Inferencia

Inferencia basada en diseño, e inferencia basada en modelo

PSI2301

Métodos de Investigación en Ciencias Sociales

Abril 11

2023

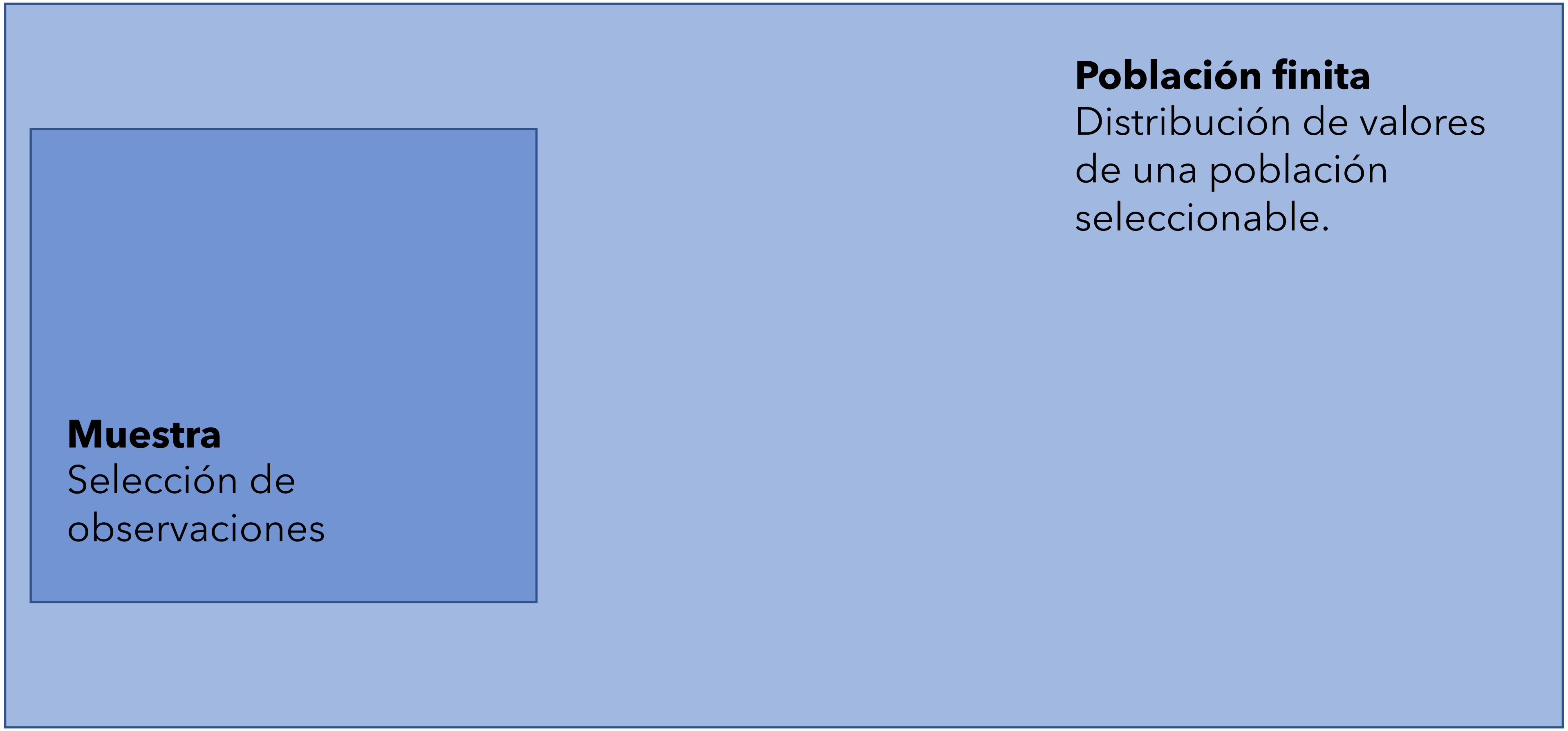
Carrasco, D., PhD,
Centro de Medición MIDE UC
Pontificia Universidad Católica de Chile

