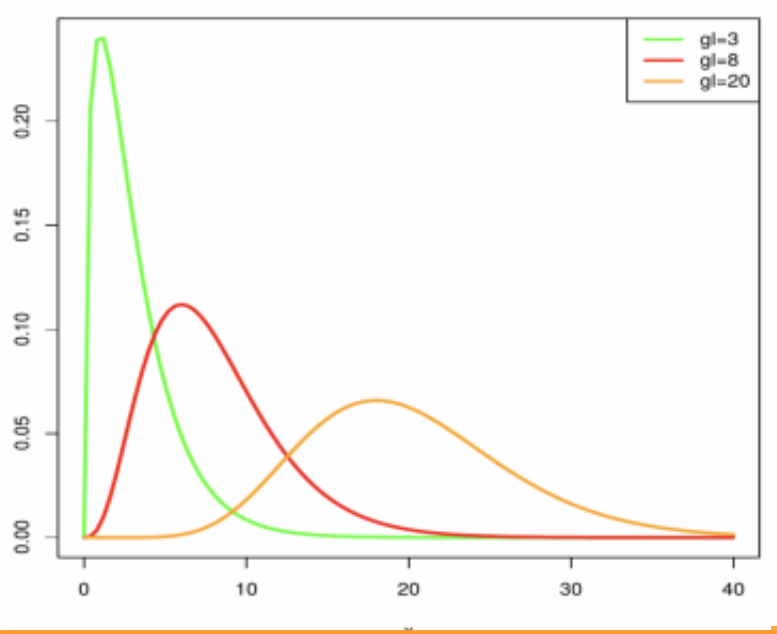
$$x_{obs}^2 = \frac{(n-1)S^2}{\sigma^2} \sim x_{n-1}^2$$

Es decir, $x_{obs}^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{(n-1)}{\sigma^2} * \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2$ siendo la definición de una variable Chi-Cuadrado dado que se define como la suma de variables normales estandarizadas al cuadrado

Chi-Cuadrado: $X \sim x_n^2$ E(X) =n V(X)= 2*n

Donde n son los grados de libertad
Característica: es una distribución asimétrica a derecha.
Requiere uso de tabla

Si X=$\{Z_1, \ldots, Z_n\}$ son una muestra de variables aleatorias normales con media 0 y desvío 1, y son independientes e idénticamente distribuidas (i.i.d.) entonces: X= $Z_1^2, \ldots, Z_n^2 \sim x_n^2$



A medida que n crece, tiende a una distribución normal

Ejemplos Distribuciones Muestrales uso de tablas

Ejemplo 1: Supongamos $X \sim x_n^2$ con 8 grados de libertad. Hallar la probabilidad de que X sea menor 2,18.

La P(X < 2,18) = 0,025

Tablas. Cátedra: BIANCO, María José.

Distribución Chi Cuadrado con "GL" grados de libertad

(Áreas acumuladas a izquierda)

GL	0,005	0,01	0,02	0,025	0,05	0,1	0,25	0,3	0,4	0,5	0,6	0,7	0,75	0,9	
1	4E-05	2E-04	6E-04	1E-03	0,004	0,016	0,102	0,148	0,275	0,455	0,708	1,074	1,323	2,706	3
2	0,01	0,02	0,04	0,051	0,103	0,211	0,575	0,713	1,022	1,386	1,833	2,408	2,773	4,605	5
3	0,072	0,115	0,185	0,216	0,352	0,584	1,213	1,424	1,869	2,366	2,946	3,665	4,108	6,251	7
4	0,207	0,297	0,429	0,484	0,711	1,064	1,923	2,195	2,753	3,357	4,045	4,878	5,385	7,779	9
5	0,412	0,554	0,752	0,831	1,145	1,61	2,675	3	3,655	4,351	5,132	6,064	6,626	9,236	1
6	0,676	0,872	1,134	1,237	1,635	2,204	3,455	3,828	4,57	5,348	6,211	7,231	7,841	10,64	1
7	0,989	1,239	1,564	1,69	2,167	2,833	4,255	4,671	5,493	6,346	7,283	8,383	9,037	12,02	1
8	1,344	1,646	2,032	2,18	2,733	3,49	5,071	5,527	6,423	7,344	8,351	9,524	10,22	13,36	1
9	1,735	2,088	2,532	2,7	3,325	4,168	5,899	6,393	7,357	8,343	9,414	10,66	11,39	14,68	1
10	2,156	2,558	3,059	3,247	3,94	4,865	6,737	7,267	8,295	9,342	10,47	11,78	12,55	15,99	1

Ejemplo : Consideremos una medición de rendimiento financiero de un activo, en donde el interés recae en la variabilidad de su cotización. Supongamos que, basados sobre la experiencia, la medición es una variable aleatoria normalmente distribuida con media 10 y desvío standard igual a 0,1 unidades. Si se toma una **muestra** aleatoria de tamaño **25**, ¿cuál es la probabilidad de que el valor de la **variabilidad muestral** sea mayor de 0,013833 unidades cuadradas?

