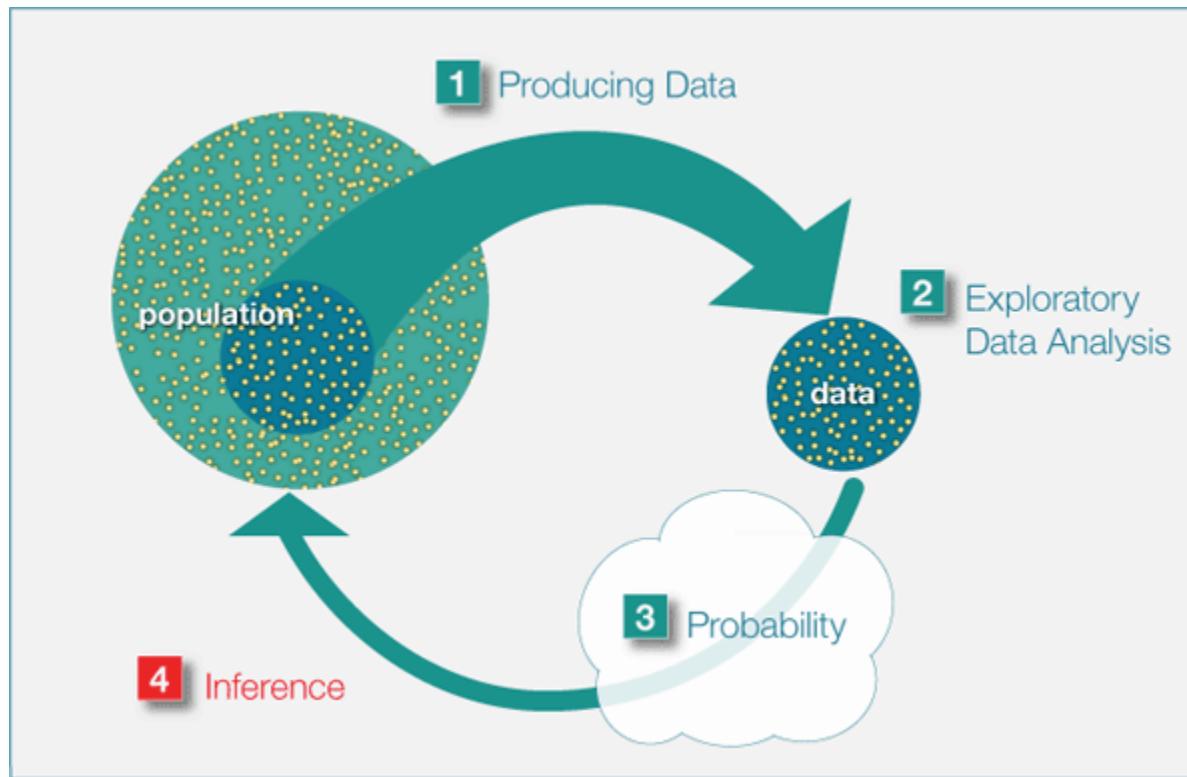


Inferencia Estadística



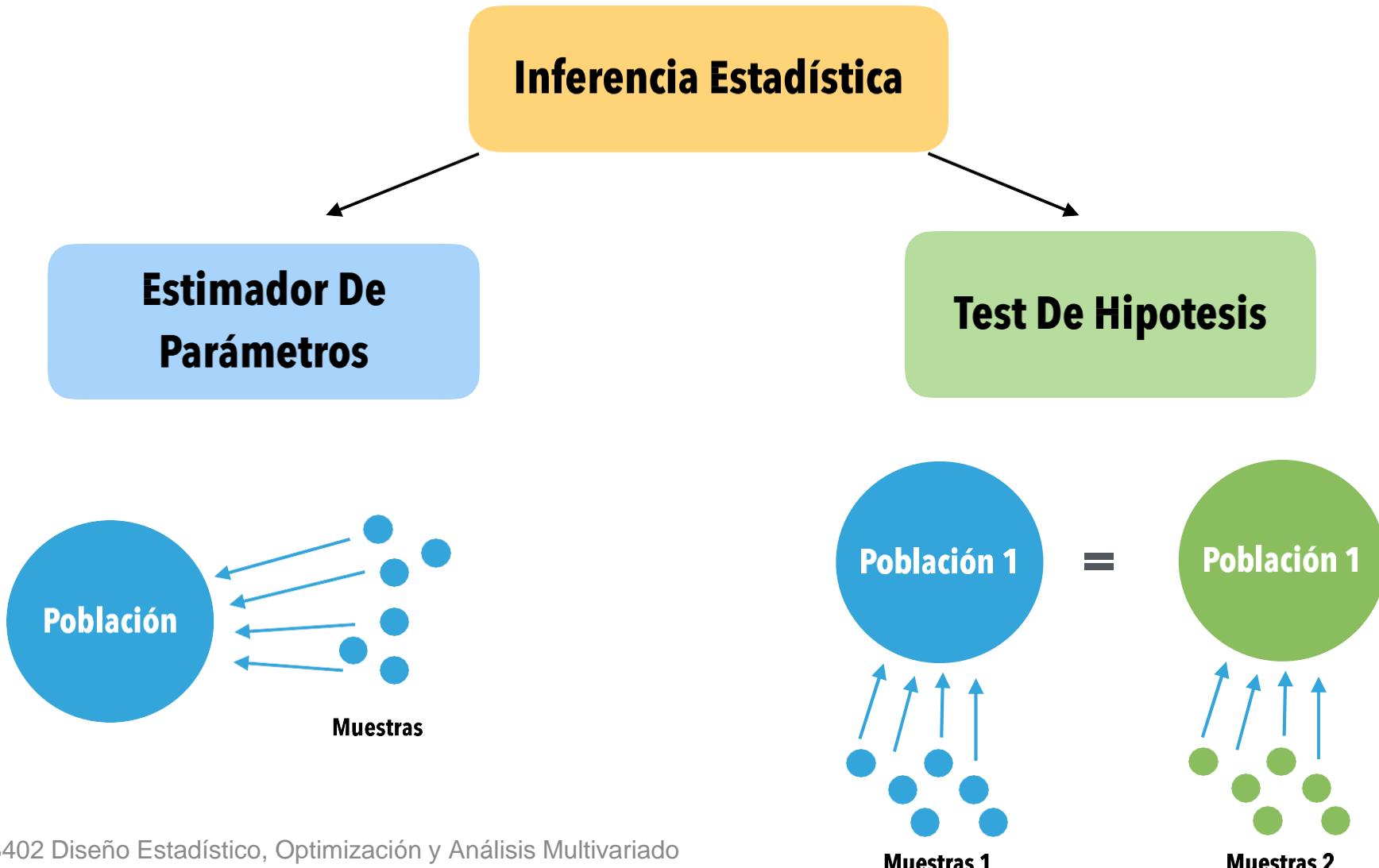
Profesor: Pedro Saa (pnsaa@uc.cl)
Año: 1-2025

OBJETIVOS DE APRENDIZAJE

- ▶ **O1:** Comprender que la estimación puntual y es un estimador de un parámetro poblacional tiene **variabilidad** (error estándar) y la distribución de esta variabilidad construye la **distribución muestral**
 - ▶ **O2:** Entender que el teorema del límite central es sobre la **distribución muestral** y, que bajo ciertas condiciones, esta es casi **normal**
 - ▶ **O3:** Definir y explicar que un intervalo de confianza es un **rango plausible de valores** para un **parámetro de población**.
-
- ▶ **O4:** Evaluar cuando es necesario aplicar un test de hipótesis de una o dos colas
 - ▶ **O5:** Aplicar test de hipótesis para contrastar hipótesis utilizando **intervalos de confianza o valor-p**
 - ▶ **O6:** Reconocer que toda test de hipótesis conlleva un **error medible (tipo I o tipo II)**

La inferencia estadística es el conjunto de métodos que permiten **estimar**, a través de una **muestra**, el **comportamiento** de una determinada **población** con un riesgo de **error medible** en términos de probabilidad

Inferencia estadística busca tomar decisiones sobre el comportamiento de la población, esta se puede apertura en inferencia sobre **parámetros de población** o **test de hipótesis**