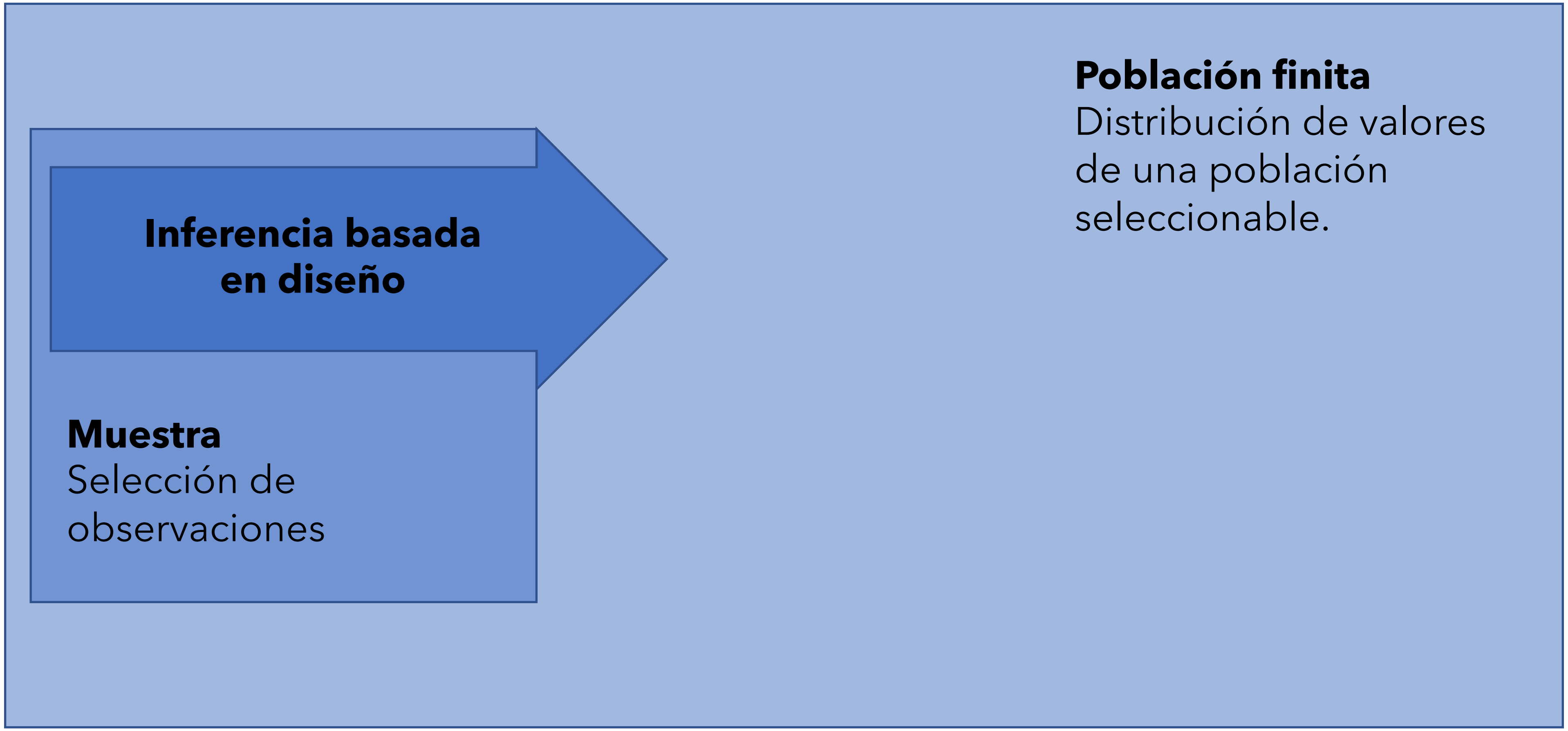
Centro UC
Medición - MIDE

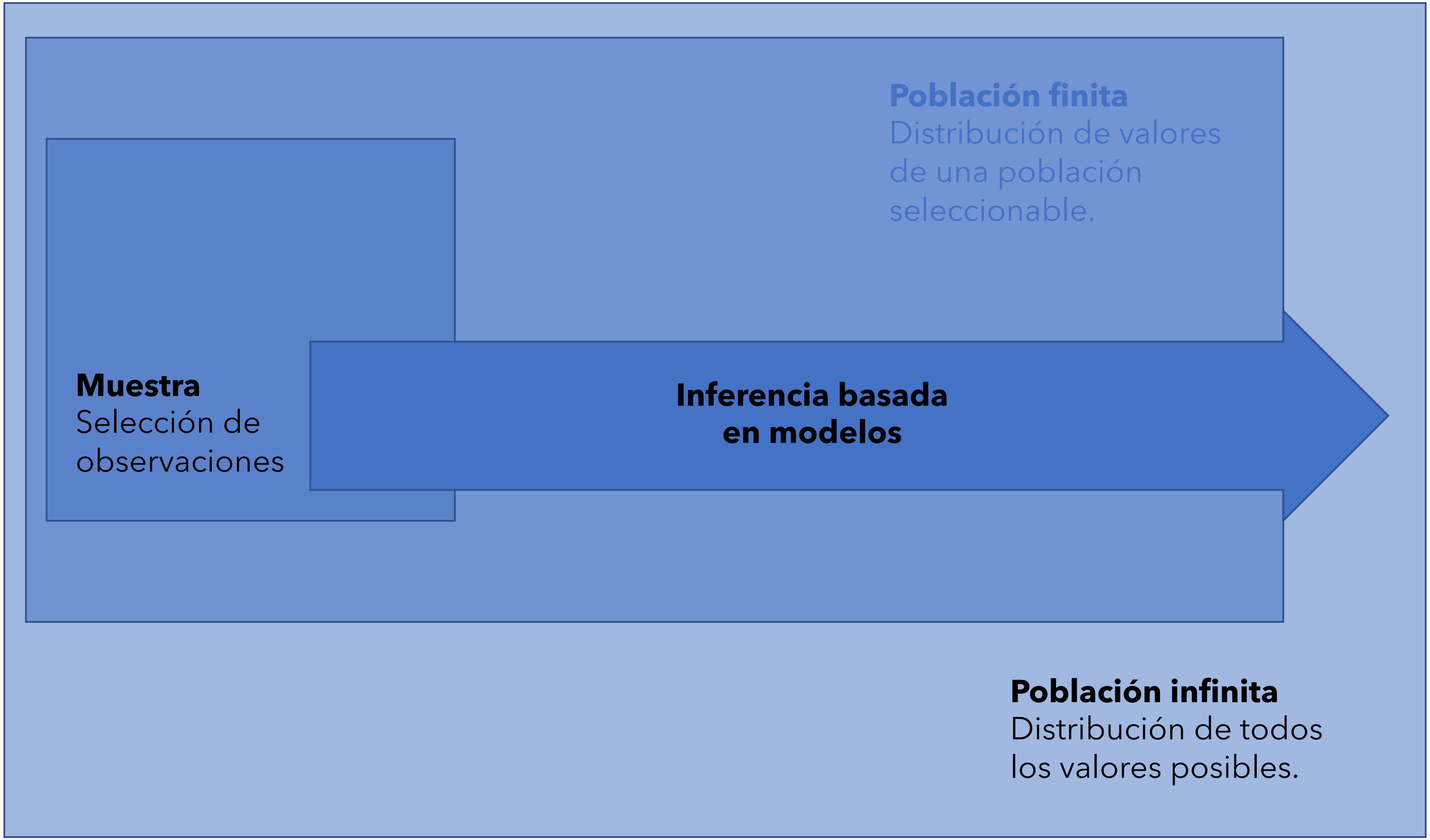
Metodología Cuantitativa

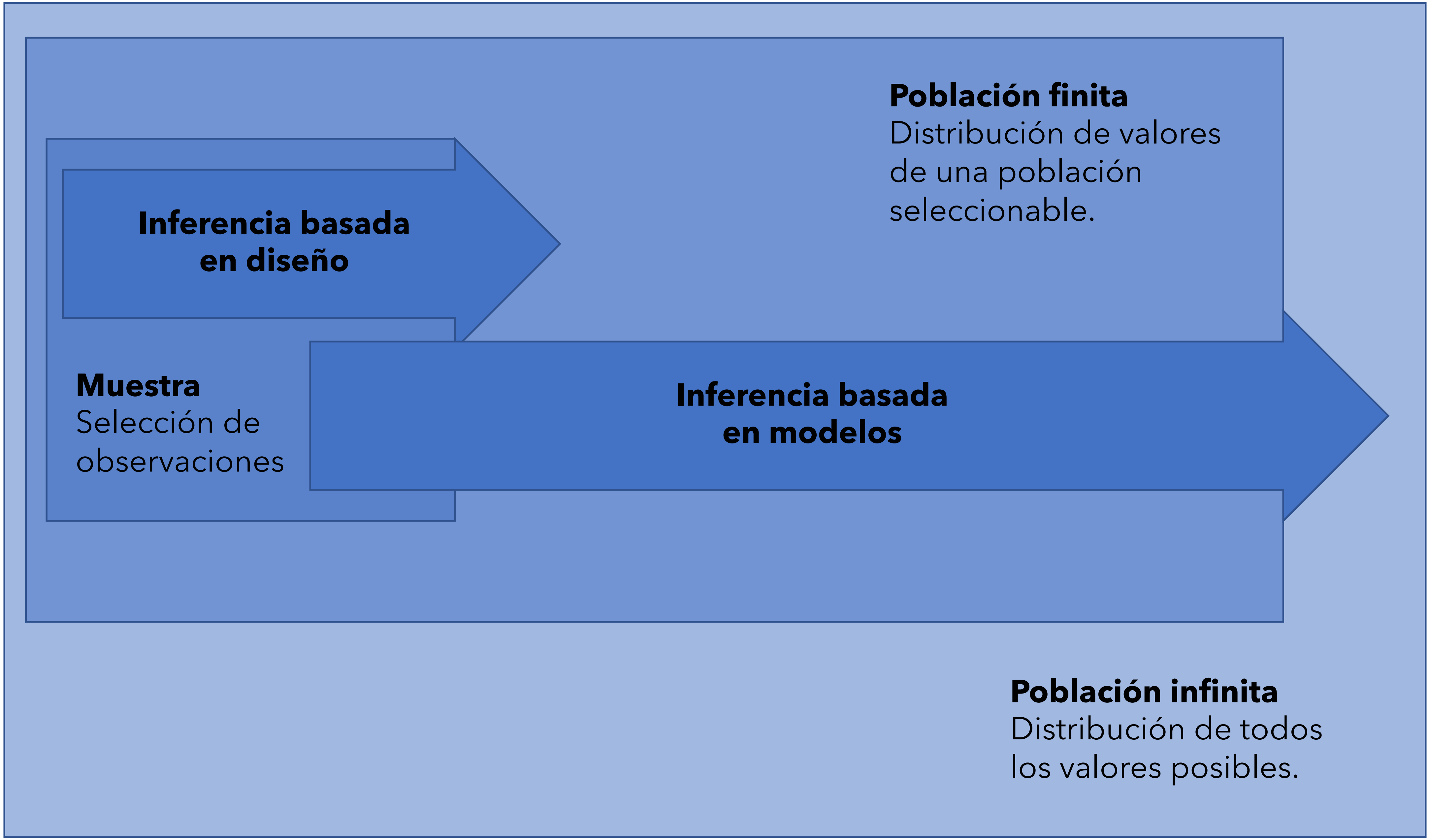
Inferencias

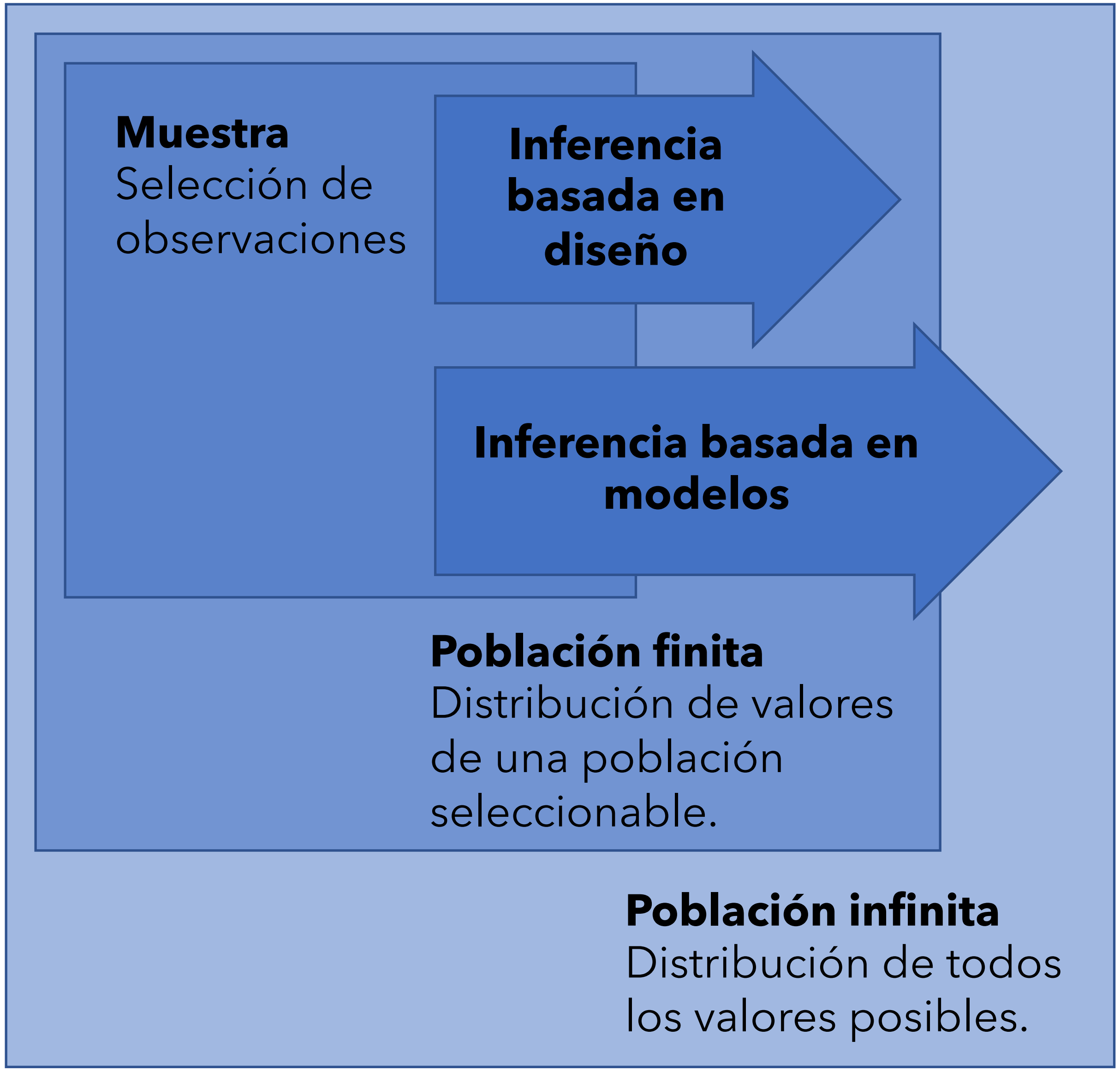
Dos tipos de inferencia: basada en diseño, y basada en modelo


Población finita
Distribución de valores de una población seleccionable.
Muestra
Selección de observaciones


Población finita
Distribución de valores de una población seleccionable.
Inferencia basada en diseño
Muestra
Selección de observaciones


Población finita
Distribución de valores de una población seleccionable.
Muestra
Selección de observaciones
Inferencia basada en modelos
Población infinita
Distribución de todos los valores posibles.


Población finita
Distribución de valores de una población seleccionable.
Inferencia basada en diseño
Muestra
Selección de observaciones
Inferencia basada en modelos
Población infinita
Distribución de todos los valores posibles.



Nota: Tambien es posible realizar inferencias hibridas, basadas en modelos y en diseño. Este tipo de inferencias excede los contenidos del curso. Este es el punto central del atículo de Sterba (2009).

Dos formas de inferencia

