

DigitalHouse >
Coding School

DATA SCIENCE

MÓDULO 2

Inferencia Estadística

1

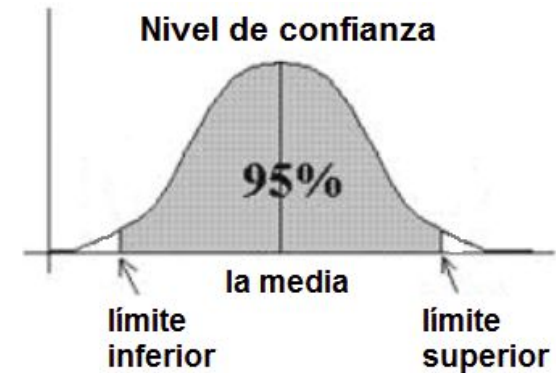
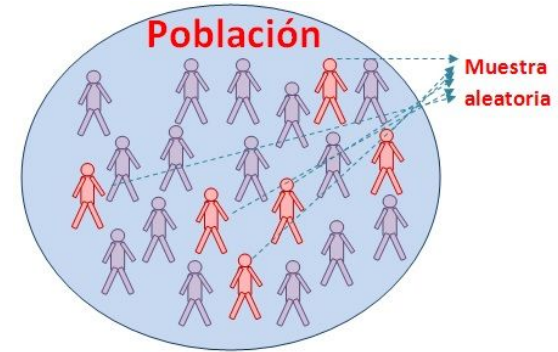
Repasar los términos población-muestra;
parámetro-estimador

2

Comprender el concepto de muestreo y de
distribución muestral

3

Calcular estimaciones puntuales y por
Intervalos de Confianza bajo Muestreo Aleatorio
Simple



POBLACIÓN-MUESTRA



POBLACIÓN

- En los problemas de diferentes disciplinas se estudia el comportamiento de varias variables definidas sobre un conjunto de objetos. **El conjunto de objetos será denominado población.**
- Una forma de simbolizar las “N” unidades de la población es:
 - $\{U_i \text{ donde } i = 1...N\} = (U_1, U_2, \dots, U_i, \dots, U_N)$
- La población se define en relación al problema de investigación a abordar.
 - Ejemplo: los participantes del curso de Data Science, se busca analizar su altura promedio.

TIPOS DE POBLACIÓN

Población finita: el número de individuos que componen a la población es finito. Por ejemplo los habitantes de una ciudad, las unidades producidas por una planta industrial por día, etc.

Población infinita: el número de individuos que componen a la población es infinito. Por ejemplo, el conjunto de números positivos, el resultado de un experimento que, al menos teóricamente, se puede repetir tantas veces como se quiera.

También se consideran como poblaciones infinitas las que tienen un número extremadamente grande de elementos: granos de arena en una playa, átomos en el universo, etc.

TIPOS DE POBLACIÓN

Población real: es todo el grupo de elementos concretos, como las personas de Argentina que se dedican a Data Science.

Población hipotética: es el conjunto de situaciones posibles imaginables en que puede presentarse un suceso, como por ejemplo las formas de reaccionar de una persona ante una catástrofe.

Otras: **estable vs. inestable, binomial vs. polinomial**, etc.

MUESTRA

- **Muestra:** es seleccionada de la población (definida en relación al problema de investigación). Es el subconjunto de unidades seleccionadas de la población definida.
- En ésta recae la realización de las observaciones, mediciones, etc. Las “n” unidades o Muestra seleccionada de una Población de “N” se simbolizan:
 - $\{ u_i \text{ donde } i = 1...n \} = (u_1, u_2, \dots, u_i, \dots, u_n)$
- Si tomamos como muestra a toda la población, entonces se dice **censo**. ¿Por qué no es práctico o conveniente realizar siempre un censo?

ALGUNOS SESGOS MUESTRALES

Por conveniencia: los individuos más accesibles tienen más probabilidades de ser seleccionados.

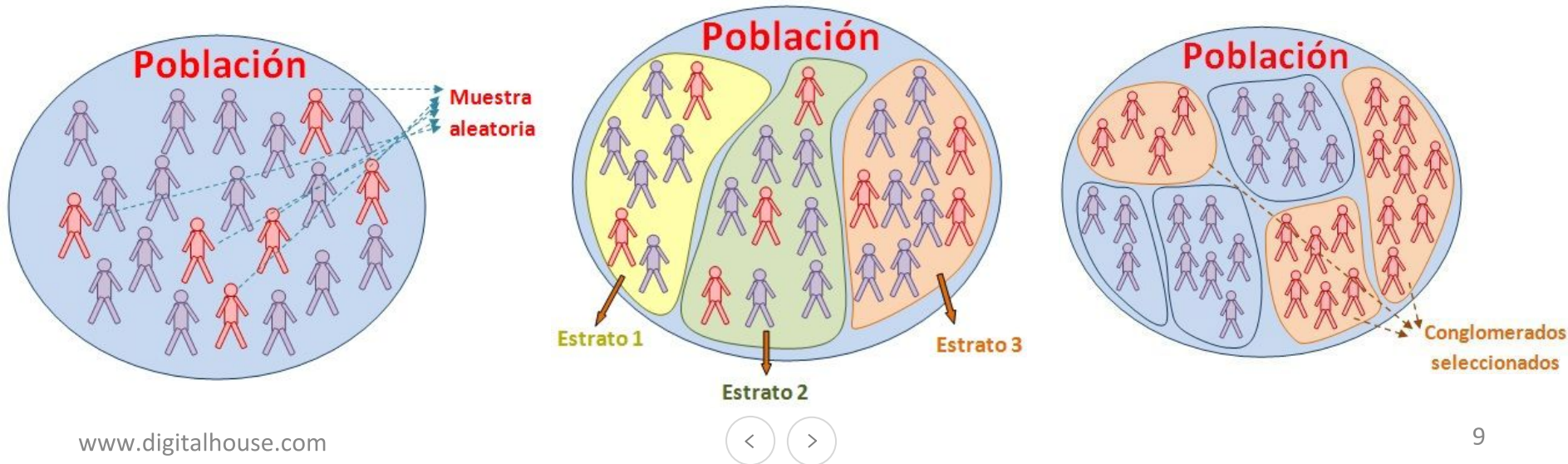
No respuesta: si no tenemos datos de una porción no aleatoria de una muestra aleatoria (por ejemplo, porque no responden la encuesta), entonces la muestra ya no será representativa de la población.