En esta clase vamos a trabajar con los datos de bienes raíces del estado de Ames en Iowa, USA. Estos corresponden a los datos de **venta de casas**, con sus características durante el periodo 2016-2010



```
1 path = kagglehub.dataset_download("marcopale/housing") # descargamos el set de datos  
2 datos = pd.read_csv(path+'/AmesHousing.csv')[['PID', 'Lot Area', 'SalePrice']]  
3 datos.head() # por defecto son las primeras 5 filas
```

Descripción de los datos que vamos a utilizar

PID: ID de la casa

Lot Area: Área de la casa

SalePrice: Precio de Venta de la casa

	PID	Lot Area	SalePrice
0	526301100	31770	215000
1	526350040	11622	105000
2	526351010	14267	172000
3	526353030	11160	244000
4	527105010	13830	189900

Vamos a considerar que estos datos representan la **población completa** de venta de casas

Variabilidad en estimadores

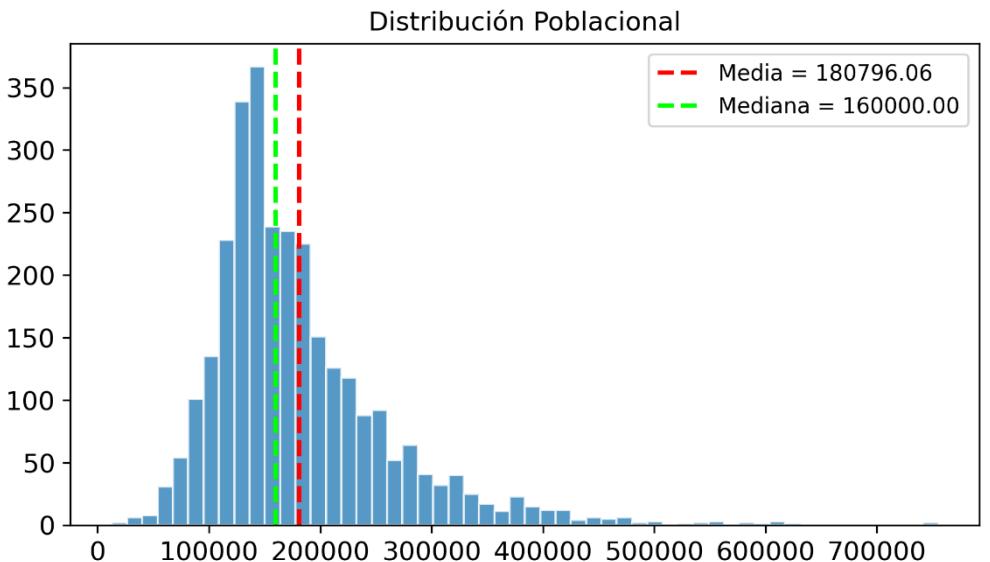
Preguntas a resolver

¿Cuál es el área promedio de una casa?

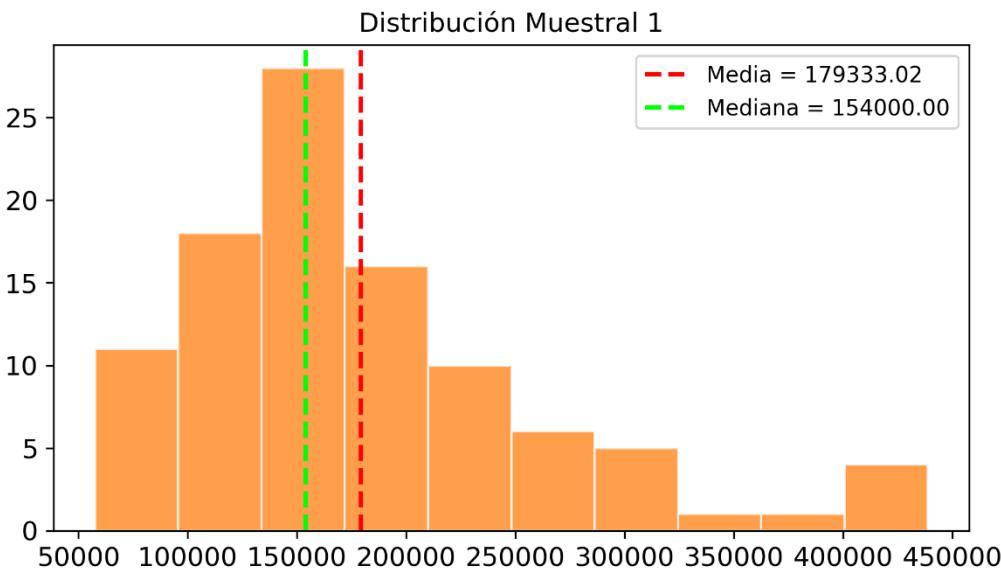
¿Cuál es el precio promedio de una casa?

Dado que no podemos consultar toda la población (¿Por qué?), es necesario tomar una muestra (\bar{x} o s) y a partir de ella determinar el valor de la población (μ o σ)

La **estimación puntual** corresponde a un valor único de nuestra muestra que mejor representa nuestro, por ejemplo, promedio poblacional (μ)



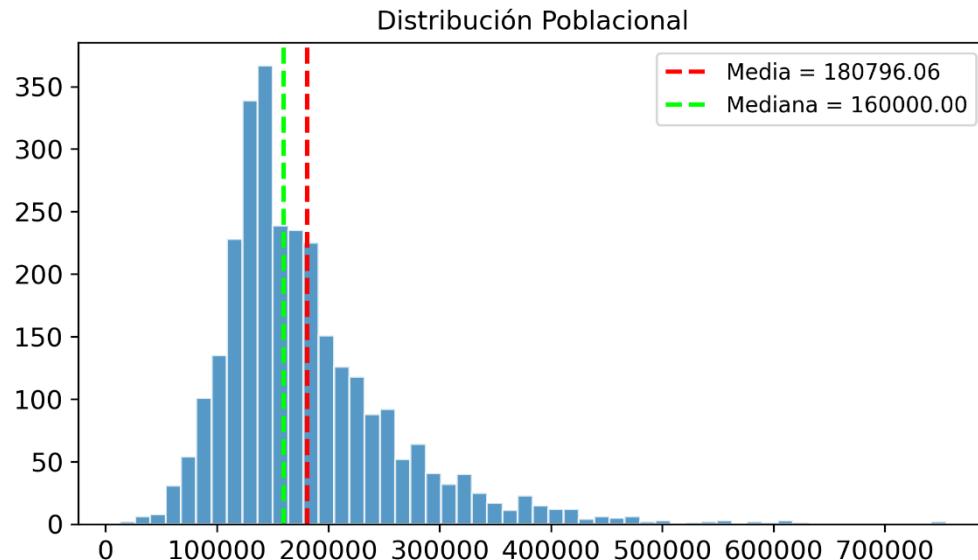
```
● ● ●  
1 prices_df = datos['SalePrice']  
2 histogram_described(prices_df, mode = False,  
density=False)
```