Solución:

$X_i \sim$ N(μ;σ) => $X \sim$ N(10;0,1) y n=25 => $x_{obs} = \frac{(25-1)S^2}{\sigma^2} \sim x^2_{n-1}$

Queremos calcular P(S^2 >0,013833)

P(S^2 >0,014)= P($\frac{(n-1)S^2}{\sigma^2} > \frac{24*0,013833}{(0,1)^2}$)= P($\frac{(n-1)S^2}{\sigma^2}$ >33,2)= 1-P($\frac{(n-1)S^2}{\sigma^2}$ <33,2) => busco en la

tabla de Chi-Cuadrado para 24 grados de libertad y un valor de 33,2 => 0,9

1-0,9=0,1

La probabilidad de que el valor de la varianza muestral sea mayor de 0,013833 unidades cuadradas es 10%.

Proporción Muestral

$$\frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

- Si trabajamos con una proporción de la población que cumple determinada característica

$$\bar{p} = \frac{X}{n}$$

donde X representa la cantidad de casos favorables que se obtuvo en la muestra

- Entonces X $\sim Bi(n,p)$ con E(X)=n*p y D(X)=$\sqrt{p(1-p)}$

- Como $\bar{p}$ es el promedio de las variables Bernoulli, aplicando el Teorema Central del Límite, para un n suficientemente grande, entonces la variable aleatoria es:

$Z_{obs} = \frac{\sum_{i=1}^{n} x_i - \mu}{\sigma\sqrt{n}} = \frac{X - np}{\sqrt{p(1-p)}\sqrt{n}}$ dividiendo el numerador y denominador por n, entonces:

$$= \frac{\frac{X}{n} - \frac{np}{n}}{\frac{\sqrt{p(1-p)}\sqrt{n}}{n}} = \frac{\bar{p} - p}{\frac{\sqrt{p(1-p)}\sqrt{n}}{n}} = \frac{\bar{p} - p}{\sqrt{p(1-p)}n^{1/2}n^{-1}} = \frac{\bar{p} - p}{\sqrt{p(1-p)}n^{-1/2}} = \frac{\bar{p} - p}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}} =$$

$$Z_{obs} = \frac{\bar{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

Ejemplo : Supongamos que, en una empresa, la proporción de productos fallados en la producción diaria es de 0,10. ¿Cuál es la probabilidad de que, si se extraen 100 artículos, la proporción muestral de artículos fallados sea menor al 6%?

Solución:

Tenemos que $\bar{p}$ =0,06 y n=100. Como n es lo suficientemente grande entonces por Teorema central del Límite

$$Z_{obs} = \frac{\bar{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

Queremos calcular P($\bar{p}$<0,06)=$P\left(\frac{\bar{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < \frac{0,06-0,1}{\sqrt{\frac{0,1(1-0,1)}{100}}}\right) = P\left(\frac{\bar{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < \frac{-0,04}{0,03}\right) = P(Z< -1,333)$

=>=1- $P(Z<1,333)$ => busco en la tabla Normal Estándar => = 1- 0,9082= 0,0918

La probabilidad de que la proporción de artículos fallados sea menor al 6% es 9,18%

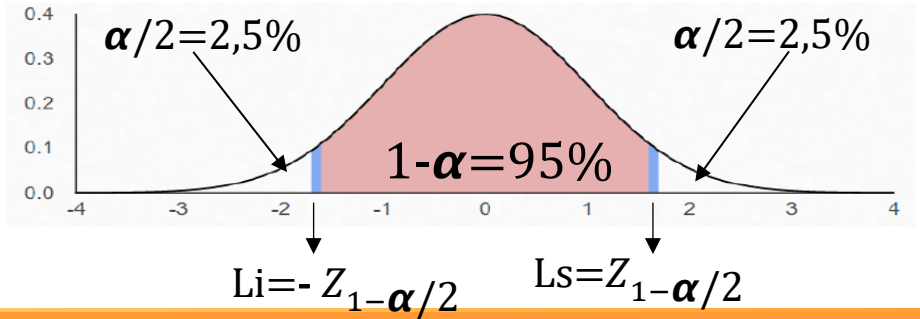
Intervalo de confianza

Es un rango de valores posibles dentro del cual se encuentra el parámetro poblacional que se quiere estimar bajo un cierto nivel de confianza.

Los estadísticos y estimadores son variables aleatorias ya que su valor varía de una muestra a otra. En consecuencia, por más que usemos el mismo estimador, la estimación que realicemos variará de una muestra a otra. Pero si conocemos la distribución de muestreo del estimador (lo que vimos la clase pasada, entonces utilizaremos un estadístico), podremos hallar un intervalo estimado que con cierta probabilidad contendrá al parámetro poblacional.