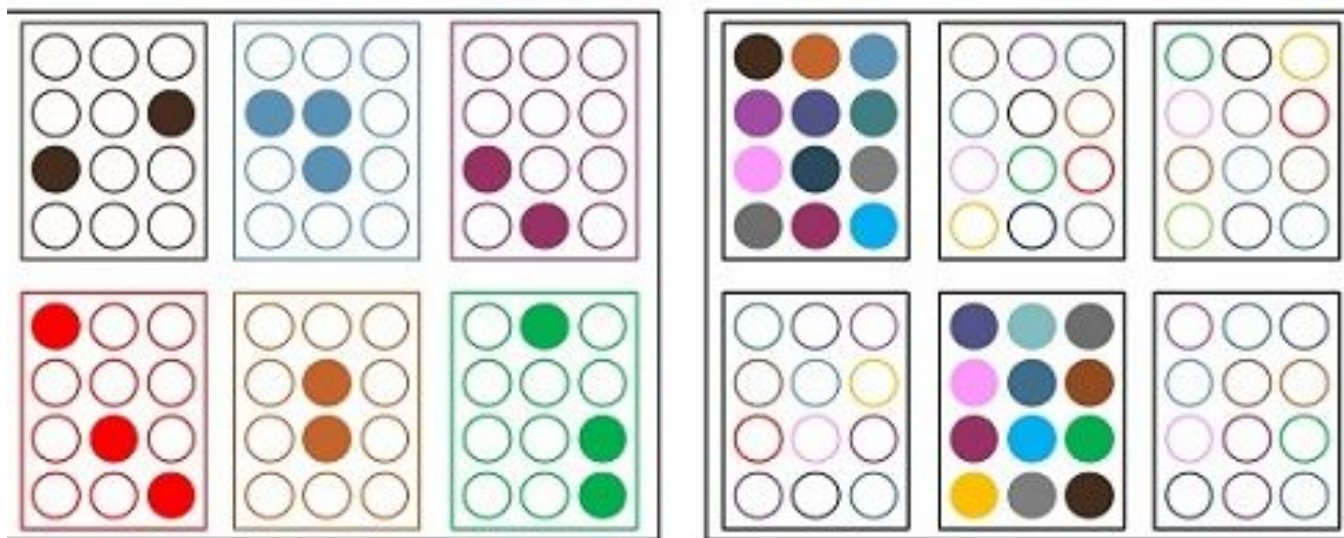
Respuesta voluntaria: la muestra está compuesta solamente por personas que se ofrecen a responder (tienden a ser la personas con opiniones más fuertes).

- **Probabilísticas:** puedo calcular la probabilidad de selección de cada una de las unidades de la muestra => Puedo calcular una medida del error.

Algunos tipos:

1. [Muestreo aleatorio simple](#)
2. [Muestreo aleatorio estratificado](#)
3. [Muestreo por clusters.](#)
4. Muestreo aleatorio por etapas múltiples (combina 3 y 2)





Stratified Sampling Vs Cluster Sampling

- **No probabilísticas:** la muestra no probabilística no es un producto de un proceso de selección aleatoria. Los sujetos en una muestra no probabilística generalmente son seleccionados en función de su accesibilidad o a criterio personal e intencional del investigador.
 - **Muestreo por conveniencia**
 - **Muestreo discrecional**

Ejemplo: Un hospital desea hacer un estudio para testar la eficacia de su nueva vacuna contra la gripe que acaba de patentar un laboratorio farmacéutico. **Realizan el estudio sobre sus pacientes porque así al hospital le supone menos costes económicos.**

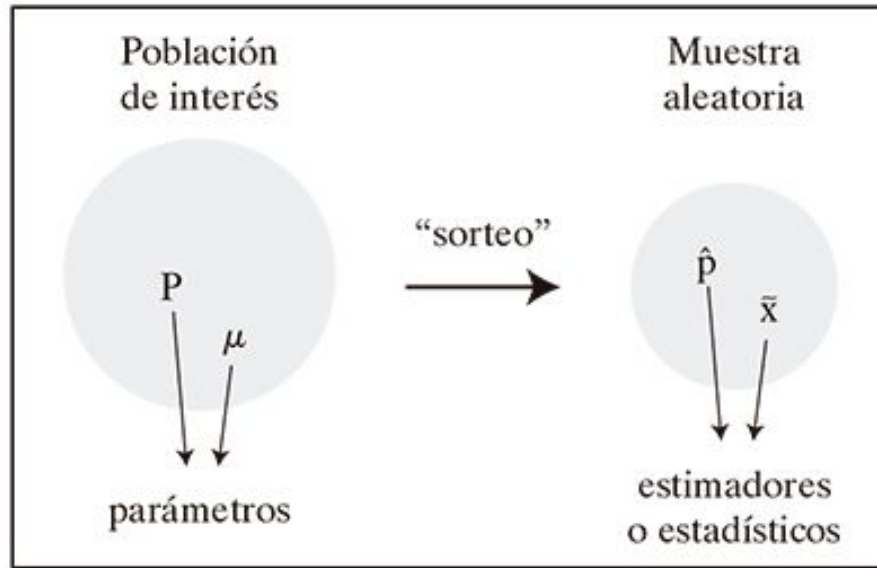


PARÁMETROS, ESTIMACIONES Y ESTIMADORES



- El objetivo siempre es **“estimar”** alguna característica de la población. que se supone fija (no aleatoria).
- Pueden ser características simples como:
 - una media, una proporción, una varianza
- O medidas más complejas, como por ejemplo:
 - los coeficientes de una regresión o la asociación entre variables
- A esta característica se la llama **parámetro**.
- En general, podemos considerar a los parámetros como relativamente “constantes” (en tiempo y espacio).
 - Esto los diferencia de los **estimadores** que veremos a continuación.

- Un estimador es un estadístico (esto es, una función de la muestra) usado para estimar un parámetro desconocido de la población.
- Por ejemplo, la **media muestral** es un estimador de la **media poblacional** que se calcula tomando el promedio de los datos.



Los estimadores son **variables aleatorias**.