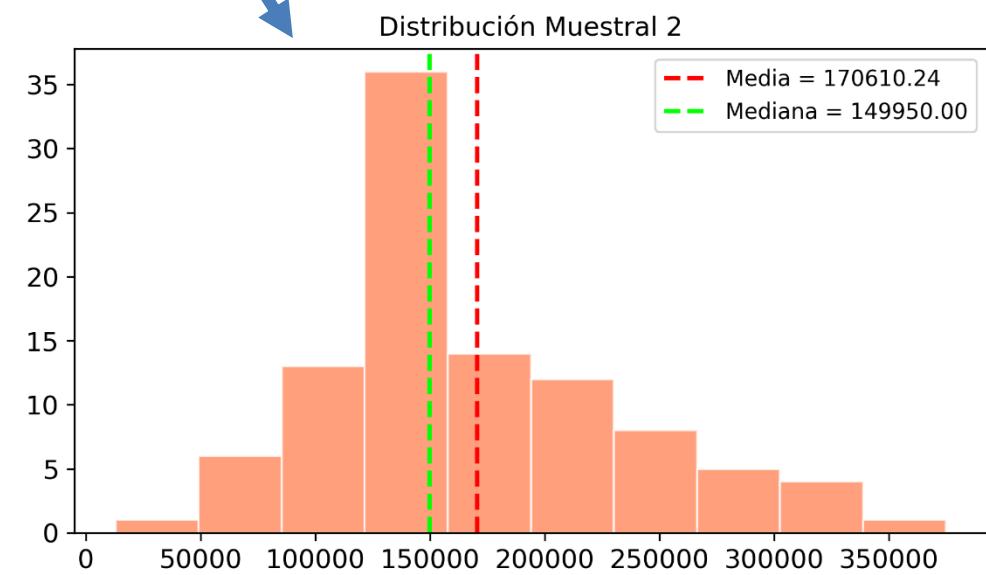
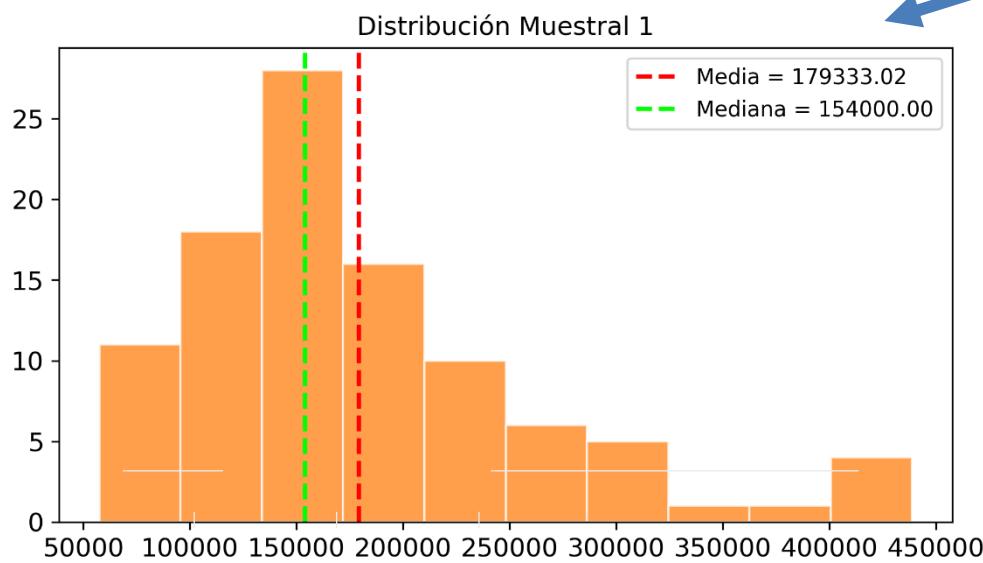
```
● ● ●  
1 muestra1 = prices_df.sample(100, random_state=41)  
2 histogram_described(muestra1, mode = False,  
density=False, hist_color = 'tab:pink')
```

*La función histogram_described la pueden encontrar en el colab de la clase

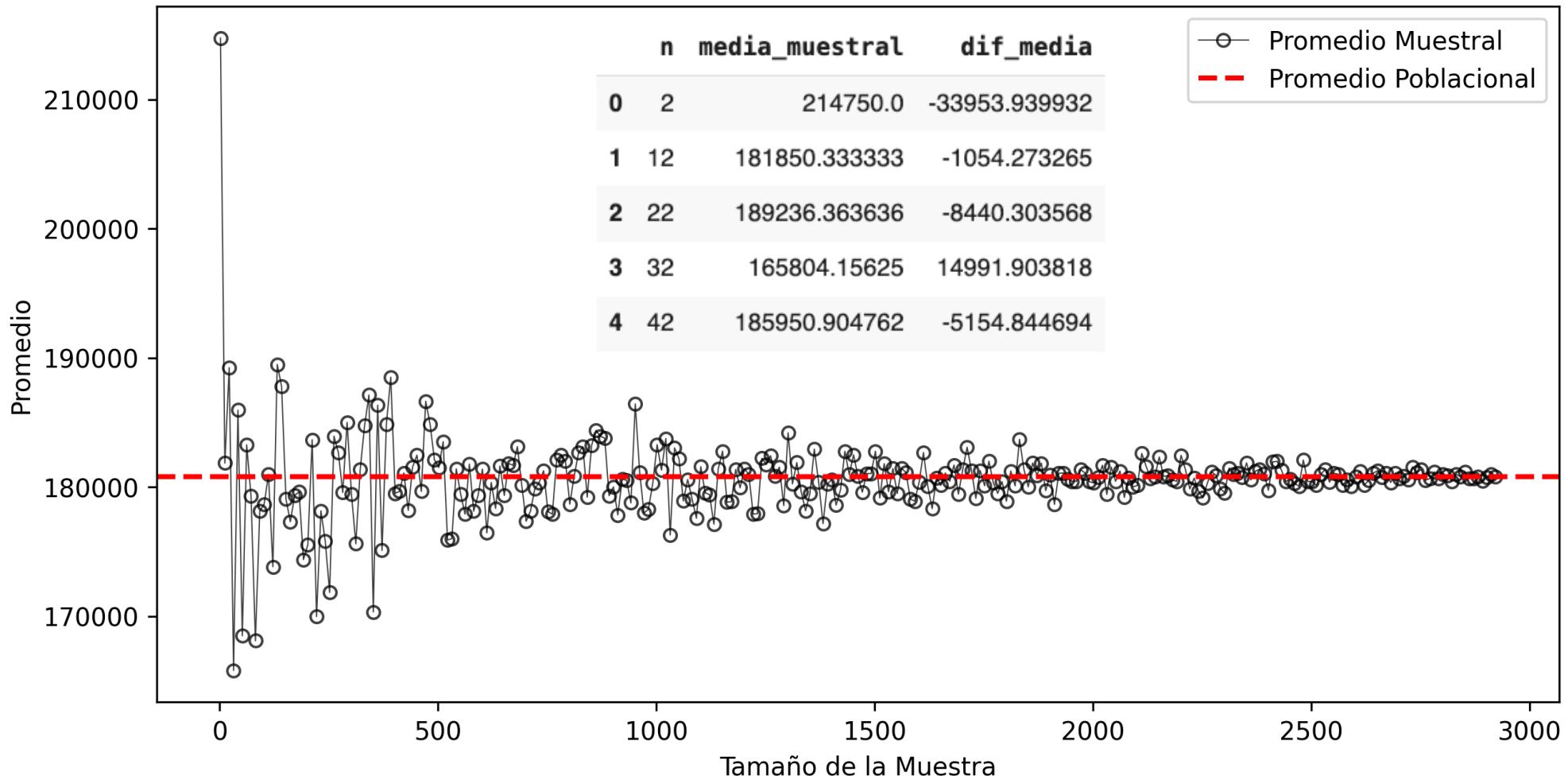
La **estimación puntual** corresponde a un valor único de nuestra muestra que mejor representa nuestro, por ejemplo, promedio poblacional (μ)



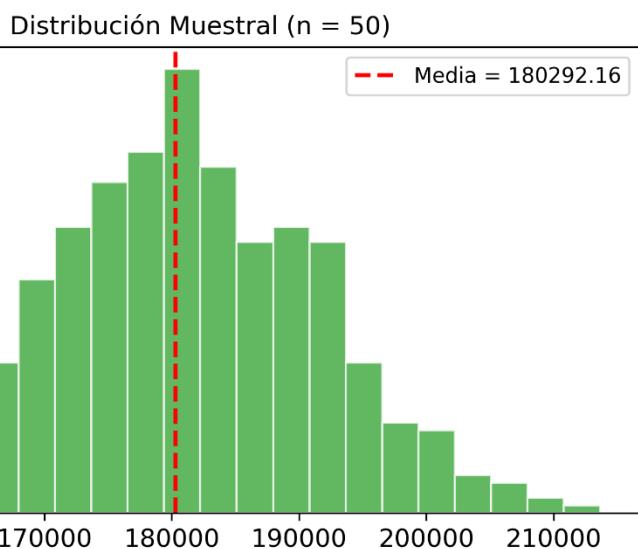
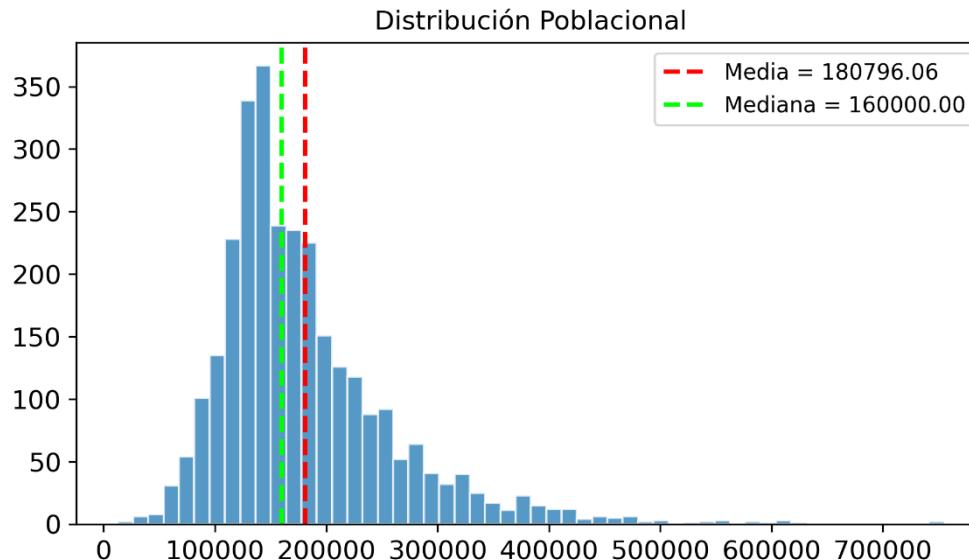
Muestras de **100** datos aleatorios