Elementos que contiene un intervalo de confianza

- **Límites**: Son los valores límites entre los cuales se encontrará el parámetro de la población bajo un cierto nivel de confianza. Señalamos con **Li** al límite inferior y **Ls** al límite superior .
- **Nivel de confianza**: es el nivel de probabilidad con el cual quiero realizar la estimación 1-α Si establezco un 95% (1-α) de confianza para estimar entonces α=5% es el nivel de significatividad. Por lo tanto 1- α es el percentil hasta el cual estoy acumulando la probabilidad de la estimación.



Por lo tanto, la estructura de un intervalo viene dada por:

$$P\Big[\underbrace{\text{Estadístico Muestral} - \text{Margen de Error}}_{\text{Límite inferior}} \leq \text{Parámetro Poblacional} \leq \underbrace{\text{Estadístico Muestral} + \text{Margen de Error}}_{\text{Límite superior}}\Big] = \underbrace{1-\alpha}_{\text{Percentil}}$$

Para nuestra estimación queremos:

- Realizarla con el mayor nivel de confianza posible
- Obtener mucha precisión (es decir, cometer el menor error posible y por tanto obtener una longitud chica del intervalo)
- Optimizar el costo (es decir, poder obtener lo anterior con el menor tamaño de muestra posible)

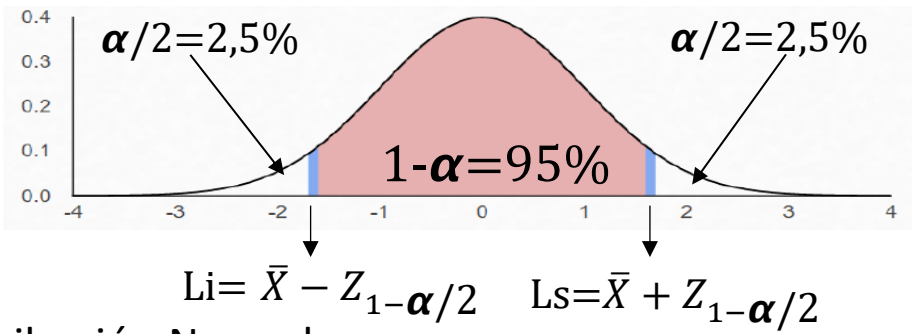
Pero como se trata de inferencia, entonces observaremos que:

- El margen de error disminuye a medida que aumenta n (tamaño de muestra)
- Si queremos mejorar el nivel de confianza (o disminuir el margen de error), debemos compensar aumentando el tamaño de la muestra

1. Intervalo de confianza para la media si la varianza poblacional es conocida

Sean $\{X_1, X_2 \ldots, X_n\}$ una muestra aleatoria de tamaños n proveniente de una población Normal con media μ_x y varianza σ_X^2. Entonces:

$$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$



Construimos el intervalo: por simetría de la distribución Normal

$$P(-Z_{1-\frac{\alpha}{2}} \leq \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq Z_{1-\frac{\alpha}{2}}) = 1-\alpha \quad \rightarrow \quad \boxed{P(\underbrace{\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}}_{\text{Li}} \leq \mu \leq \underbrace{\bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}}_{\text{Ls}}) = \underbrace{1-\alpha}_{\text{Percentil}}}$$

1.1 Tamaño de la muestra

Error: $\in = |\bar{X} - \mu| \quad \rightarrow \quad n = \left(\frac{Z_{1-\frac{\alpha}{2}} * \sigma}{\in}\right)^2$ donde $\in$ es el error de la estimación

Longitud: $L = 2 * Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \quad \rightarrow \quad n = \left(\frac{2 * Z_{1-\frac{\alpha}{2}} * \sigma}{L}\right)^2$ donde L es la longitud del IC

Respecto del Tamaño de la muestra

Error: es la diferencia entre el estadístico muestral y el parámetro poblacional, es decir

$$\in = |\bar{X} - \mu|$$

Partiendo del intervalo de confianza: $\bar{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = \mu$

$\Rightarrow \bar{X} - \mu = Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \Rightarrow \in = Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \Rightarrow \sqrt{n} = \frac{Z_{1-\frac{\alpha}{2}} * \sigma}{\in} \Rightarrow \boldsymbol{n = \left(\frac{Z_{1-\frac{\alpha}{2}} * \sigma}{\in}\right)^2}$ **Tamaño de la muestra**

Longitud: es la diferencia entre el límite superior y el límite inferior, es decir

$$L = \text{Límite superior} - \text{Límite inferior}$$

Partiendo del intervalo de confianza: $L = \bar{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} - \left(\bar{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$

$\Rightarrow L = Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \Rightarrow L = 2 * Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \Rightarrow \sqrt{n} = \frac{2 * Z_{1-\frac{\alpha}{2}} * \sigma}{L}$

$\Rightarrow \quad \boldsymbol{n = \left(\frac{2 * Z_{1-\frac{\alpha}{2}} * \sigma}{L}\right)^2}$ **Tamaño de la muestra**

Ejemplo 1: Una empresa proveedora de gas en una ciudad decide hacer un estudio para determinar la cantidad de gas gastada para calefacción casera durante un año. Con tal motivo se selecciona una muestra de 64 hogares de la ciudad. Se supone que el gasto por consumo domiciliario de gas sigue una distribución normal. La media muestral del gasto en gas para calefacción resultó de $836. Se sabe por experiencia que el desvío estándar de la población es $178. Hallar un intervalo de con fianza del 95% para el gasto promedio anual de gas en las viviendas de esta ciudad.