En clases anteriores, hemos visto el ejercicio de realizar **inferencias basadas en la idea de obtener muestras, respecto a una población finita.** Con estas muestras, queremos estimar al parámetro poblacional, por medio del cálculo de estadígrafos. Debido a **cómo se comporta una distribución muestral de un estadígrafo**, podemos estimar al parámetro poblacional, con el estadígrafo.

Existe una forma adicional de realizar inferencias, llamada **inferencia basada en modelos**. En este caso, hacemos conclusiones no respecto a una población finita, sino a una **población infinita**; respecto a la distribución total de valores posibles un parámetro (Sterba, 2009). En este caso, la muestra observada, y la población finita serían solo realizaciones de un **mecanismo generador de datos** (Rabe-Hesketh et al., 2012)

Metodología Cuantitativa

Inferencia basada en diseño

Distribución muestral de la media

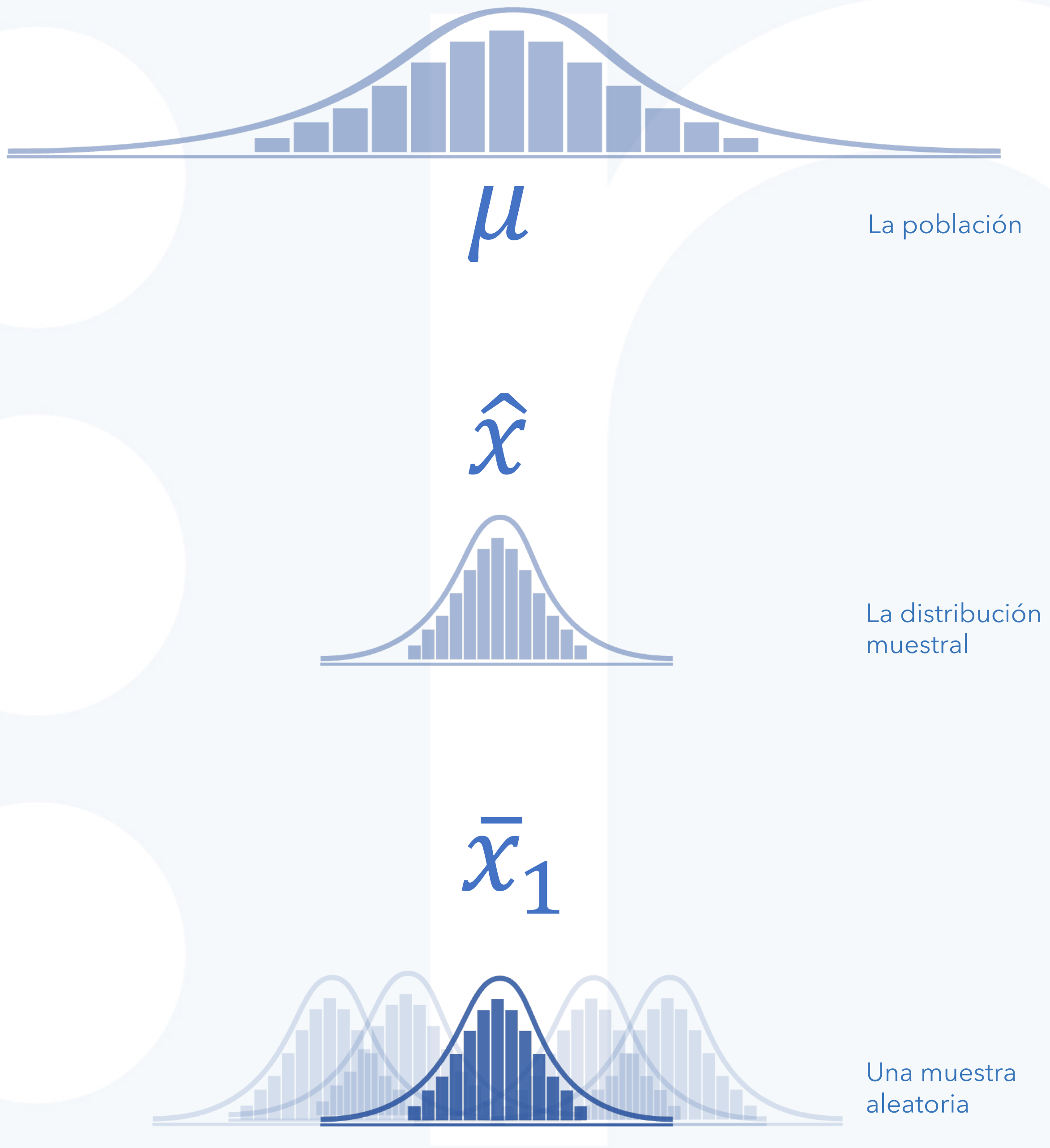
Inferencia basada en diseño

En inferencia basada en diseño, nuestro foco de inferencia son los **parámetros de una población finita**. Es decir, un conjunto de observaciones seleccionables (e.g., todos los estudiantes de sexto grado de Chile).