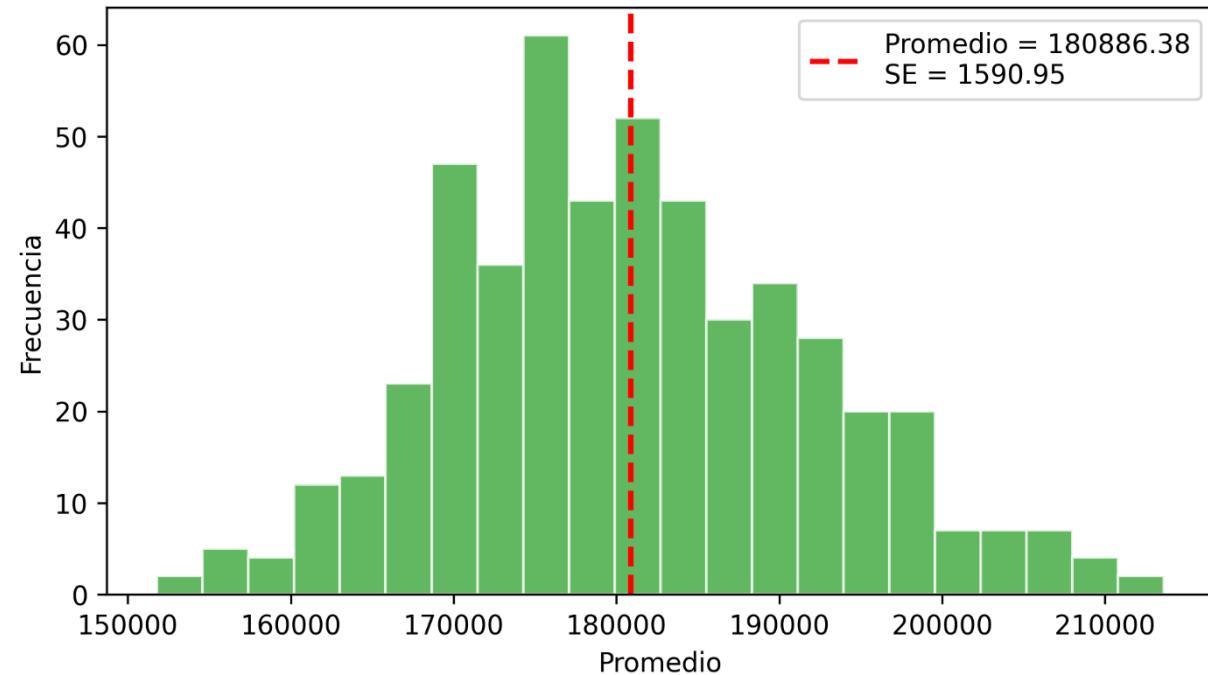
La estimación puntual no es exacta y presenta variabilidad en cada muestreo, no obstante, esta estimación mejora a medida que **disponemos de mas datos**



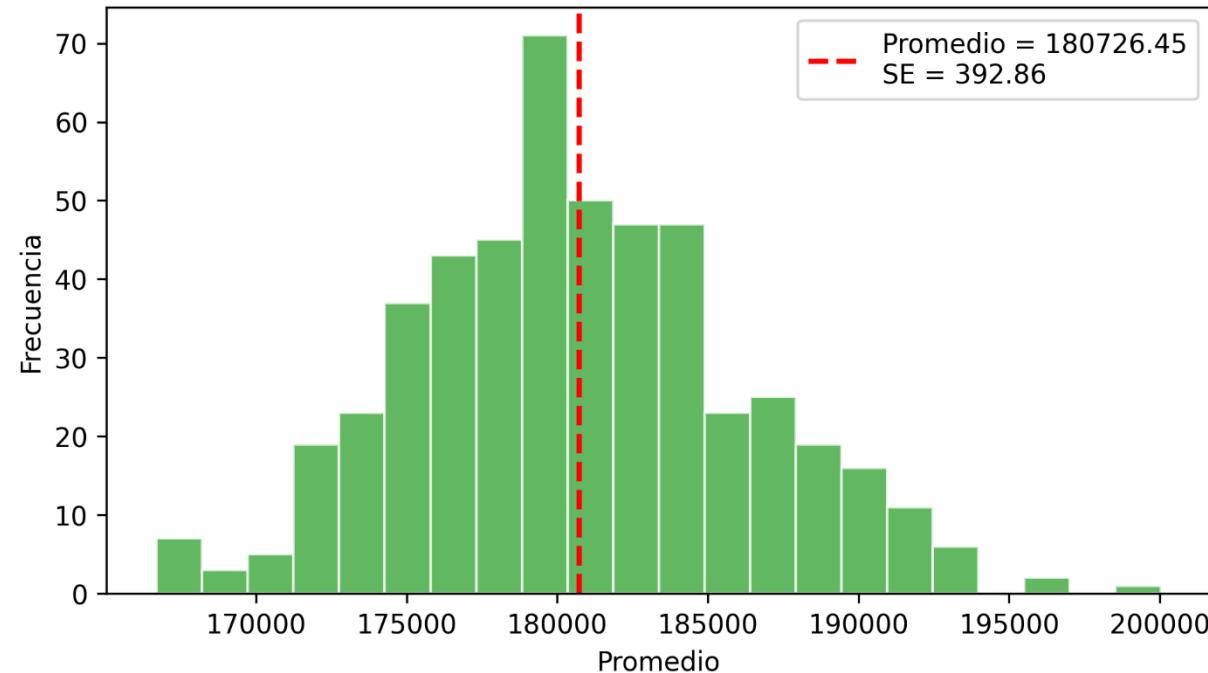
La **distribución muestral** corresponde a todas las distribuciones de estimaciones puntuales que se pueden obtener a partir de una muestra de **tamaño fijo**



Observamos que la distribución muestral (500 promedios) es unimodal y **centrado** en torno al verdadero promedio poblacional (μ).

Distribución Muestral ($n = 50$)