1. Estimación puntual:

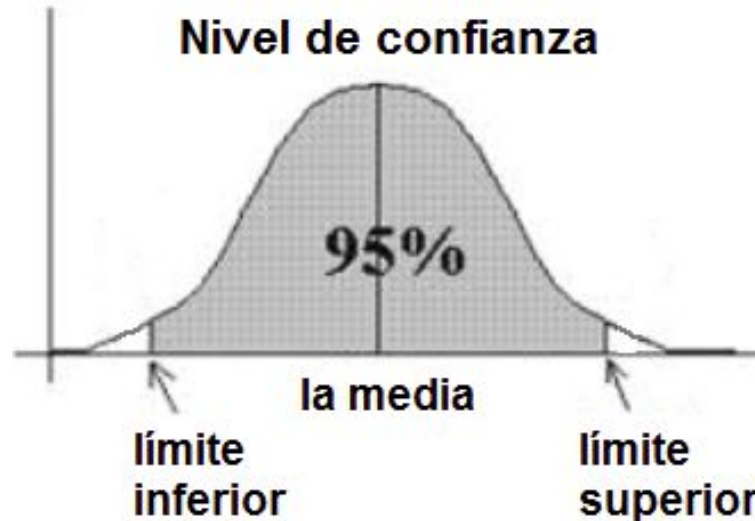
La estimación puntual consiste en utilizar el valor de un **estadístico** (alguna función de los datos) que denominaremos estimador para calcular el valor de un parámetro desconocido de una población. En la estadística clásica, esos parámetros se consideran **fijos** (no aleatorios).

- Por ejemplo, cuando usamos la media muestral para estimar la media de una población, o la proporción de una muestra para estimar el parámetro p de una distribución binomial.

Una estimación puntual de algún parámetro de una población es **un solo valor obtenido a partir de un estadístico**.

2. Estimación por intervalos de confianza

Se da una “franja” de valores posibles. Generalmente, se da un límite inferior (L_i) y otro superior (L_s) tal que la confianza de que el parámetro se encuentre entre L_i y L_s es conocida.