Nos preguntamos: el desvío de la población, ¿es conocido? SI. Entonces:

$$\boldsymbol{P(\bar{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha}$$

$$P\left(836 - Z_{1-\frac{\alpha}{2}}\frac{178}{\sqrt{64}} \leq \mu \leq 836 + Z_{1-\frac{\alpha}{2}}\frac{178}{\sqrt{64}}\right) = 95\%$$

$$Si\ 1 - \boldsymbol{\alpha} = 95\% \Rightarrow \boldsymbol{\alpha} = 5\% \Rightarrow\ Z_{1-\frac{0{,}05}{2}} = Z_{1-0{,}025} = Z_{0{,}975}$$

$$\Rightarrow busco\ en\ al\ tabla\ el\ valor\ de\ Z\ para\ 0{,}975 \Rightarrow Z = 1{,}96$$

$$P\left(836\ - 1{,}96 * \frac{178}{\sqrt{64}} \leq \mu \leq 836 + 1{,}96 * \frac{178}{\sqrt{64}}\right) = 95\% \Rightarrow\ P(792{,}39 \leq \mu \leq 879{,}61) = 95\%$$

El gasto promedio anual de gas de los hogares en la ciudad se encuentra entre 792, 39 y 879, 61 con un nivel de confianzadel95%

Ejemplo 1.1: Un investigador desea determinar mediante un intervalo de confianza del 95% y una longitud de 3000 pesos, el salario medio mensual del sector gastronómico de determinada ciudad. Por estudios anteriores se sabe que los salarios siguen una distribución normal con desvío estándar de 10000 pesos. ¿Qué tamaño de muestra debería utilizar?.

$n = \left(\frac{2 * Z_{1-\frac{\alpha}{2}} * \sigma}{L}\right)^2$ donde L es la longitud del IC

$Si\ 1 - \boldsymbol{\alpha} = 95\% \Rightarrow \boldsymbol{\alpha} = 5\% \Rightarrow\ Z_{1-\frac{0,05}{2}} = Z_{1-0,025} = Z_{0,975}$

$\Rightarrow busco\ en\ al\ tabla\ el\ valor\ de\ Z\ para\ 0,975 \Rightarrow Z = 1,96$

$$n = \left(\frac{2 * 1,96 * 10000}{3000}\right)^2 = 170,74 == 171$$

El tamaño de muestra que se deberá utilizar es de por lo menos 171 personas

2. Intervalo de confianza para la media si la varianza poblacional es desconocida

Sean $\{X_1, X_2 \ldots, X_n\}$ una muestra aleatoria de tamaños n proveniente de una población Normal con media μ_x y **varianza desconocida**. Entonces:

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$$

Construimos el intervalo: por simetría de la distribución t-student

$$P\left(t_{n-1;\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \leq t_{n-1;1-\frac{\alpha}{2}}\right) = 1 - \boldsymbol{\alpha} \rightarrow \boxed{P\left(\underbrace{\bar{X} - t_{n-1;1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}}_{\text{Li}} \leq \mu \leq \underbrace{\bar{X} + t_{n-1;1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}}_{\text{Ls}}\right) = \underbrace{1 - \boldsymbol{\alpha}}_{\text{Percentil}}}$$

Por lo tanto, si me solicitan estimar un intervalo de confianza para la media poblacional, lo que debo preguntarme es si conozco o no conozco la varianza (o desvío) poblacional para determinar que intervalo utilizar

2.1 Tamaño de la muestra

Error $\in = |\bar{X} - \mu| \longrightarrow n = \left(\frac{t_{n-1;1-\frac{\alpha}{2}} * S}{\in}\right)^2$ donde $\in$ es el error de la estimación

$L = 2 * t_{n-1;1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \longrightarrow n = \left(\frac{2 * t_{n-1;1-\frac{\alpha}{2}} * S}{L}\right)^2$ donde L es la longitud del IC

Ejemplo 2: Con el fin de determinar la edad media de alumnos que cursan una carrera determinada en un Universidad, se tomó una muestra de 20 alumnos, obteniéndose una media de 23 años y un desvío de 3,5 años. Si la edad en un estudio se distribuye normalmente, hallar un intervalo de confianza para la verdadera media con un nivel de confianza del 95%.