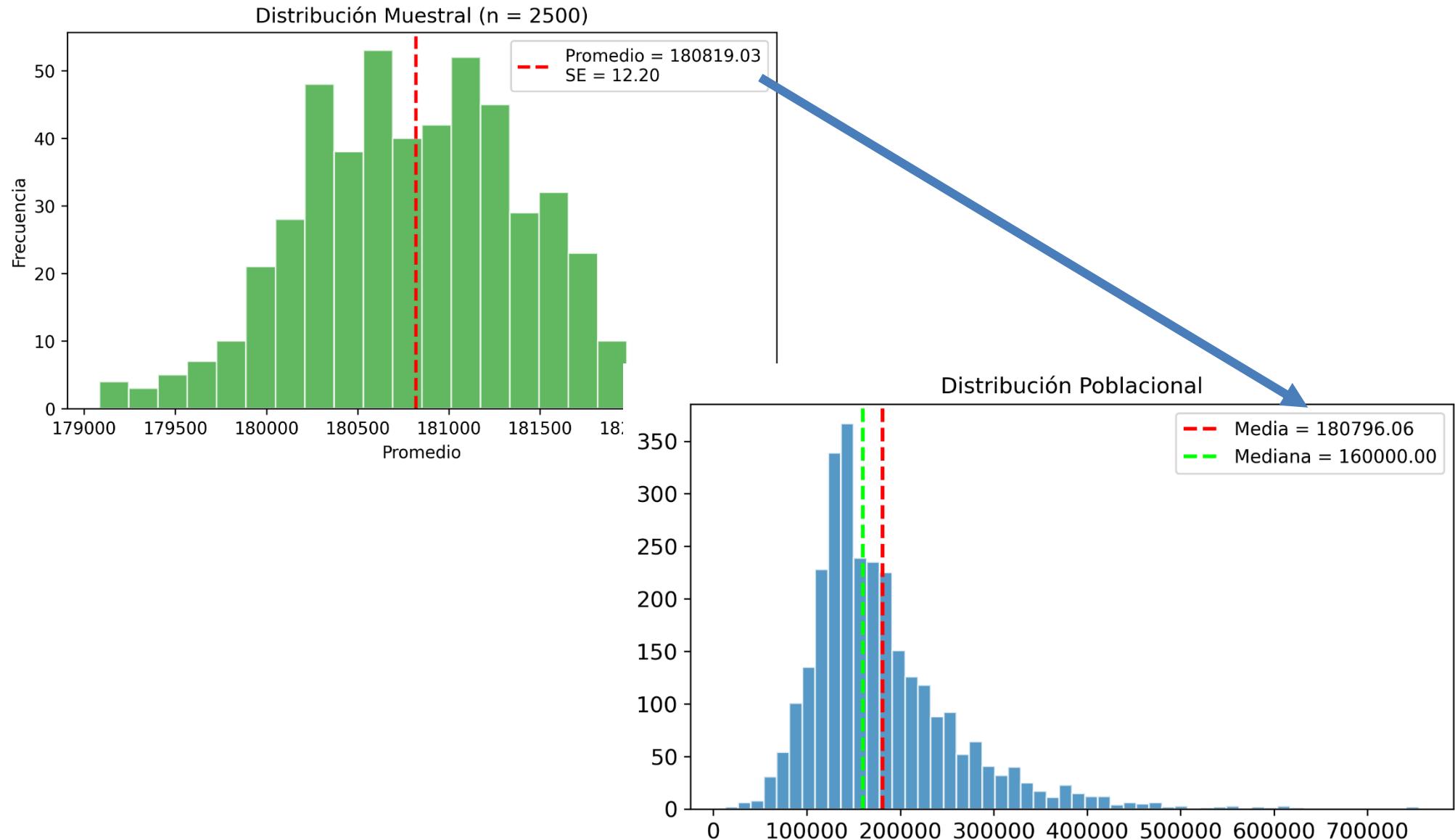
$$SE = \frac{\sigma}{\sqrt{n}}$$

Distribución Muestral ($n = 200$)

El **error standar** nos permite cuantificar la **incertidumbre** de nuestra estimación puntual de la muestra, en este caso es el **error standar del promedio muestral**

... Pero, ¿Qué sucederá si tomamos una muestra de tamaño **cercano** al tamaño poblacional?

El **error estándar del promedio muestral** se acerca a **cero**, es decir tenemos total certeza que la estimación puntual corresponde al promedio poblacional



Desviación estandar vs Error estandar

En general en la literatura científica los datos se resumen usando el promedio y la **desviación estandar** de la muestra o el promedio con el **error estandar** de la muestra

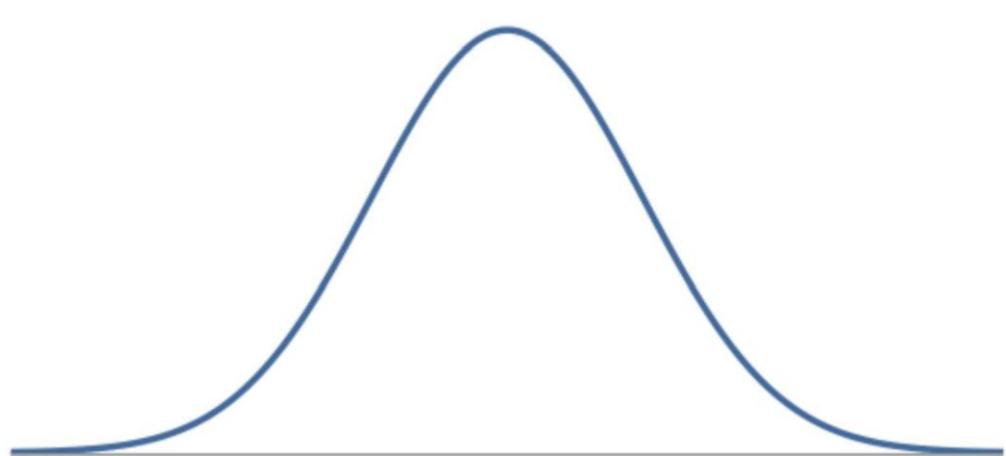
Esto genera confusiones en muchos casos por la intercambiabilidad de ambos

El promedio y la **desviación estandar** muestral son **estadísticas descriptivas de la muestra**, mientras que el **error estandar** de la muestra es descriptivo al **proceso de muestreo aleatorio**

En palabras simples, el **error estandar del promedio** de la muestra es un estimador de cuan lejos está el **promedio de la muestra al promedio de población**, en cambio, la **desviación estandar** de la muestra es la **dispersión de las observaciones** con respecto al promedio de la muestra

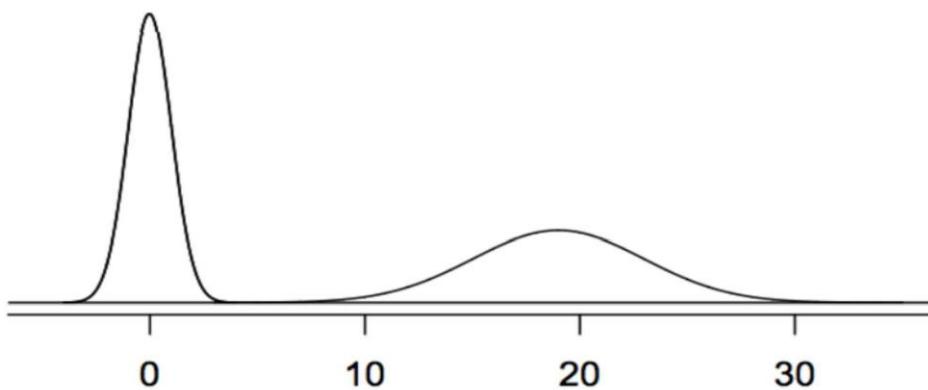
Teorema del límite central

Se llama distribución **normal** o de **Gauss** a la distribución de probabilidades que con más frecuencia aparece en la estadística