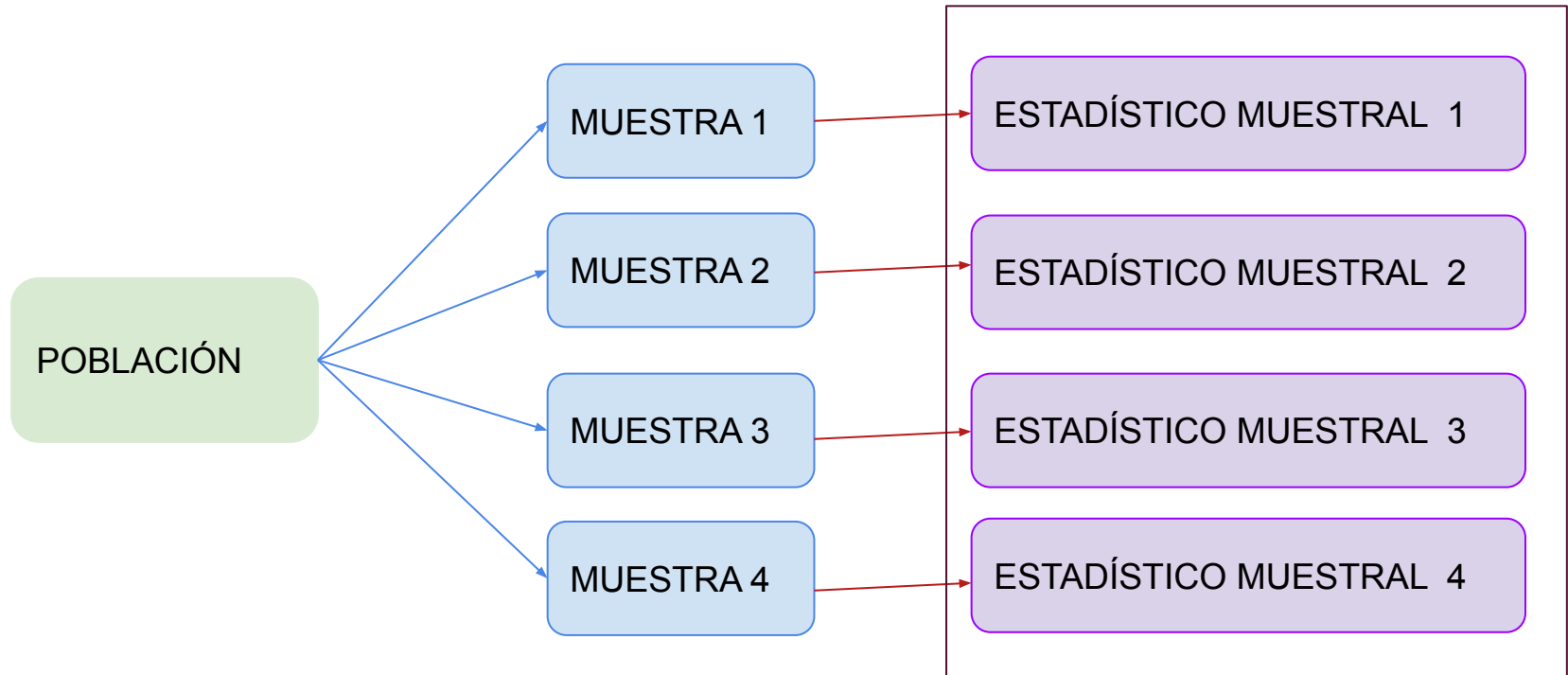
DISTRIBUCIONES MUESTRALES



Distribución Poblacional

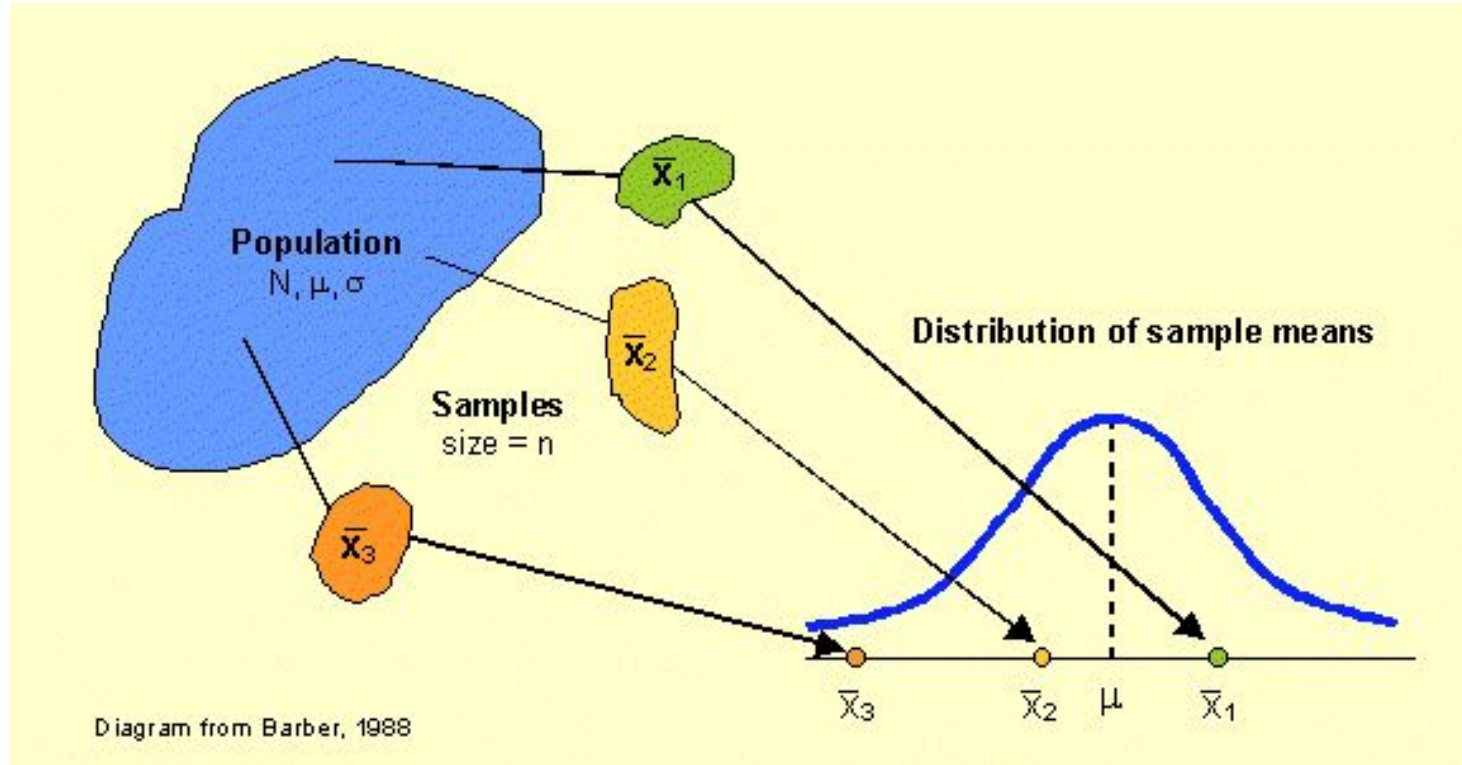
Distribución de la muestra

Distribución muestral



Distribución de las medias muestrales:

Simulación: https://gallery.shinyapps.io/CLT_mean/



LEY DE GRANDES NÚMEROS

TEOREMA CENTRAL DEL LÍMITE



— Ley de los Grandes Números

¿Por qué tiene sentido estimar la media poblacional (valor esperado) con la media muestral?

- Simplificando los planteos de **Kolmogorov**:
Supongamos que tenemos extracciones X_1, X_2, \dots, X_n .
Si las X_i son una sucesión de observaciones **independientes e idénticamente distribuidas** tales que $E(X_i)$ es igual a una constante μ finita.
Entonces el promedio de las X_i converge en probabilidad a $E(X_i)$ cuando n tiende a infinito..
- La LGN nos garantiza que, **a medida que el tamaño muestral aumenta, la media muestral se acerca a la media poblacional.**

Teorema (del límite central): Sea X_1, X_2, \dots, X_n un conjunto de variables aleatorias, independientes e idénticamente distribuidas de una distribución con media μ y varianza $\sigma^2 \neq 0$. Entonces, si n es suficientemente grande, la variable aleatoria

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

tiene aproximadamente una distribución normal con $\mu_{\bar{X}} = \mu$ y $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$.

Condiciones del TCL

- **Independencia:** las observaciones tienen que provenir de un muestreo aleatorio y deben ser independientes. Si el muestreo es sin reemplazo, entonces $n < 10\%$ de la población.
- Si la **distribución** de una población es muy **asimétrica**, **n deberá ser muy grande** (a veces se menciona un $n=30$ como aproximación, pero en realidad depende de cuán asimétrica es la población). Si la población tiene distribución normal, no hay condiciones sobre n .

- Teorema Central del Límite y muestreo:
 - El TCL nos da una distribución para nuestro estimador de la media poblacional dado por la **media muestral**.
 - El **desvío estándar** de la distribución muestral se llama **error estándar (SE)** de la media muestral.

Error Estándar:
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- Teorema Central del Límite y muestreo:
 - El TCL nos da una distribución para nuestro estimador de la media poblacional dado por la **media muestral**.
 - El **desvío estándar** de la distribución muestral se llama **error estándar (SE)** de la media muestral.
 - En el caso que no dispongamos del desvío estándar de la población, el error estándar se aproxima utilizando el **desvío estándar muestral**:

Error Estándar: $\sigma_{\bar{x}} = \frac{\cancel{\sigma}^S}{\sqrt{n}}$

- Teorema Central del Límite y muestreo:
 - El TCL nos da una distribución para nuestro estimador de la media poblacional dado por la **media muestral**.
 - El **desvío estándar** de la distribución muestral se llama **error estándar (SE)** de la media muestral.
 - En el caso que no dispongamos del desvío estándar de la población, el error estándar se aproxima utilizando el **desvío estándar muestral (S)**.
 - El **error estándar (SE)** nos permite medir dispersión respecto a la media muestral, ajustando por el tamaño muestral (podemos comparar dispersión para estimadores de la media que usan diferentes tamaños muestrales).