Los parámetros de interés de una población pueden ser la media, una proporción, un total, un ratio, u otra cifra, como la varianza, y la desviación estándar. En la población, **el parámetro se simboliza con μ** (la letra griega "mu"). En la clase de hoy, vamos a partir con la idea de una media.

Para poder realizar inferencias basadas en diseño, **descansamos en la idea de la distribución muestral del estadígrafo $\hat{x}$**. Este es el promedio de todas las medias de las muestras aleatorias posibles.

La distribución de medias, de todas las muestras posibles toma una forma particular, es **una distribución normal**.



Distribución Muestral

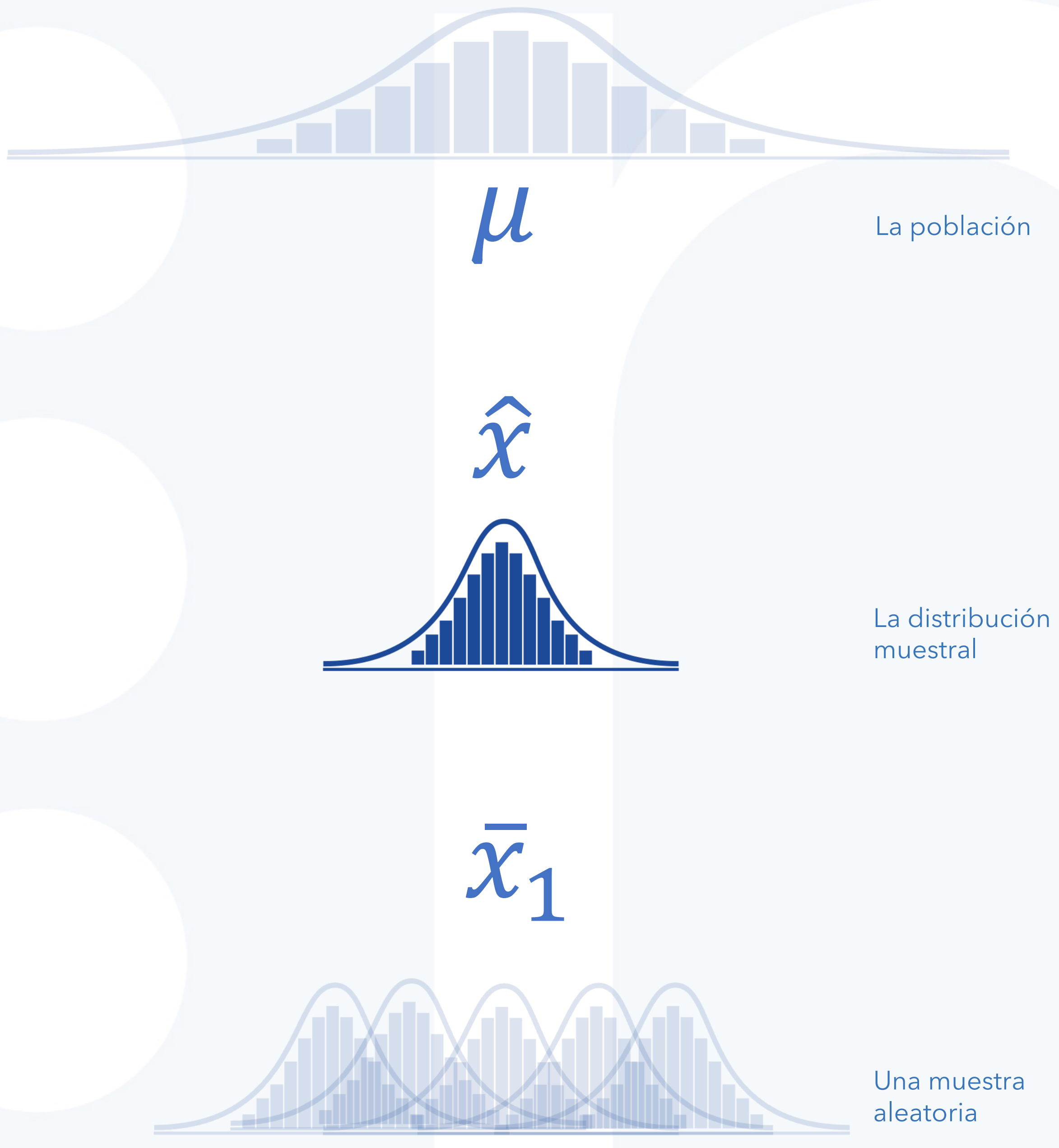
La distribución muestral de la media, es el **promedio de todas las medias, de cada muestra**.

Podemos aproximarnos a esta distribución, por medio de generar una serie de muestras aleatorias simples (con reemplazo). Es decir, un conjunto de muestras que poseen la misma probabilidad de ser seleccionadas. Calcular sus medias, y luego calcular la media de todas las muestras.

Esta distribución, **siempre toma una forma normal**. Esta afirmación, a veces es formalizada con la siguiente expresión (Devlin et al., 2018):

$$\hat{x} \xrightarrow{iid} N(\mu, \frac{\sigma^2}{n})$$

Nota: Esta notación indica que el promedio de las medias de muchas muestras (aletorias simples), toma como media al parámetro de la población, y posee una varianza conocida, y una forma normal.



Teorema del limite central

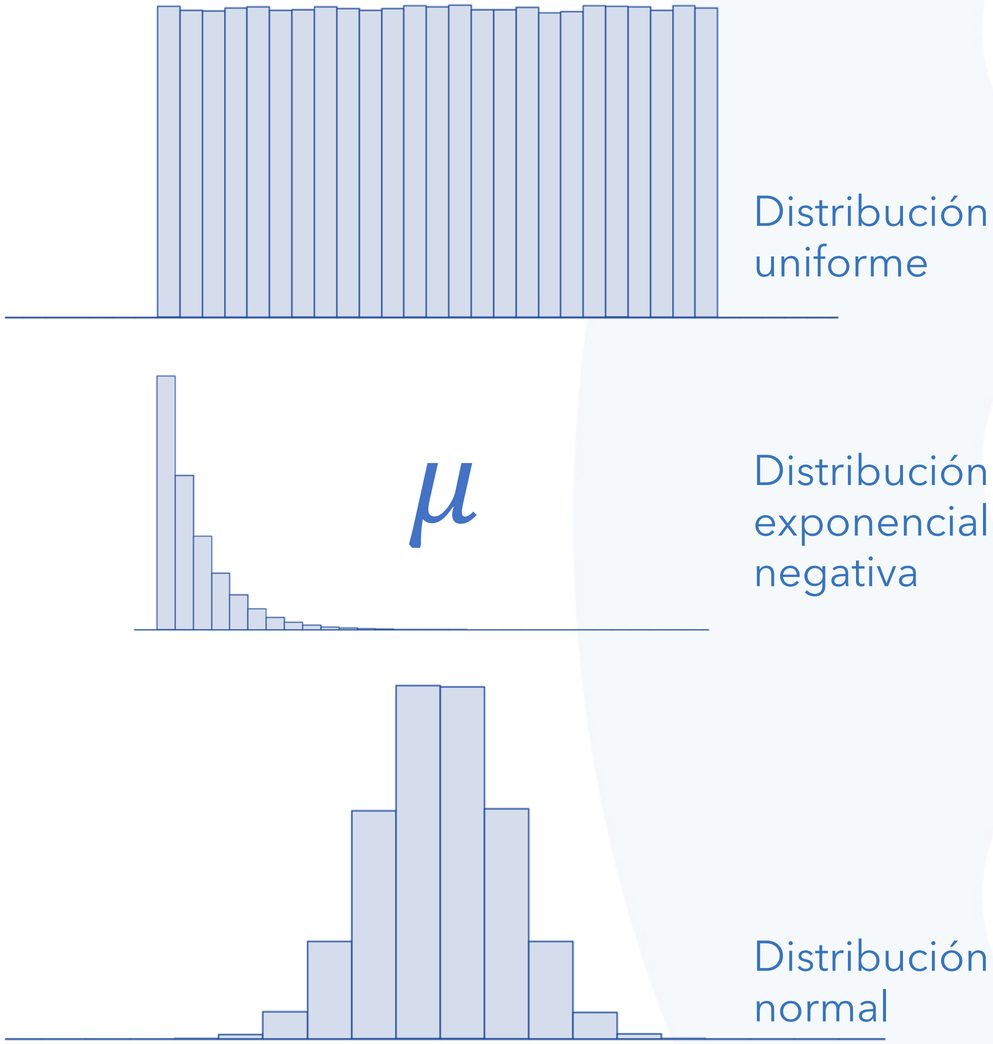
En términos verbales, la idea de que la distribución muestral de la media, tome una forma normal, es conocida como el **teorema del límite central**.