$$N(\mu = 0, \sigma = 1)$$

$$N(\mu = 19, \sigma = 4)$$

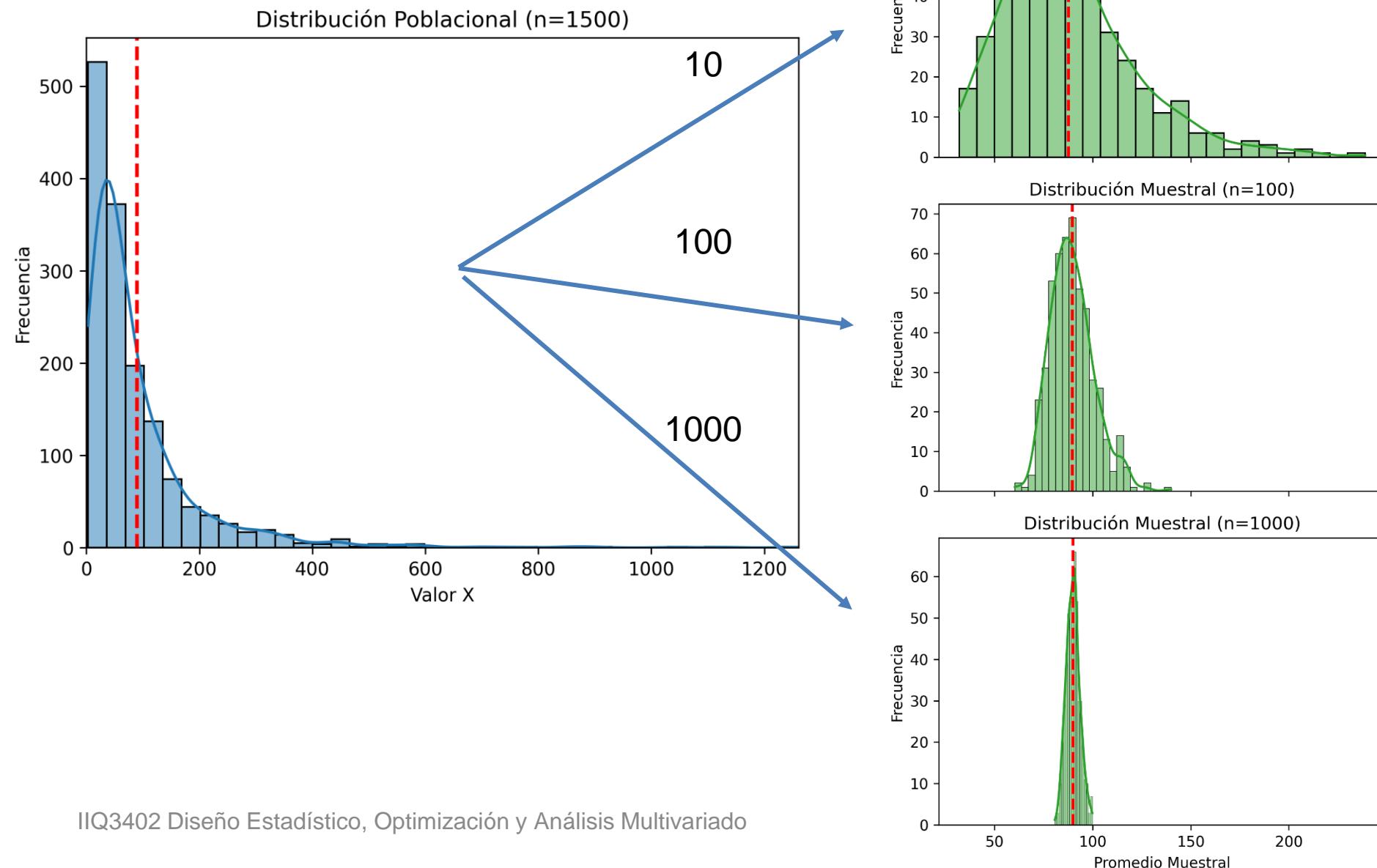


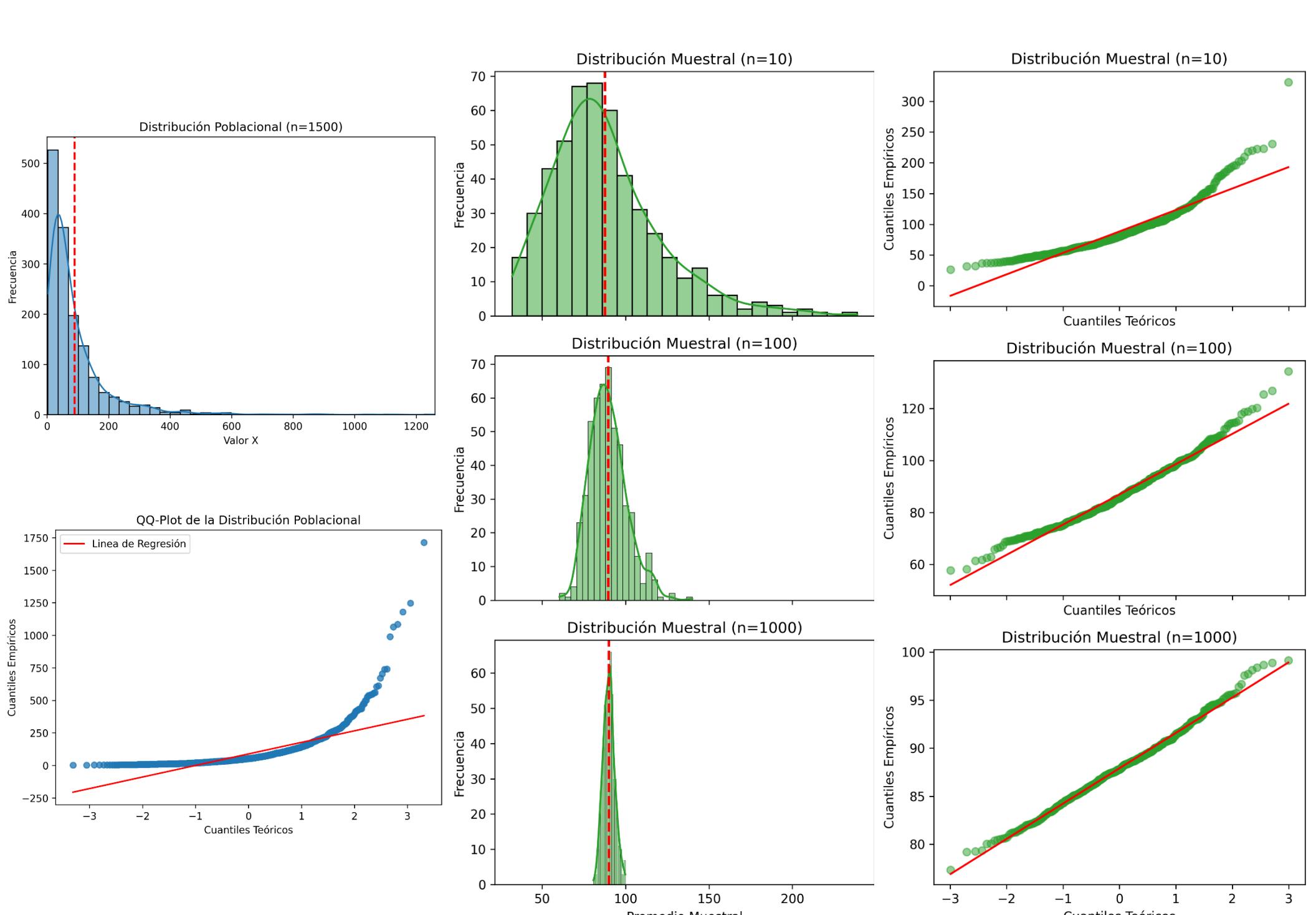
Es **Unimodal** y **simétrica**, con forma de campana

Muchas variables son “**casi normales**” y por tanto modelarlos con esta distribución nos permite obtener Insights del fenómeno

Se anota como $N(\mu, \sigma)$: normal con **promedio μ** y **desviación estándar σ**