Error Estándar:
$$\sigma_{\bar{x}} = \frac{\overset{\text{S}}{\cancel{\sigma}}}{\sqrt{n}}$$

ALGUNOS ESTIMADORES PUNTUALES



POPULATION PARAMETER	ESTIMATOR	ESTIMATE
Mean (μ)	$\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$	\bar{x}
Variance (σ^2)	$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$	s^2
StandartDeviation (σ)	$S = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$	s
Proportion (P)	$\hat{P} = \frac{X}{n}$	\hat{p}

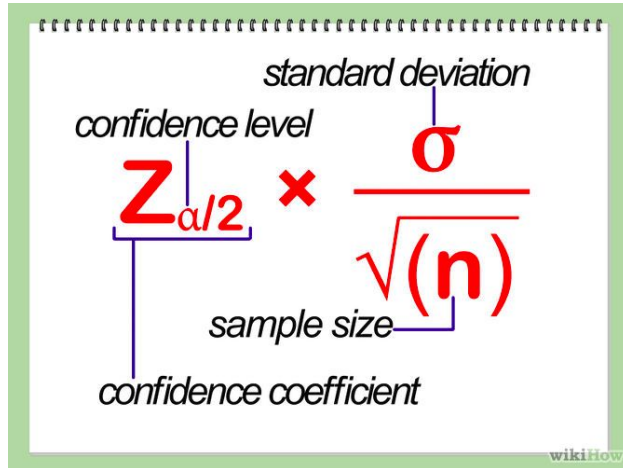
INTERVALOS DE CONFIANZA



- **Media de la muestra:** estimación puntual => es el valor del estadístico en esa muestra.
 - No sirve de mucho. ¿Por qué?
- **Estimación por intervalos:** la idea es dar un rango de valores posibles para el parámetro con un valor de confianza asociado.
 - «El parámetro de la población está entre 4,5 y 8,2 con una confianza del 95%»
- **¿Cómo se logra?** A partir de poder obtener un estadístico cuya distribución sea conocida y no dependa de parámetros desconocidos.

$$\bar{x} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{(n)}}$$

Margen de Error:



A diagram of the Margin of Error formula, $Z_{\alpha/2} \times \frac{\sigma}{\sqrt{(n)}}$, enclosed in a green rectangular frame. The formula is written in red. Labels with blue lines pointing to the components are: 'confidence level' pointing to $Z_{\alpha/2}$, 'confidence coefficient' pointing to the same $Z_{\alpha/2}$, 'standard deviation' pointing to σ , and 'sample size' pointing to (n) inside the square root. A 'wikiHow' watermark is visible in the bottom right corner of the frame.

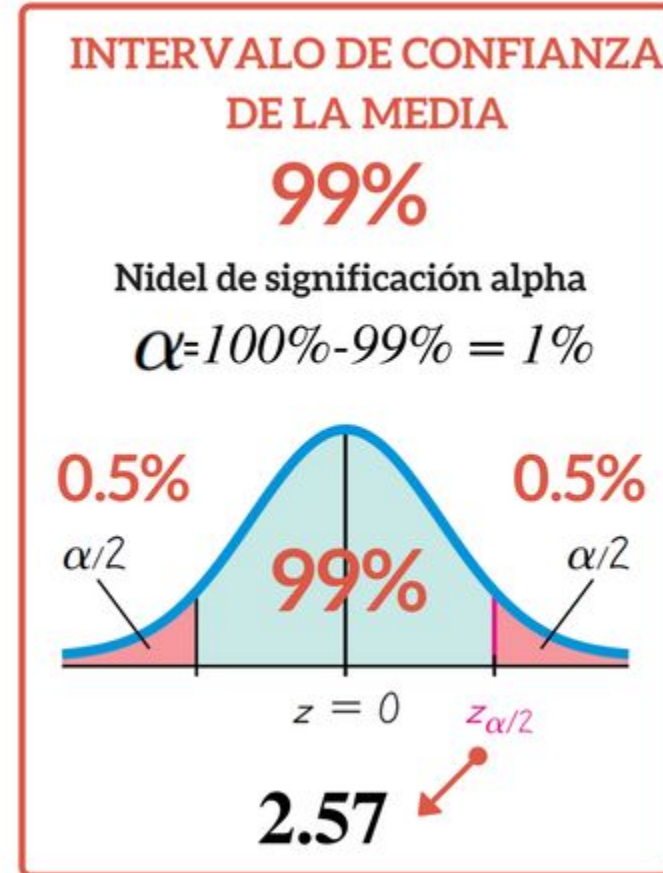
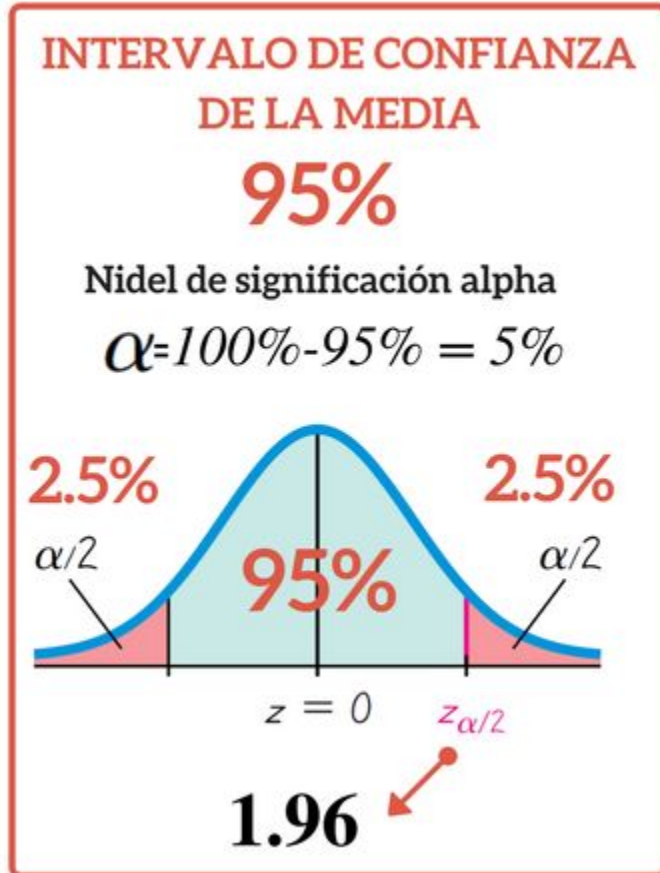
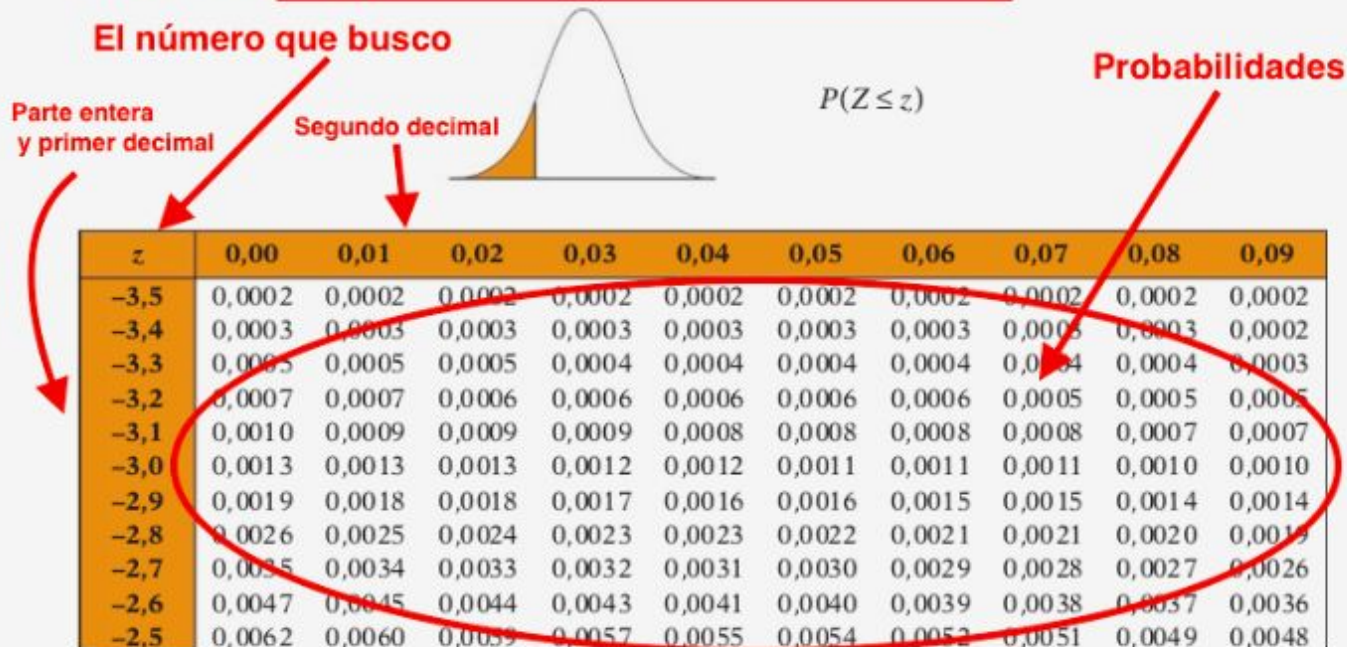