A medida de que el **tamaño muestral crece, la distribución de medias**, de todas las muestras posibles, **toma una forma normal** (Huck, 2008), sin importar la forma original de la distribución de la población (Devlin et al., 2018).

No obstante, un aspecto a considerar es que la **varianza** de la distribución muestral de la media **disminuye** en la medida de que el **tamaño muestral es mayor**. Para el ejercicio de realizar inferencias con una sola muestra aleatoria, esto es importante. Porque implica que, muestras de tamaños más grandes (e.g., 100, 200, 300, 400 casos), provienen de distribuciones muestrales con menos dispersión, en contraste a muestras de menor tamaño (e.g., 10, 20, 30).

No importa la forma de la distribución de la población, la distribución muestral **nos permite recuperar parámetros de la población**.

Teorema del limite central



No importa la forma de la distribución de la población, la distribución muestral **nos permite recuperar parámetros de la población**.

Dispersión de la distribución muestral

La varianza de la distribución muestral tiene como expresión a sigma cuadrado, partido por el tamaño muestral.

$$\hat{x} \xrightarrow{iid} N(\mu, \frac{\sigma^2}{n})$$

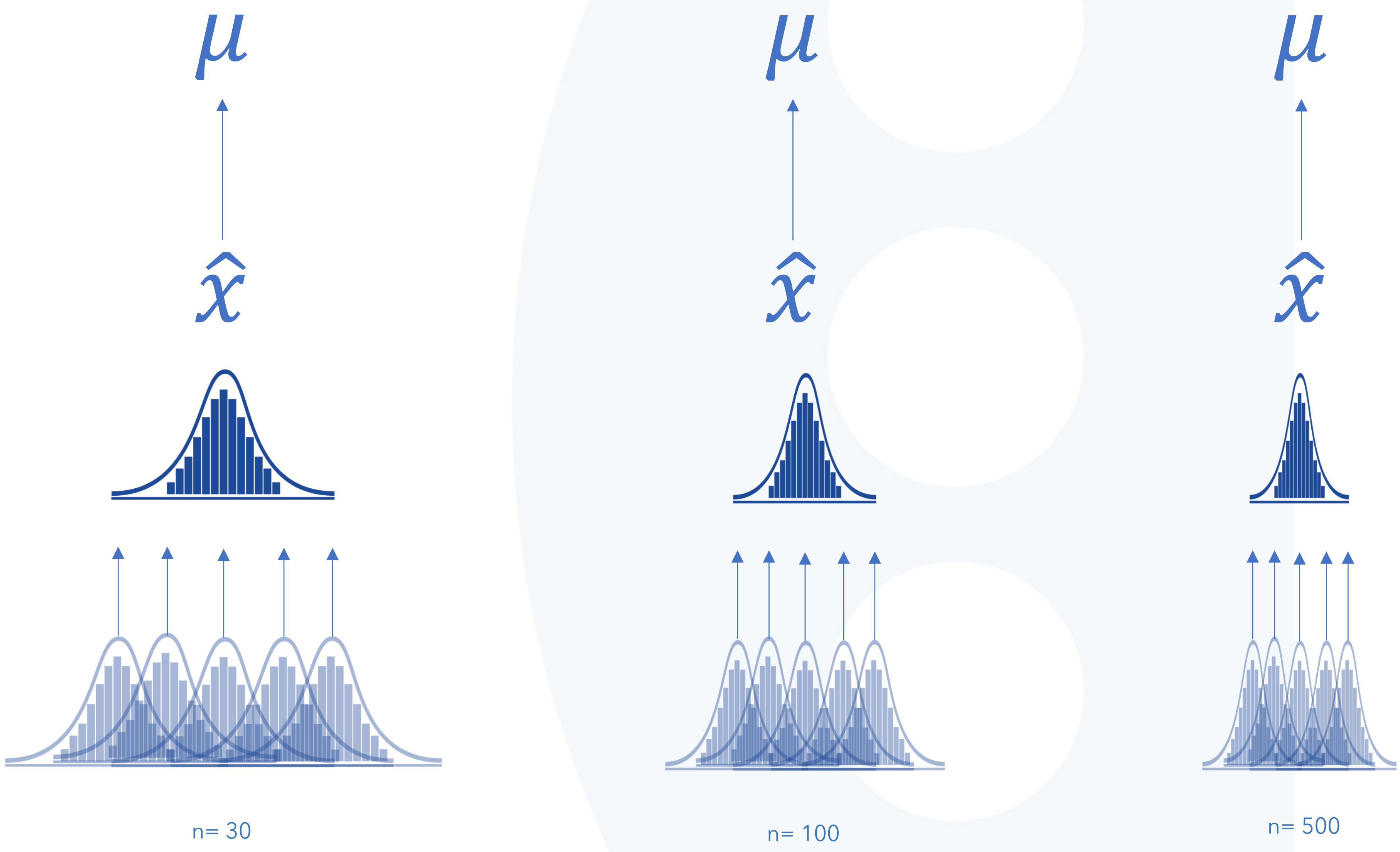
La desviación estándar de la distribución muestral de la media, tiene por expresión a

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

En consecuencia, la precisión de una muestra aleatoria no se incrementa de forma lineal. Dicho de otro modo, si se quiere duplicar la precisión de un muestra aleatoria, **el tamaño muestral podría requerír ser cuadruplicado** (Stigler, 2016).

No importa la forma de la distribución de la población, la distribución muestral **nos permite recuperar parámetros de la población**.