Nos preguntamos: el desvío de la población, ¿es conocido? NO. Entonces:

$$P(\bar{X} - t_{n-1;1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1;1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}) = 1 - \boldsymbol{\alpha}$$

$$P\left(23 - t_{n-1;1-\frac{\alpha}{2}} \frac{3{,}5}{\sqrt{20}} \leq \mu \leq 23 + t_{n-1;1-\frac{\alpha}{2}} \frac{3{,}5}{\sqrt{20}}\right) = 95\%$$

$$Si\ 1 - \boldsymbol{\alpha} = 95\% \Rightarrow \boldsymbol{\alpha} = 5\% \Rightarrow t_{20-1;1-\frac{0{,}05}{2}} \Rightarrow t_{20-1;1-0{,}025} \Rightarrow t_{19;0{,}975} \Rightarrow$$

$$\Rightarrow busco\ en\ al\ tabla\ el\ valor\ de\ t\ para\ 0{,}975\ y\ 19\ GL \Rightarrow t = 2{,}093$$

$$P\left(23 - 2{,}093 * \frac{3{,}5}{\sqrt{20}} \leq \mu \leq 23 + 2{,}093 * \frac{3{,}5}{\sqrt{20}}\right) = 95\% \Rightarrow P(21{,}36 \leq \mu \leq 24{,}64) = 95\%$$

La edad promedio de los alumnos en una universidad se encuentra entre 21,36 y 24,64 con un nivel de confianza del 95%

3. Intervalo de confianza para la varianza en poblaciones normales

Sean $\{X_1, X_2 \ldots, X_n\}$ una muestra aleatoria de tamaños n proveniente de una población Normal con media **conocida**. Entonces:

$$\frac{(n-1)S^2}{\sigma^2} \sim x^2_{n-1}$$

Construimos el intervalo:

$$P(x^2_{n-1;\frac{\alpha}{2}} \leq \frac{(n-1)S^2}{\sigma^2} \leq x^2_{n-1;1-\frac{\alpha}{2}}) = 1 - \boldsymbol{\alpha}$$

$$P(\underbrace{\frac{(n-1)S^2}{x^2_{n-1;1-\frac{\alpha}{2}}}}_{\text{LI}} \leq \sigma^2 \leq \underbrace{\frac{(n-1)S^2}{x^2_{n-1;\frac{\alpha}{2}}}}_{\text{LS}}) = 1 - \boldsymbol{\alpha}$$

Percentil

Ejemplo 3: Una agencia de alquiler de autos quiere estudiar la distribución de los kilómetros diarios que realiza su flota. Para eso, a lo largo de varios días, se anotan los recorridos de cien vehículos de su flota y se obtiene que la media muestral es de 165km/día y que el desvío estándar muestral es 6km/día. Suponiendo que la distribución de los km recorridos es normal con media conocida, hallar un intervalo de con confianza del 90% para la varianza de la distribución.

Nos preguntamos: ¿conocemos la media de la población? SI. Entonces:

$$P\left(\frac{(n-1)S^2}{x^2_{n-1;1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S^2}{x^2_{n-1;\frac{\alpha}{2}}}\right) = 1-\boldsymbol{\alpha} \quad P\left(\frac{(100-1)*6^2}{x^2_{100-1;1-\frac{0,1}{2}}} \leq \sigma^2 \leq \frac{(100-1)*6^2}{x^2_{n-1;\frac{0,1}{2}}}\right) = 90\%$$

$$Si\ 1-\boldsymbol{\alpha} = 90\% \Rightarrow \boldsymbol{\alpha} = 10\% \Rightarrow x^2_{100-1;1-\frac{0,1}{2}} \Rightarrow x_{99;0,95} \Rightarrow$$

$$\Rightarrow busco\ en\ al\ tabla\ el\ valor\ de\ x\ para\ 0{,}95\ y\ 99\ GL \Rightarrow x = 123{,}2$$

$$Si\ 1-\boldsymbol{\alpha} = 90\% \Rightarrow \boldsymbol{\alpha} = 10\% \Rightarrow x^2_{100-1;\frac{0,1}{2}} \Rightarrow x_{99;0,05} \Rightarrow$$

$$\Rightarrow busco\ en\ la\ tabla\ el\ valor\ de\ x\ para\ 0{,}05\ y\ 99\ GL \Rightarrow x = 77{,}5$$

$$P\left(\frac{(100-1)*6^2}{123{,}2} \leq \sigma^2 \leq \frac{(100-1)6^2}{77{,}5}\right) = 90\% \qquad P(28{,}93 \leq \sigma^2 \leq 45{,}99) = 90\%$$

La varianza de la distribución de km recorridos por día se encuentra entre 28,93 y 45,99 con un nivel de confianza del 90%