TABLA III: DISTRIBUCIÓN NORMAL TIPIFICADA





$$P(Z \leq z)$$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767

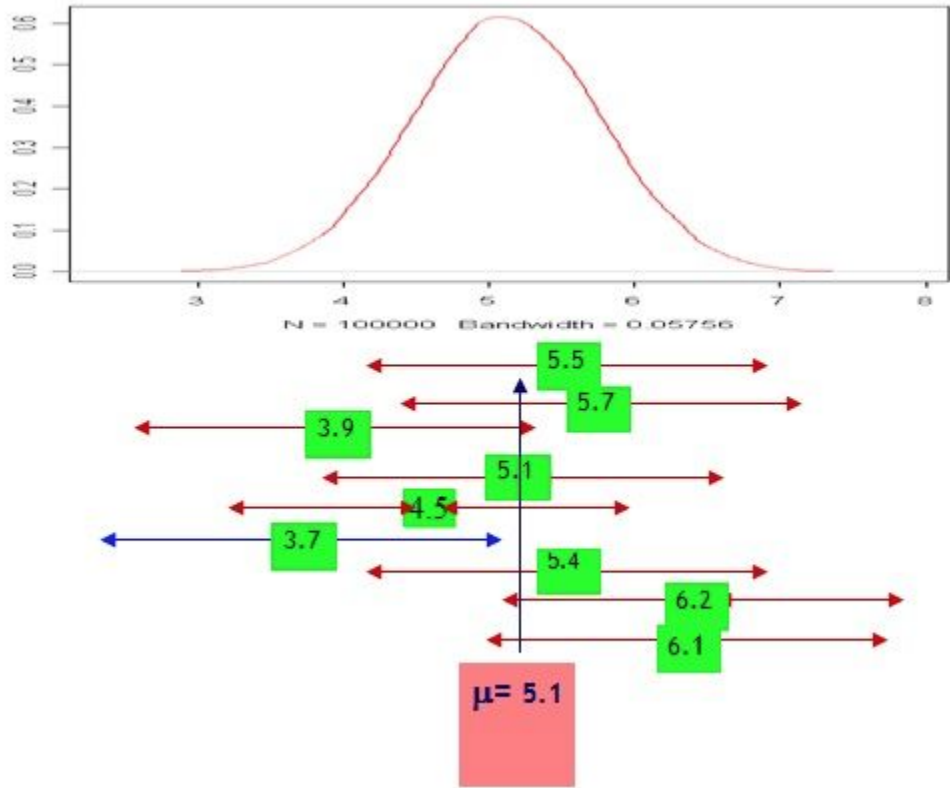
- En nuestro caso de varianza poblacional conocida el largo del intervalo de confianza resulta

$$Length = \left[X + z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) \right] - \left[X - z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right) \right]$$

$$Length = 2z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

- El largo del intervalo de confianza depende de
 - la varianza poblacional
 - el tamaño muestral
 - el nivel de confianza

El intervalo de confianza nos dice que de todas las veces que tomemos una muestra y calculemos el IC, $(1- \alpha)\%$ tenderá a contener al verdadero parámetro poblacional.