Dispersión de la distribución muestral

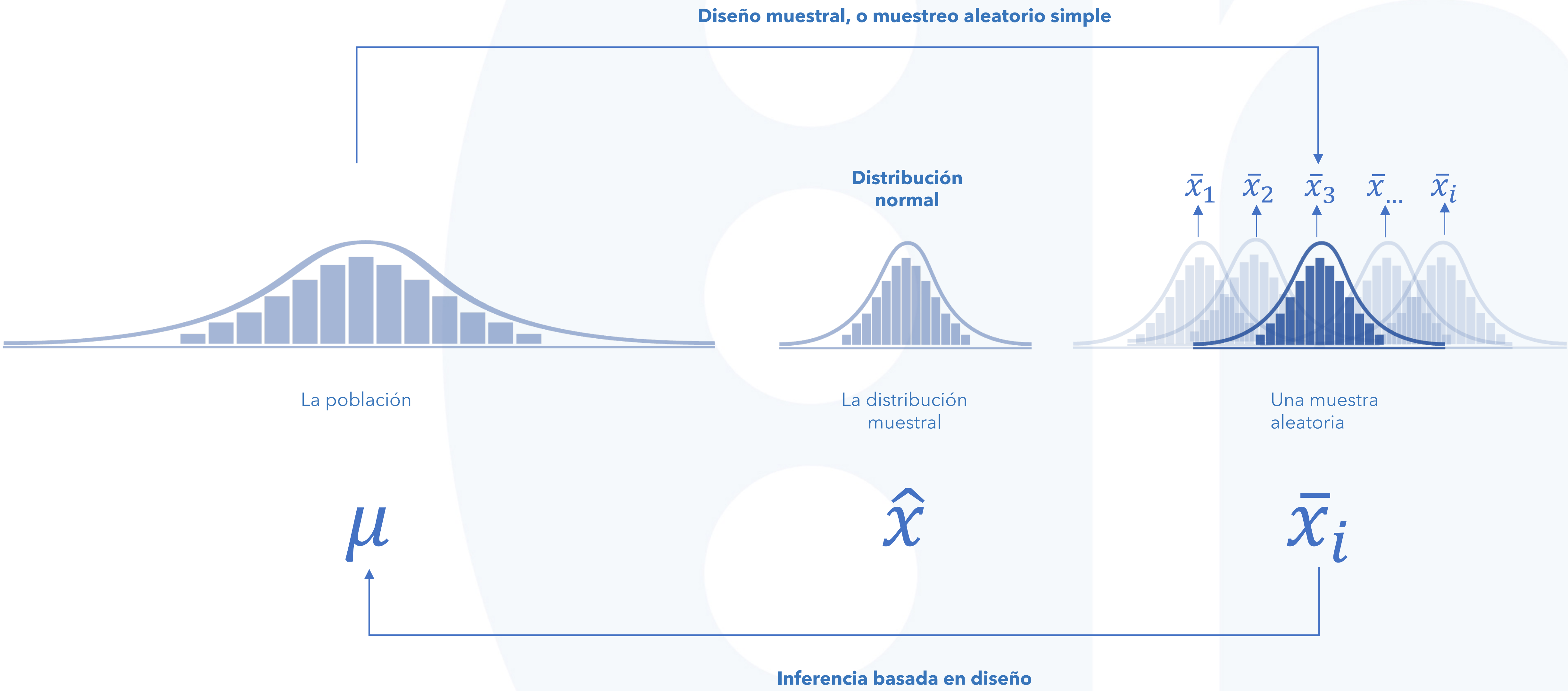


A medida que el tamaño muestral aumenta, la dispersión de la distribución muestral de las medias se hace más estrecha.

En otras palabras, la distribución de medias, de la distribución muestral, es menos dispersa.

En consecuencia se plantea que, muestras aleatorias de mayor tamaño podrían ser más precisas para realizar inferencias a la población.

Inferencia basada en diseño



Generemos una distribución muestral!

- Población
- Distribución muestral
- Varios tamaños
  - 10 a 1000 observaciones

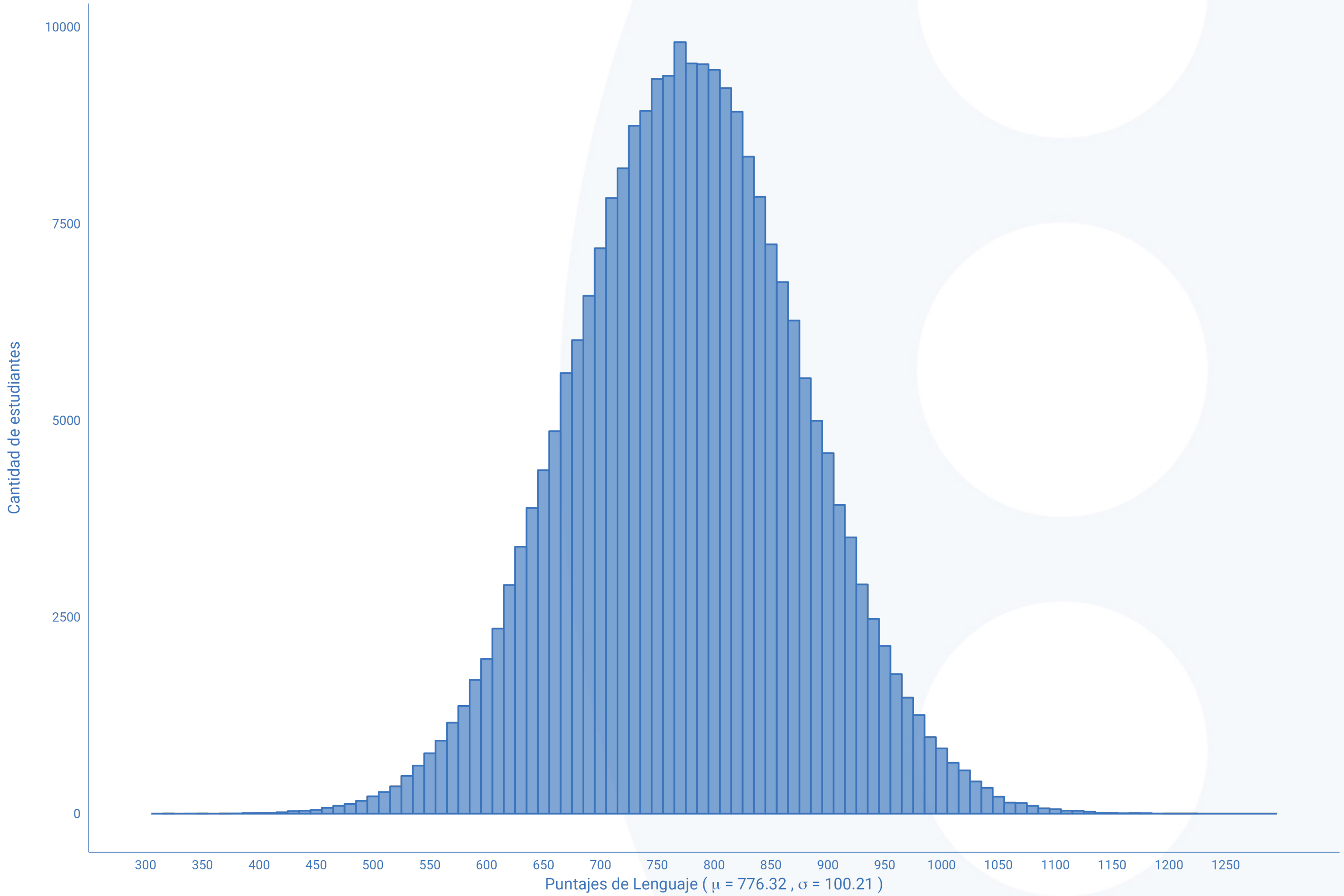
Generación de una distribución muestral

Vamos a crear una población de datos realista.

Vamos a crear una población de 242122, de estudiantes. Esta es la matricula de estudiantes de Chile de sexto grado.

Esta población tendrá una media de 775.320 puntos, y una varianza de 10043.550. Estos son los resultados obtenidos por Chile en 2013, en la prueba TERCE.