4. Intervalo de confianza para proporciones en poblaciones binomiales

Sean $\{X_1, X_2 \ldots, X_n\}$ una muestra aleatoria de tamaños n proveniente de una población Normal con media **conocida**. Entonces:

$$\frac{\bar{p} - p}{\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}} \sim N(0;1)$$

Construimos el intervalo:

Por simetría de la distribución normal $P(Z_{\frac{\alpha}{2}} \leq \frac{\bar{p} - p}{\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}} \leq Z_{1-\frac{\alpha}{2}}) = 1 - \boldsymbol{\alpha}$

tomando un $\tilde{p}$ ya que no conocemos a p

$$P(\underbrace{\tilde{p} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}}_{\text{Li}} \leq p \leq \underbrace{\tilde{p} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}}_{\text{Ls}}) = \underbrace{1 - \boldsymbol{\alpha}}_{\text{Percentil}}$$

Cuando hablamos de proporciones nos referimos a la proporción de una población que cumple con determinadas características. Entonces, p será la cantidad de casos favorables que se obtuvo en la muestra. Además, los límites de los intervalos nunca podrán ser negativos ni mayor que 1

Tamaño de la muestra

Tamaño de la muestra : $n = \left(\frac{Z_{1-\frac{\alpha}{2}}}{\in}\right)^2 \tilde{p}(1-\tilde{p})$

Ejemplo 4: En una investigación de mercados que involucró encuestar a 100 personas se detectó que el 40% de los consumidores prefieren los productos de la marca A frente a los de otras marcas sustitutas.

Construir un intervalo de confianza del 95% para la proporción poblacional.

$$P\left(\tilde{p} - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} \leq p \leq \tilde{p} + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}\right) = 1 - \boldsymbol{\alpha}$$

$$P\left(0{,}4 - Z_{1-\frac{0{,}05}{2}}\sqrt{\frac{0{,}4(1-0{,}4)}{100}} \leq p \leq 0{,}4 + Z_{1-\frac{0{,}05}{2}}\sqrt{\frac{0{,}4(1-0{,}4)}{100}}\right) = 95\%$$

$$Si\ 1 - \boldsymbol{\alpha} = 95\% \Rightarrow \boldsymbol{\alpha} = 5\% \Rightarrow Z_{1-\frac{0{,}05}{2}} \Rightarrow Z_{0{,}975}$$

$$\Rightarrow busco\ en\ al\ tabla\ el\ valor\ de\ Z\ para\ 0{,}975\ \Rightarrow Z = 1.96$$

$$P\left(0{,}4\ - 1{,}96 * \sqrt{\frac{0{,}4(1-0{,}4)}{100}} \leq p \leq 0{,}4 + 1{,}96 * \sqrt{\frac{0{,}4(1-0{,}4)}{100}}\right) = 95\% \Rightarrow$$

$$P(0{,}304 \leq p \leq 0{,}496) = 95\%$$

***La proporción poblacional se encuentra entre* $0{,}304$ *y* $0{,}496$ *con un nivel de confianza del* 95%**

Ejemplo 4.1: Continuando con el ejemplo anterior, supongamos que antes de recopilar los datos se especificó que la estimación a través de un intervalo del 95% de confianza debería tener una precisión del 5% ¿Cuál tendría que ser el tamaño de la muestra que debería tomarse?

$Sabiamos\ que\ \bar{p} = 0{,}4\ entonces$:

Tamaño de la muestra : $n = \left(\frac{Z_{1-\frac{\alpha}{2}}}{\in}\right)^2 \tilde{p}(1-\tilde{p})$

$n = \left(\frac{1{,}96}{0{,}05}\right)^2 0{,}4(1-0{,}4)$= 368,7936==369

Habría que tomar una muestra de 369 personas

Intervalo de confianza para comparar dos poblaciones

Muchas veces se desea comparar poblaciones para saber si sus medias o varianzas o proporciones son iguales o no.

Por ejemplo, el gobierno Nacional podría estar interesado en comparar el ingreso medio de los habitantes de dos provincias para saber si son iguales. O una empresa podría estar interesada en comprar la proporción de artículos defectuosos que generan dos máquinas para utilizar aquélla que tenga mejor funcionamiento.

Una herramienta útil en estos casos son los intervalos de confianza que nos permiten realizar estas comparaciones.