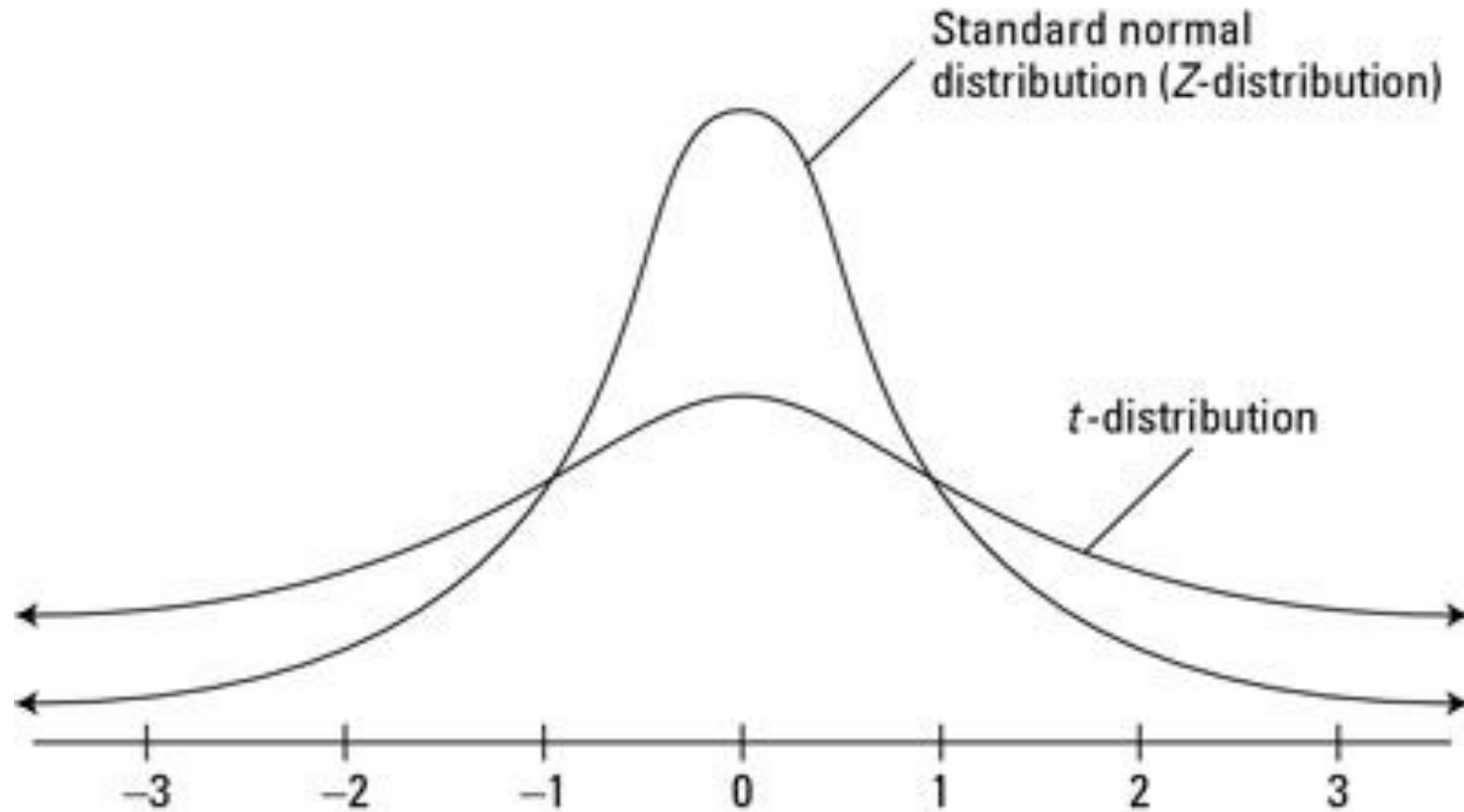


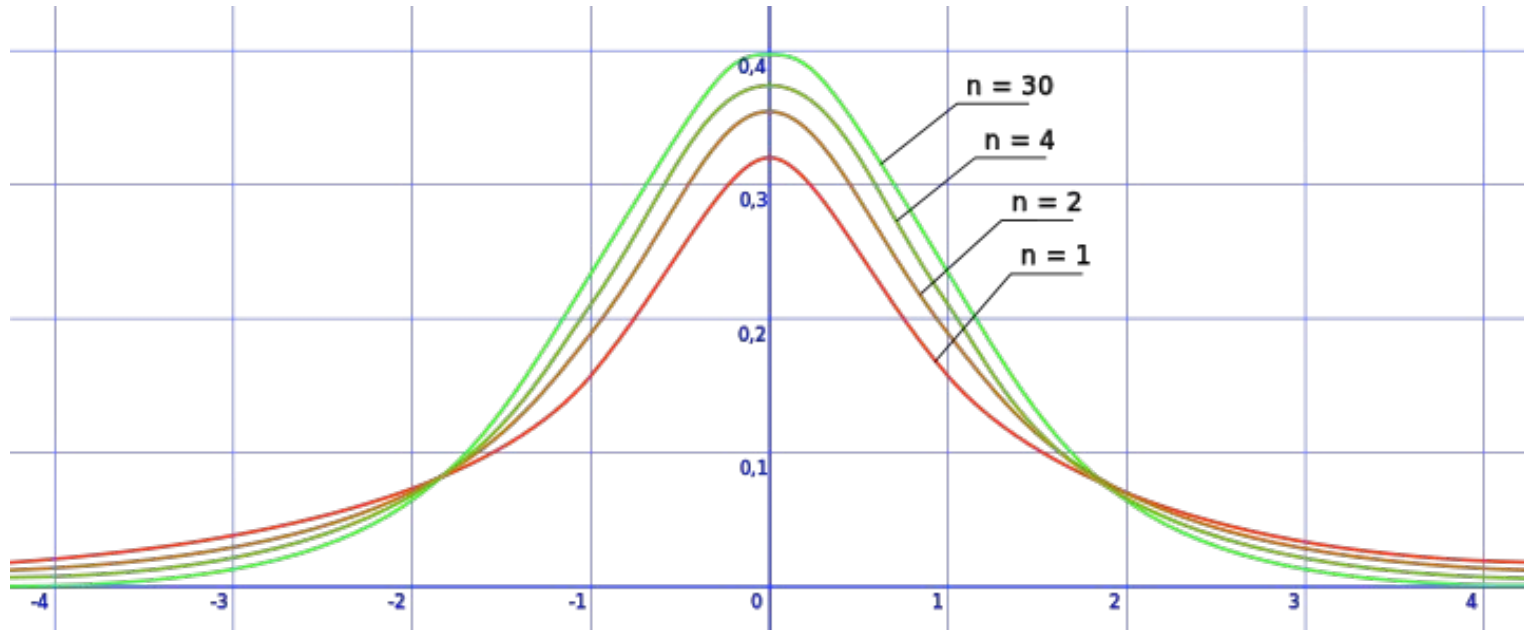
- Se desea analizar el ingreso promedio de los científicos de datos. A partir de una muestra al azar de 100 individuos, cuyo ingreso promedio es de 30 mil dólares y desvío estándar de aproximadamente 7 mil dólares, se desea construir un intervalo de confianza del 90% para el ingreso promedio poblacional.
- Dado que n puede ser considerado "grande" $\Rightarrow \bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$
- El borde inferior del intervalo resulta $30 - 1.645 * 7 / \text{raíz}(100) = 28.85$
- El borde superior del intervalo resulta $30 + 1.645 * 7 / \text{raíz}(100) = 31.15$

Con una confianza del 90%, el intervalo que va de 28.85 a 31.15 contendrá la media poblacional de los ingresos.