Histograma de la población

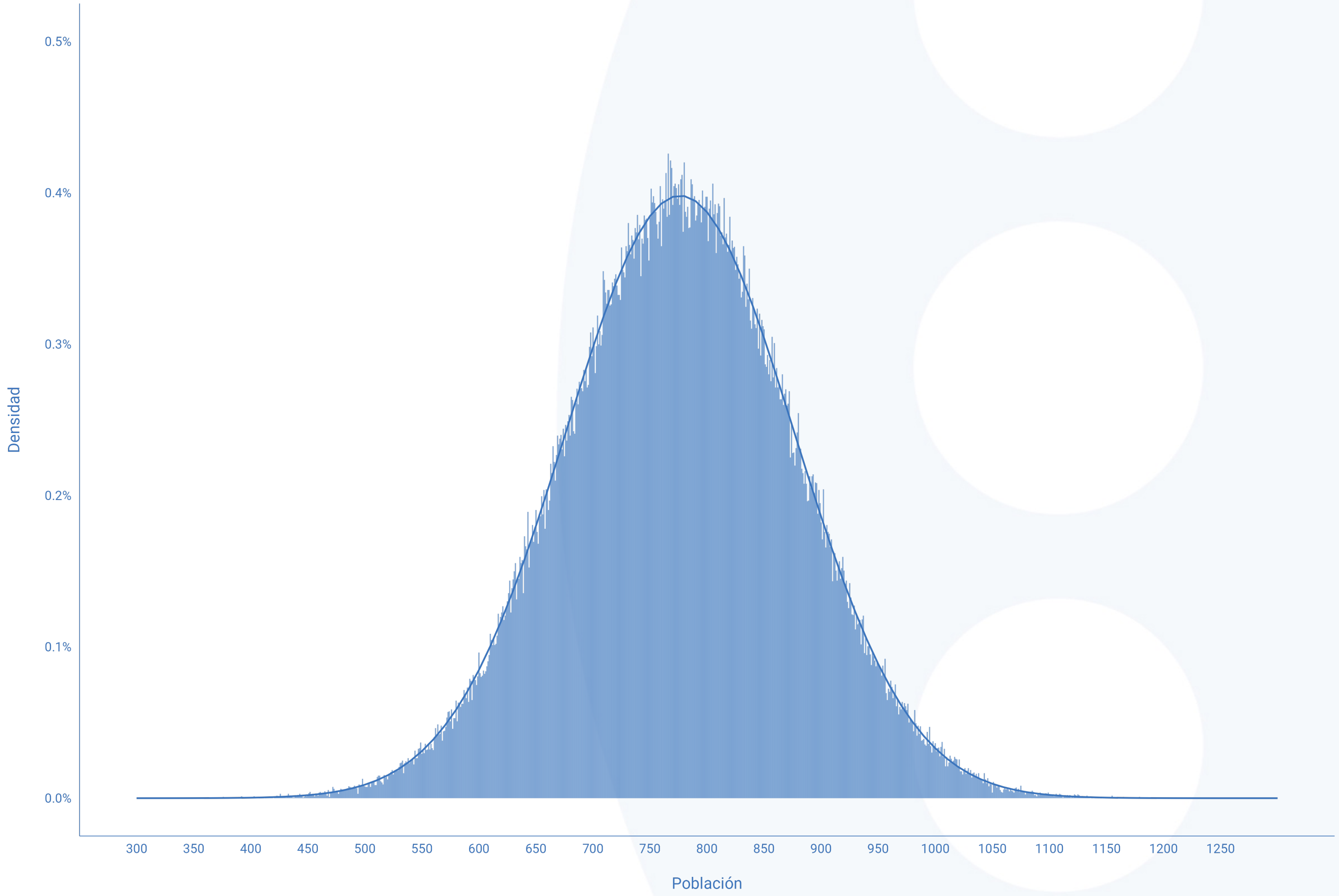


Tenemos una población de valores. Estos son los resultados de los estudiantes de sexto grado de Chile, en una prueba de lenguaje.

Esta es una población finita de 242.122 estudiantes de sexto grado. Presenta una media empírica de 776.32 puntos, y una desviación estándar de 100.21.

Estos puntajes, presentan una media ideal de 700 puntos, y una desviación estándar de 100 puntos. Sobre esta población, vamos a generar muestras de diferentes tamaños.

Histograma de la población como densidad



Re-escalamos al histograma en términos de densidad. En esta escala, la suma de todas las densidades del histograma suman a 1. Esta es una manera de representar a distribuciones de probabilidad.

Para los siguientes gráficos, nos servimos de esta escala, de modo que los histogramas de la distribución muestral, y de las muestras generadas, se encuentren en una escala común.

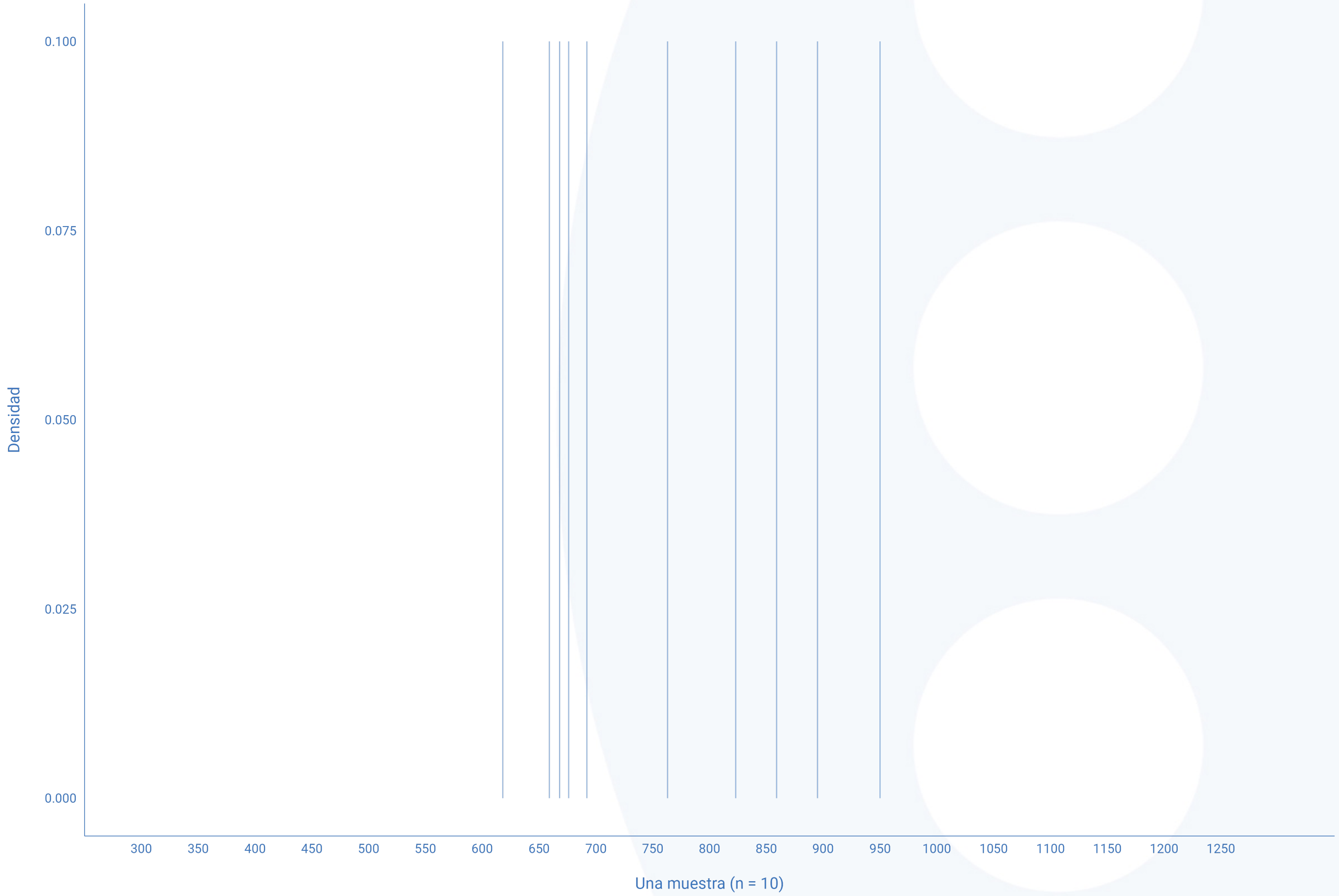
Nota: la población generada es simulada. Se toma el total de estudiantes de sexto grado de 2018 en Chile. Y se generan los puntajes con la media y desviación estándar obtenida por los estudiantes de sexto grado de Chile en TERCE 2013.