5. Intervalo de confianza para diferencia de medias de dos poblaciones con varianzas poblaciones conocidas

Sean $\{X_1, X_2 \ldots, X_n\}$ y $\{Y_1, Y_2 \ldots, Y_n\}$ dos muestras aleatorias de tamaños n y m, respectivamente, provenientes de dos poblaciones Normales independientes con medias μ_x y μ_y, y varianzas σ_X^2 y σ_Y^2. Entonces:

$$\frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0,1)$$

Construimos el intervalo correspondiente

$$P\left(-Z_{1-\frac{\alpha}{2}} \leq \frac{(\overline{X} - \overline{Y}) - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \leq Z_{1-\frac{\alpha}{2}}\right) = 1 - \boldsymbol{\alpha}$$

$$P\left((\overline{X} - \overline{Y}) - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \leq (\mu_x - \mu_y) \leq (\overline{X} - \overline{Y}) + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}\right) = 1 - \boldsymbol{\alpha}$$

Li — Ls — Percentil

Ejemplo 5: Supongamos que se extrae una muestra con con 20 de varones y 30 de mujeres. Las medias es las muestra son 1,8 mts y 1,7 mts respectivamente. Si la distribución de estaturas de mujeres es Normal con desvío estándar de 15 cm (o 0,15 mts) e independiente de las estaturas de varones con desvío estándar 10cm (o 0,10 m), ¿cuál es el IC al 99% para la diferencia de las medias?

$$P((\overline{X} - \overline{Y}) - Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}} \leq (\mu_x - \mu_y) \leq (\overline{X} - \overline{Y}) + Z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}) = 1 - \boldsymbol{\alpha}$$

Definimos:

X= estatura de los varones Y = estatura de las mujeres

Nota: el valor más chico siempre es Y

$$Si\ 1 - \boldsymbol{\alpha} = 99\% \Rightarrow \boldsymbol{\alpha} = 1\% \Rightarrow Z_{1-\frac{0,01}{2}} \Rightarrow Z_{0,995}$$

$$\Rightarrow busco\ en\ al\ tabla\ el\ valor\ de\ Z\ para\ 0,995\ \Rightarrow Z = 2,576$$

$$P\left((1,8 - 1,7) - 2,576\sqrt{\frac{0,15^2}{20} + \frac{0,10^2}{30}} \leq (\mu_x - \mu_y) \leq (1,8 - 1,7) + 2,576\sqrt{\frac{0,15^2}{20} + \frac{0,10^2}{30}}\right) = 99\%$$

$$P(0,0016 \leq (\mu_x - \mu_y) \leq 0,1984) = 99\%$$

Por lo tanto, en base a esta muestra, podemos decir con un 99% de confianza que la estatura media de los varones es superior a la estatura media de las mujeres.

6. Intervalo de confianza para diferencia de medias de dos poblaciones con varianzas poblaciones desconocidas e iguales

Sean $\{X_1, X_2 \dots, X_n\}$ y $\{Y_1, Y_2 \dots, Y_n\}$ dos muestras aleatorias de tamaños n y m, respectivamente, provenientes de dos poblaciones Normales independientes con medias μ_x y μ_y, y varianzas **desconocidas pero iguales**. Entonces:

$$\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2} \text{ con } S_p = \sqrt{S_p^2} \quad \text{donde} \quad S_p^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

Construimos el intervalo correspondiente

$$P\left(-t_{n+m-2;1-\frac{\alpha}{2}} \leq \frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \leq t_{n+m-2;1-\frac{\alpha}{2}}\right) = 1-\boldsymbol{\alpha}$$

$$\downarrow$$

$$P\left((\bar{X}-\bar{Y}) - t_{n+m-2;1-\frac{\alpha}{2}} * S_p\sqrt{\frac{1}{n}+\frac{1}{m}} \leq (\mu_x-\mu_y) \leq (\bar{X}-\bar{Y}) + t_{n+m-2;1-\frac{\alpha}{2}} * S_p\sqrt{\frac{1}{n}+\frac{1}{m}}\right) = 1-\boldsymbol{\alpha}$$

Li Ls Percentil

Ejemplo 6: Supongamos que se extrae una muestra con 20 de varones y 30 de mujeres. Las medias de la estatura en las muestra son 1,8 mts y 1,7 mts respectivamente. Los desvíos muestrales son 0,064 y 0,09 respectivamente. Si la distribución de estaturas de personas es Normal, ¿cuál es el IC al 95% para la diferencia de las medias?

$$P((\overline{X} - \overline{Y}) - t_{n+m-2;1-\frac{\alpha}{2}} * S_p \sqrt{\frac{1}{n}+\frac{1}{m}} \leq (\mu_x - \mu_y) \leq (\overline{X} - \overline{Y}) + t_{n+m-2;1-\frac{\alpha}{2}} * S_p \sqrt{\frac{1}{n}+\frac{1}{m}} = 1 - \boldsymbol{\alpha}$$

Definimos:

X= estatura de los varones Y = estatura de las mujeres

Nota: el valor más chico siempre es Y

Desconocemos los desvío poblaciones entonces se asumen iguales.