- Estamos más interesados, por supuesto, en obtener intervalos de confianza lo más estrechos posible.
- Después de todo, ¿Cuál de las siguientes afirmaciones es más útil?
 - *Podemos estar seguros al 95% de que la cantidad promedio de dinero que se gasta mensualmente en alquileres se encuentra entre 300 y 3300 USD.*
 - *Podemos tener una confianza del 95% en que la cantidad promedio de dinero que se gasta mensualmente en alquileres se encuentre entre 1100 y 1300 USD.*

- ¿Qué pasa cuando no conocemos el desvío estándar poblacional?
 - Cuando **n** es **grande**, entonces **el desvío estándar de la muestra S** es una **buena aproximación a sigma**.
 - Cuando **la muestra es chica**, entonces usamos la distribución **T-Student**. Normalmente, se considera chica una muestra donde $n < 30$ pero en realidad depende de la distribución muestral.
 - La distribución T-Student es simétrica pero dada una probabilidad $\alpha/2$ que se busca acumular en cada cola, el valor t de la distribución T-Student que acumula esa probabilidad es mayor en valor en la normal estándar. ¿Intuición? Esto genera **intervalos más anchos** para incorporar el hecho de que el desvío estándar muestral es un estimador del poblacional.
 - **Usar la distribución T-Student es más conservador, por lo que si sigma es desconocido, lo recomendable es usar la distribución T.**





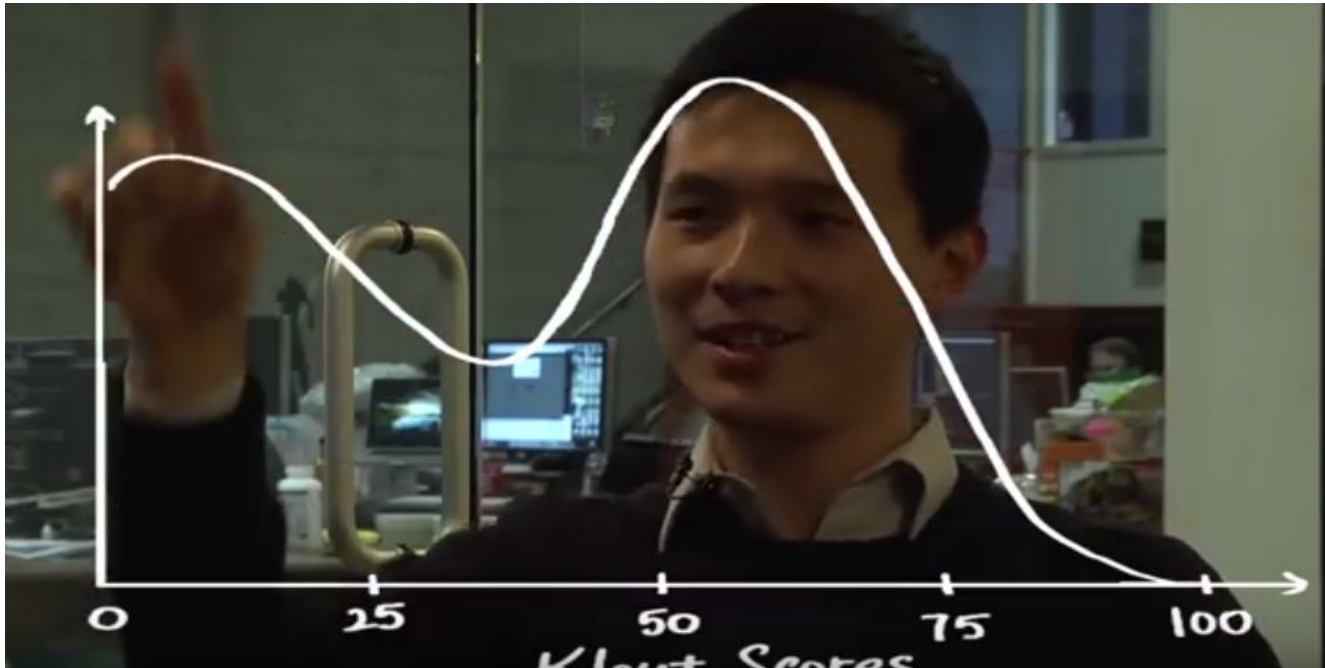
- La curva t tiene media = 0
- La curva t es más dispersa que la normal
- A medida que aumentan los grados de libertad, la dispersión disminuye y la curva t se aproxima a la normal. grados de libertad: $\mathbf{v = n - 1}$

- Observaciones:
 - El **intervalo de confianza** es siempre **función de la muestra que es aleatoria**. ¡Entonces los bordes del intervalo también son variables aleatorias! En este caso el borde del intervalo depende de la media muestral (que cambia con cada muestra).
 - El intervalo de confianza no puede depender de parámetros que sean desconocidos. En nuestro ejemplo usamos un estimador del desvío poblacional o asumimos conocido el desvío poblacional.

PRÁCTICA GUIADA



- Imagine que nos interesa estudiar la distribución de los Klout scores. La población completa tiene un histograma como el de la imagen tal como muestra su creador.
 - ¿Qué observa sobre esta distribución?



- Tomamos como **población** las **1048 observaciones de Klout scores** con las que contamos (klout_scores.csv).

¿Qué pasa si...

- Obtenemos una muestra de **$n=5$** y tomamos el promedio.
Repetimos este paso 1000 veces para poder hacer un histograma de las medias.
- Obtenemos una muestra de **$n=10$** y tomamos el promedio.
Repetimos este paso 1000 veces.
- Obtenemos una muestra de **$n=100$** y tomamos el promedio.
Repetimos este paso 1000 veces.

¿Qué observamos sobre el **histograma de la media** a medida que aumentamos el tamaño muestral?

- Tomamos como población las 1048 observaciones de Klout scores:
 - **Intuición:**

El promedio es la suma de variables aleatorias dividido por el tamaño muestral. En este caso cada Klout score de la muestra es una variable aleatoria porque no conozco si determinada observación estará o no en la muestra.
 - Al sumar variables aleatorias, los valores extremos de esa suma se vuelven menos frecuentes y eso lleva la distribución (histograma) hacia una forma acampanada y más simétrica.