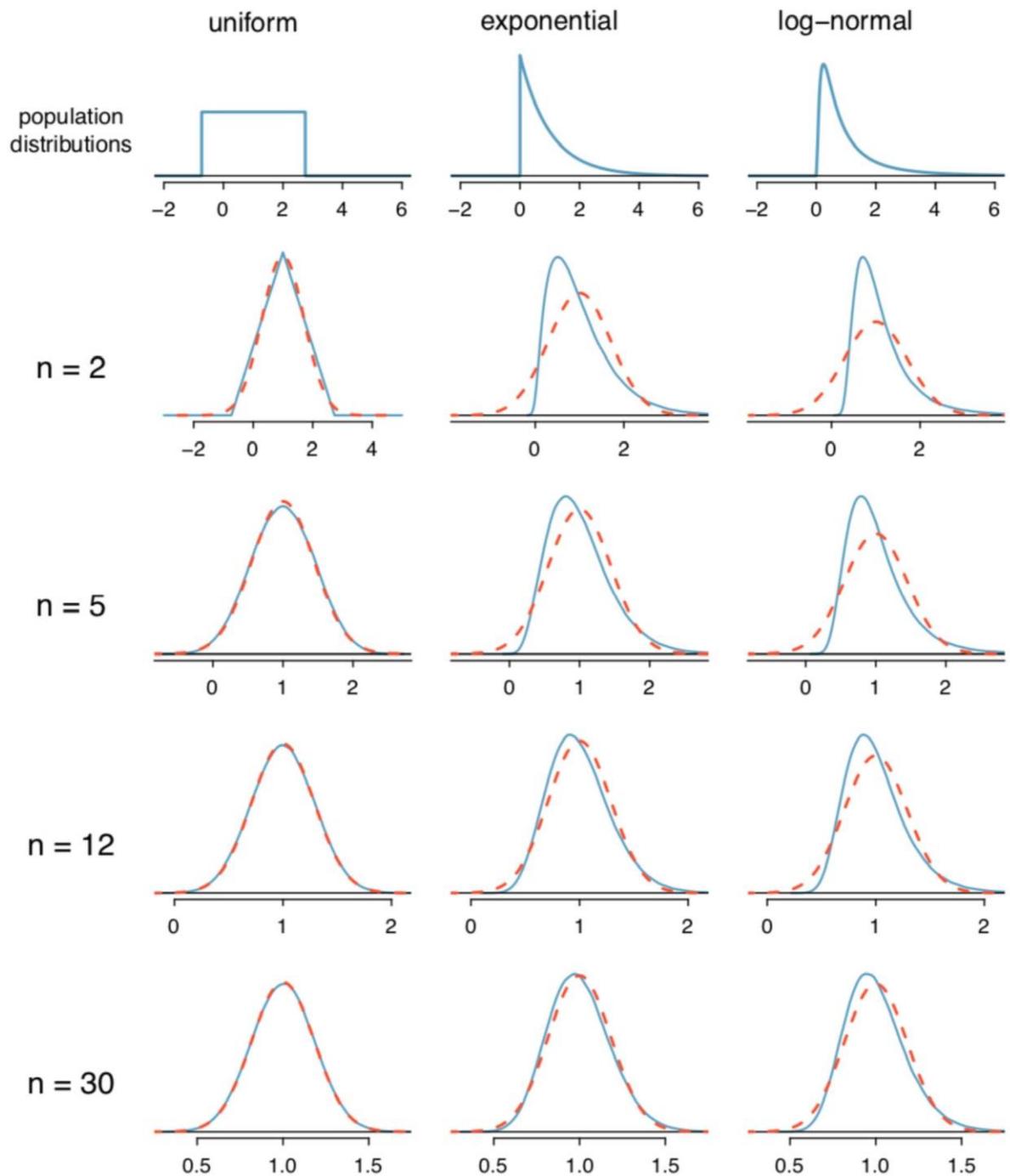
Observamos que en general la distribución muestral se **normaliza** en torno al promedio real (poblacional)





La distribución muestral se acerca a una **distribución normal**, a medida que **aumenta el tamaño de la muestra**

https://gallery.shinyapps.io/CLT_mean



El **teorema de límite central** dice que la distribución muestral es **aproximadamente normal**, centrado en torno al promedio poblacional, considerando que el **tamaño de la muestra** es suficientemente grande para compensar la asimetría

Regla general

Independencia

**La muestra consiste en observaciones independientes
(asignación aleatoria o muestra aleatoria)**

**Si se muestrea sin reemplazo que esta no supere el 10% de la
población**

Asimetría

**La muestra contiene al menos $n \geq 30$ observaciones y la
distribución no es fuertemente sesgada (de lo contrario
aumentar el número de observaciones)**

Si el teorema de límite central se cumple podemos modelar el problema usando la **distribución normal** y determinar la **probabilidad usando el z-score**

Theorem 7-2:
The Central Limit Theorem

If X_1, X_2, \dots, X_n is a random sample of size n taken from a population (either finite or infinite) with mean μ and finite variance σ^2 , and if \bar{X} is the sample mean, the limiting form of the distribution of

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad (7-6)$$

as $n \rightarrow \infty$, is the standard normal distribution.

An electronics company manufactures resistors that have a mean resistance of 100 ohms and a standard deviation of 10 ohms. The distribution of resistance is normal. Find the probability that a random sample of $n = 25$ resistors will have an average resistance less than 95 ohms.

Note that the sampling distribution of \bar{X} is normal, with mean $\mu_{\bar{X}} = 100$ ohms and a standard deviation of

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{25}} = 2$$

Therefore, the desired probability corresponds to the shaded area in Fig. 7-7. Standardizing the point $\bar{X} = 95$ in Fig. 7-7, we find that

$$z = \frac{95 - 100}{2} = -2.5$$

and therefore,

$$\begin{aligned} P(\bar{X} < 95) &= P(Z < -2.5) \\ &= 0.0062 \end{aligned}$$

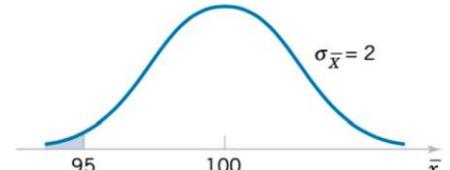


Figure 7-7 Probability for Example 7-13.

Intervalos de confianza

La estimación de un parámetro poblacional a partir de una estimación puntual contiene errores inherentes debido al procesos de muestreo, siendo estos dependientes del tamaño de la muestra

Los **intervalos de confianza** nos permite entregar un **rango de valores** donde estimamos, con un cierto **nivel de confianza**, que el parámetro poblacional se encuentre en él

El error estandar mide la incertidumbre asociado a nuestra **estimación puntual**, también nos provee de una guía para estimar el **intervalo de confianza**



Estimación Puntual



Intervalo de confianza

Si el intervalo se estira **2 SE** de nuestra estimación puntual, entonces podemos decir que estamos **95% seguros** que capturamos el parámetro poblacional

$$Estimacion\ Puntual \pm 2 * SE \quad \text{Margen de error}$$

¿Pero qué significa 95% de confianza?

	mean_precio	se_precio	ci_inf	ci_sup	contiene_real
0	178121.44	9435.994067	159249.451867	196993.428133	True
1	172865.68	9546.162364	153773.355273	191958.004727	True
2	177546.38	9345.630549	158855.118903	196237.641097	True
3	197234.34	11425.212748	174383.914504	220084.765496	True
4	174084.96	10794.213276	152496.533447	195673.386553	True
5	187697.96	10001.897471	167694.165059	207701.754941	True
6	174477.2	13838.084491	146801.031017	202153.368983	True
7	185118.34	11670.466497	161777.407005	208459.272995	True
8	182421.32	13073.091137	156275.137727	208567.502273	True
9	180597.66	10912.886806	158771.886388	202423.433612	True



```
1 muestral_ci['contiene_real'].astype(int).mean()  
2 np.float64(0.948)
```

El **95%** de intervalos de confianza
computados **contienen** el
promedio poblacional

¿Qué es una incorrecta interpretación del intervalo de confianza?

Existe un 95% de probabilidad que el promedio poblacional se encuentre en el intervalo

Falso: El nivel de confianza no refiere a la probabilidad de encontrar el parámetro real en el intervalo (está o no está!), sino habla del método, si repito infinitas veces el muestreo tengo la confianza de encontrar el parámetro poblacional el 95% de esas veces

95% de los datos caen dentro del intervalo de confianza

Falso: El intervalo de confianza esta enfocado en capturar el **parámetro real de la población**, no en observaciones individuales

¿Qué es una incorrecta interpretación del intervalo de confianza?

El intervalo de confianza captura todo el error asociado al experimento

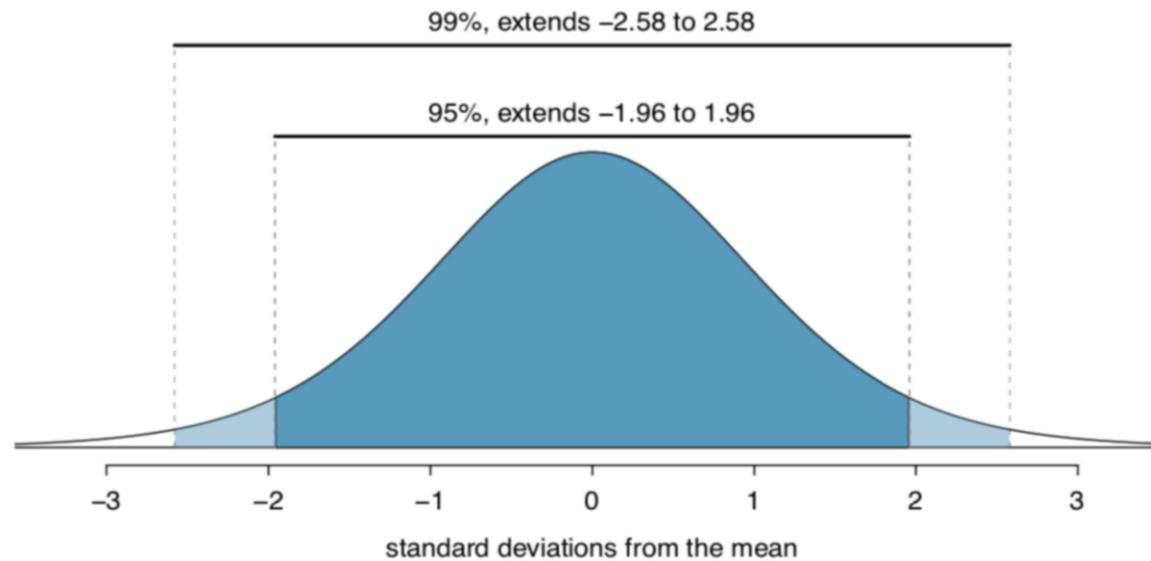
Falso: Existen otras fuentes de error que pueden incluirse en el análisis estadístico, por ejemplo, errores en el diseño experimental o en el muestreo.