$$Si\ 1 - \boldsymbol{\alpha} = 95\% \Rightarrow \boldsymbol{\alpha} = 5\% \Rightarrow t_{20+30-2;0,975}$$

$$\Rightarrow busco\ en\ al\ tabla\ el\ valor\ de\ t\ para\ 0{,}975\ y\ 48\ gl \Rightarrow t = 2{,}011$$

$$S_p^2 = \frac{(20-1)0{,}064^2+(30-1)0{,}09^2}{20+30-2} = 0{,}006515 \Rightarrow S_p = \sqrt{0{,}006515} = 0{,}0807$$

$$P\left((1{,}8-1{,}7) - 2{,}011 * 0{,}0807 \sqrt{\frac{1}{20}+\frac{1}{30}} \leq (\mu_x - \mu_y) \leq (1{,}8-1{,}7) + 2{,}011 * 0{,}0807 \sqrt{\frac{1}{20}+\frac{1}{30}}\right) = 95\%$$

$$P(0{,}05315 \leq (\mu_x - \mu_y) \leq 0{,}14685) = 95\%$$

Por lo tanto, en base a esta muestra, podemos decir con un 95% de confianza que la estatura media de los varones es superior a la estatura media de las mujeres.

7. Intervalo de confianza para diferencia de proporciones de dos poblaciones

Sean $\{A_1, A_2 \dots, A_n\}$ y $\{B_1, B_2 \dots, B_n\}$ dos muestras aleatorias de tamaños n y m, respectivamente, provenientes de dos poblaciones Bernoulli con parámetros p_A y p_B. Sean X e Y la cantidad de éxitos de cada una de ellas. Entonces:

$$\frac{(\tilde{p}_A - \tilde{p}_B) - (p_A - p_B)}{\sqrt{\frac{\tilde{p}_A(1-\tilde{p}_A)}{n} + \frac{\tilde{p}_B(1-\tilde{p}_B)}{m}}} \sim N(0,1)$$

Construimos el intervalo correspondiente

$$P\left(-Z_{1-\frac{\alpha}{2}} \leq \frac{(\tilde{p}_A - \tilde{p}_B) - (p_A - p_B)}{\sqrt{\frac{\tilde{p}_A(1-\tilde{p}_A)}{n} + \frac{\tilde{p}_B(1-\tilde{p}_B)}{m}}} \leq Z_{1-\frac{\alpha}{2}}\right) = 1 - \boldsymbol{\alpha}$$

$$\downarrow$$

$$P\left(\underbrace{(\tilde{p}_A - \tilde{p}_B) - Z_{1-\frac{\alpha}{2}} * S_p}_{\text{Li}} \leq (p_A - p_B) \leq \underbrace{(\tilde{p}_A - \tilde{p}_B) + Z_{1-\frac{\alpha}{2}} * S_p}_{\text{Ls}} = \underbrace{1 - \boldsymbol{\alpha}}_{\text{Percentil}}\right)$$

$$S_p^2 = \frac{\tilde{p}_A(1-\tilde{p}_A)}{n} + \frac{\tilde{p}_B(1-\tilde{p}_B)}{m}$$

Ejemplo 7: Suponga que, en una encuesta tomada a 500 personas en la ciudad A, 350 respondieron que apoyan al candidato A. Supongamos que en otra ciudad B se encuesta a 300 personas, y resulta que 150 están a favor del candidato A. ¿Cuál es el IC al 95% para la diferencia entre las proporciones de ambas ciudades?

$$P((\tilde{p}_A - \tilde{p}_B) - Z_{1-\frac{\alpha}{2}} * S_p \leq (p_A - p_B) \leq (\tilde{p}_A - \tilde{p}_B) + Z_{1-\frac{\alpha}{2}} * S_p) = 1 - \boldsymbol{\alpha}$$

Calculamos:

$\tilde{p}_A$ = 350/500= 0,7 $\tilde{p}_B$ = 150/300= 0,5

$$Si\ 1 - \boldsymbol{\alpha} = 95\% \Rightarrow \boldsymbol{\alpha} = 5\% \Rightarrow Z_{1-\frac{0,05}{2}} \Rightarrow Z_{0,975}$$

$$\Rightarrow busco\ en\ al\ tabla\ el\ valor\ de\ Z\ para\ 0,975\ \Rightarrow Z = 1,96$$

$$S_p^2 = \frac{0,7(1-0,7)}{500} + \frac{0,5(1-0,5)}{300} = 0,00125 \Rightarrow S_p = 0,035$$

$$P((0,7 - 0,5) - 1,96 * 0,035 \leq (p_A - p_B) \leq (0,7 - 0,5) - 1,96 * 0,035 = 1 - \boldsymbol{\alpha}$$

$$P(0,1314 \leq (p_A - p_B) \leq 0,2686) = 95\%$$

Por lo tanto, en base a las muestras, con un 95% de confianza podemos afirmar que la diferencia entre la proporción de personas favorables al candidato A en las ciudades A y B, está entre el 13,14% y el 26,86%. Esto evidencia el mayor apoyo en la ciudad A, ya que el intervalo de la diferencia de proporciones no incluye al cero, por lo que, más allá de cuál sea la diferencia, podemos afirmar con un 95% de confianza que $p_A > p_B$

Pueden realizar la práctica 6