Si aplicamos el **teorema del límite central**, podemos analizar la **distribución normal** bajo la estadística de distribución **normal**

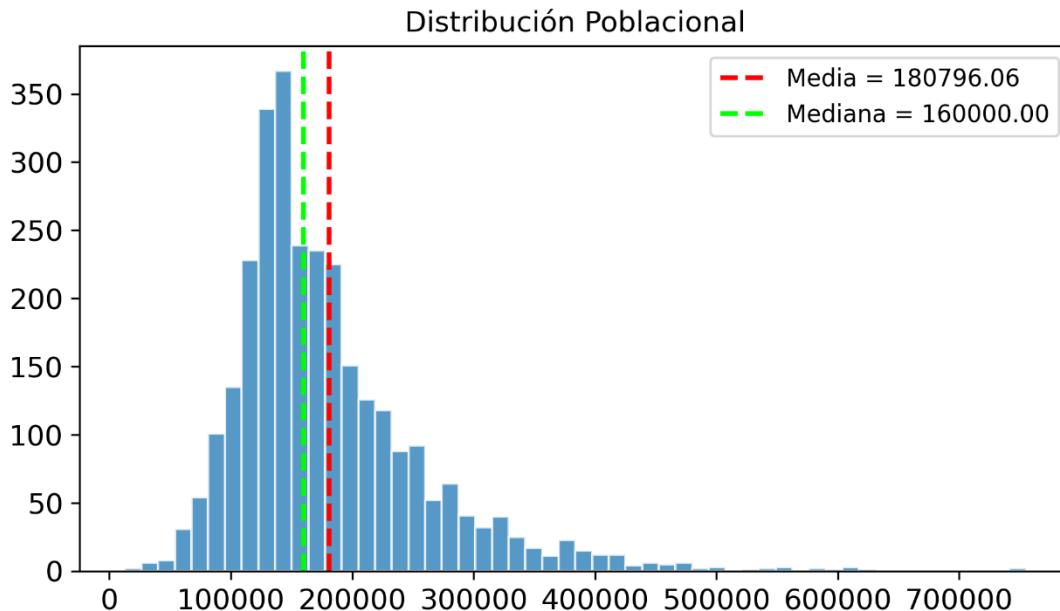
```
1 z_critico = st.norm.ppf(0.975) # 1 - alpha/2  
2      1.9599
```

Estimacion Puntual $\pm 1.96 * SE$ **95% de confianza**

¿Si aumento mi nivel de confianza mi intervalo aumenta o disminuye?



Volviendo a nuestro ejemplo del precio de las casas y aplicamos el teorema del límite central para el cálculo del intervalo de confianza...



n = 10
...¿asimetría?



```
1 muestral_ci['contiene_real'].astype(int).mean()  
2 np.float64(0.89)
```

n = 200



```
1 muestral_ci['contiene_real'].astype(int).mean()  
2 np.float64(0.97)
```

¿Cómo determino el tamaño de muestra
para lograr un **95% de confianza**?

Podemos usar la formula de z-score, y tomar ciertos supuestos, por ejemplo, el **error tolerable** y la **variabilidad de la población**

Definition

If \bar{x} is used as an estimate of μ , we can be $100(1 - \alpha)\%$ confident that the error $|\bar{x} - \mu|$ will not exceed a specified amount E when the sample size is

$$n = \left(\frac{z_{\alpha/2} \sigma}{E} \right)^2 \quad (8-8)$$

Supuestos

Dado que no tienes la variabilidad de tu población tomas una muestra de 20 datos y la desviación estándar es tu mejor aproximación. Luego, buscas que el error de tu promedio de muestra no sea superior a \$20.000 dólares.

$$n = \left(\frac{Z_{\alpha/2} \sigma}{E} \right)^2 = \left(\frac{1.96 * 67,800}{20,000} \right)^2$$

$$n = 44.148 \approx 45$$

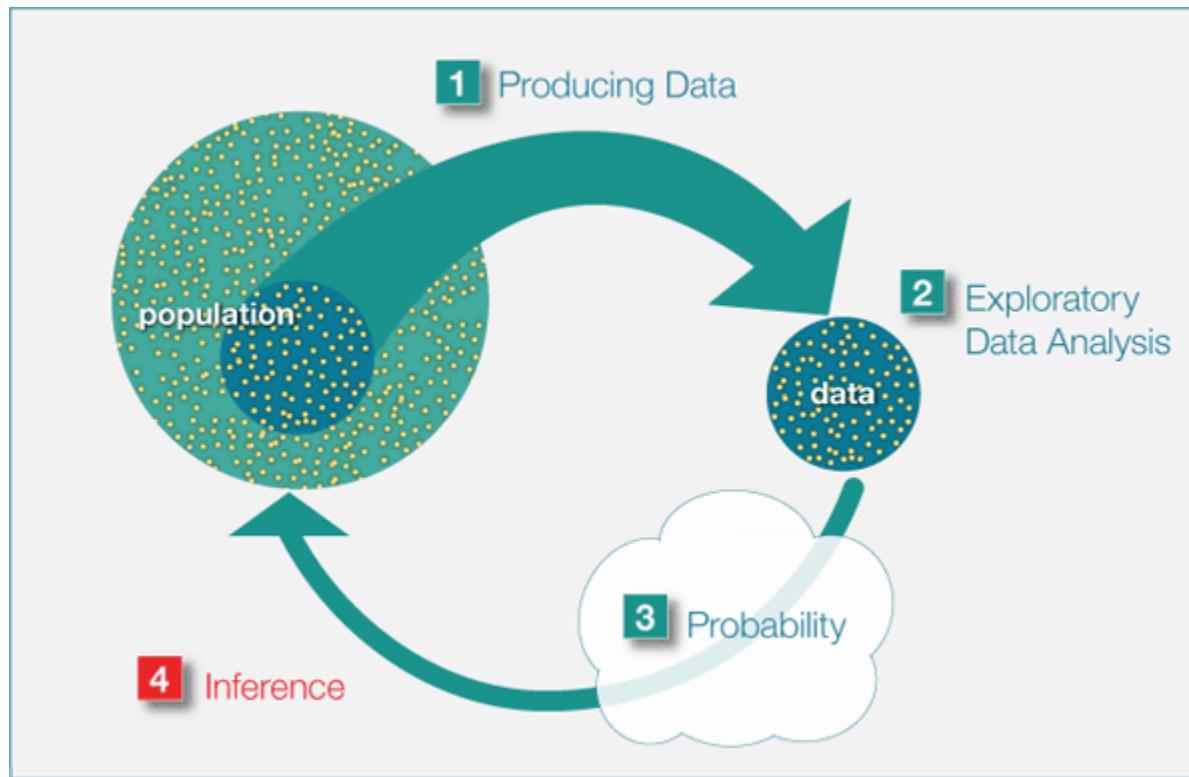
$$E = \bar{x} - \mu = \$20.000$$

$$\sigma = \$67.800$$

Resumen

- La inferencia estadística se ocupa de plantear y verificar hipótesis respecto de los datos.
- El objetivo es inferir parámetros poblacionales a partir de muestras finitas.
- La incertidumbre de la inferencia se puede describir por el intervalo de confianza, estableciendo una relación directa con tests de hipótesis.
- Para la construcción del intervalo de confianza debemos determinar el error estándar que no es lo mismo que la desviación estándar.

Inferencia Estadística



Profesor: Pedro Saa (pnsaa@uc.cl)
Año: 1-2025