Metodología Cuantitativa

Generación de muestras

Colecciones de muestras de diferente tamaño

Una muestra de 10 casos

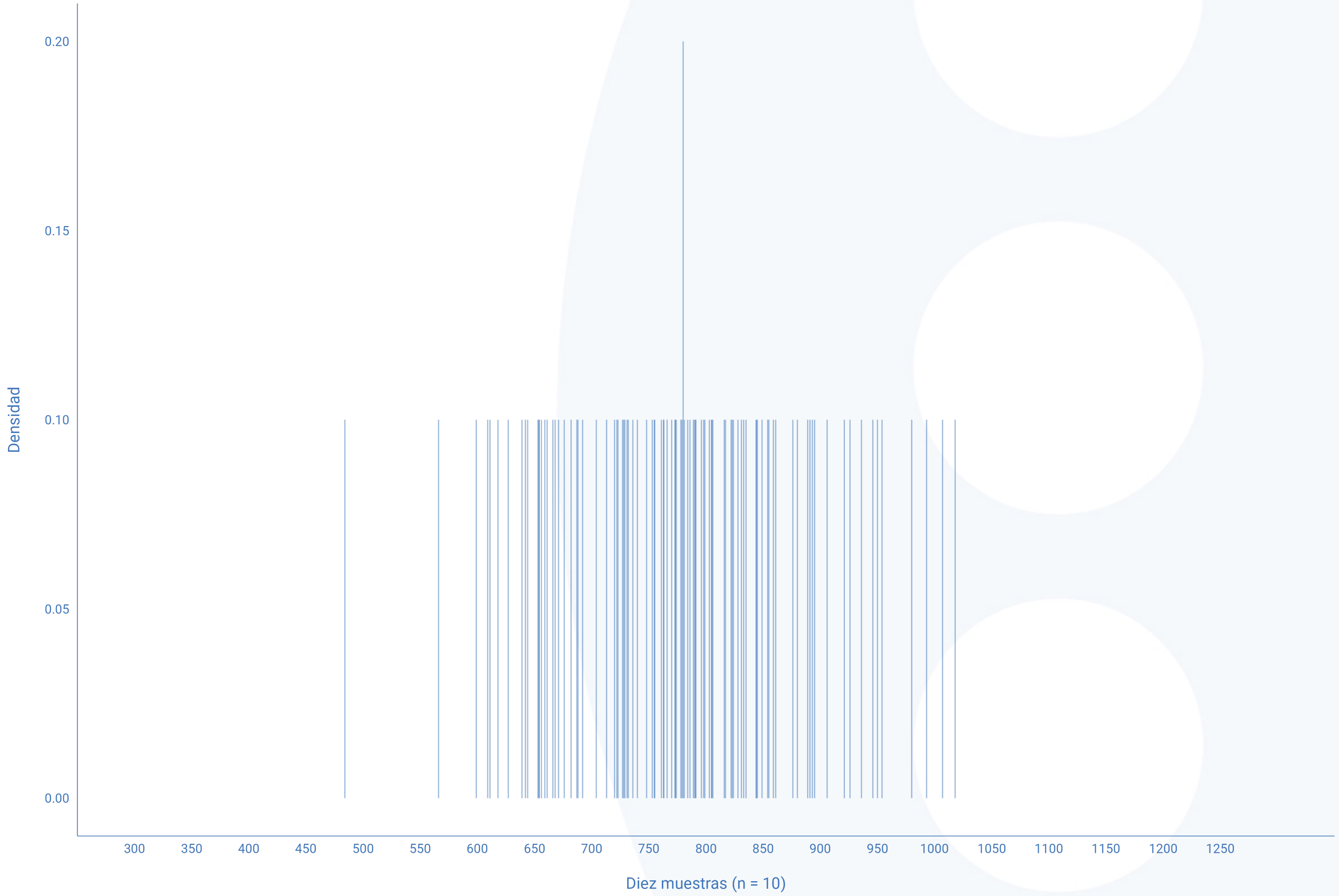


Vamos a ir generando la distribución muestral de medias de a poco.

Partamos por la intuición más básica. Generemos una muestra de casos.

En un histograma de densidad, una sola muestra de diez casos al azar solo serían diez líneas.

Diez muestras de 10 casos

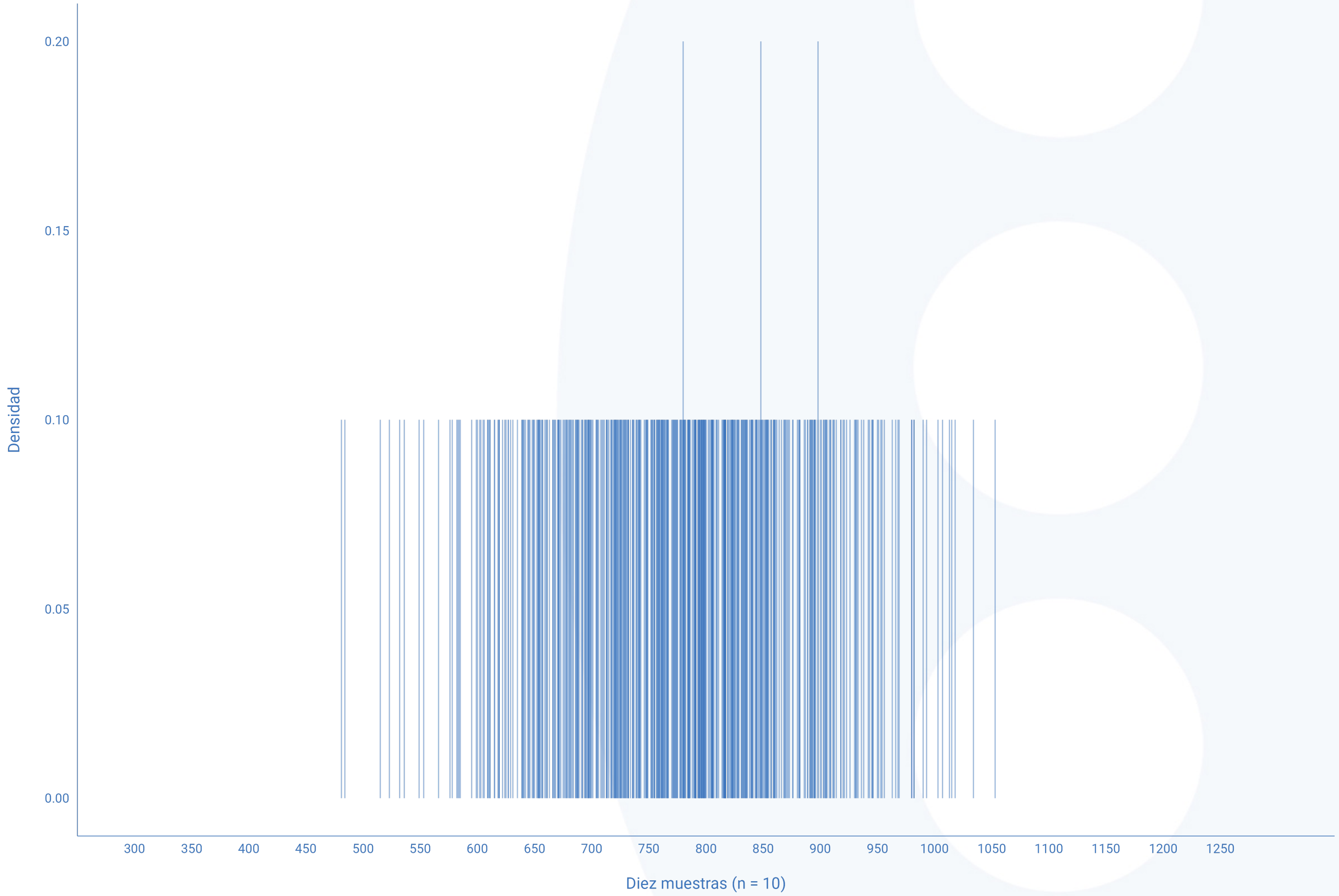


Si aumentamos la cantidad de muestras de 10 casos, a 10 muestras, ahora tendremos más líneas.

Cada una de las líneas, representa un caso, con un valor determinado.

Cuando tenemos más de un caso, con el mismo valor, entonces nuestras líneas crecen hacia arriba.

50 muestras aleatorias de 10 casos



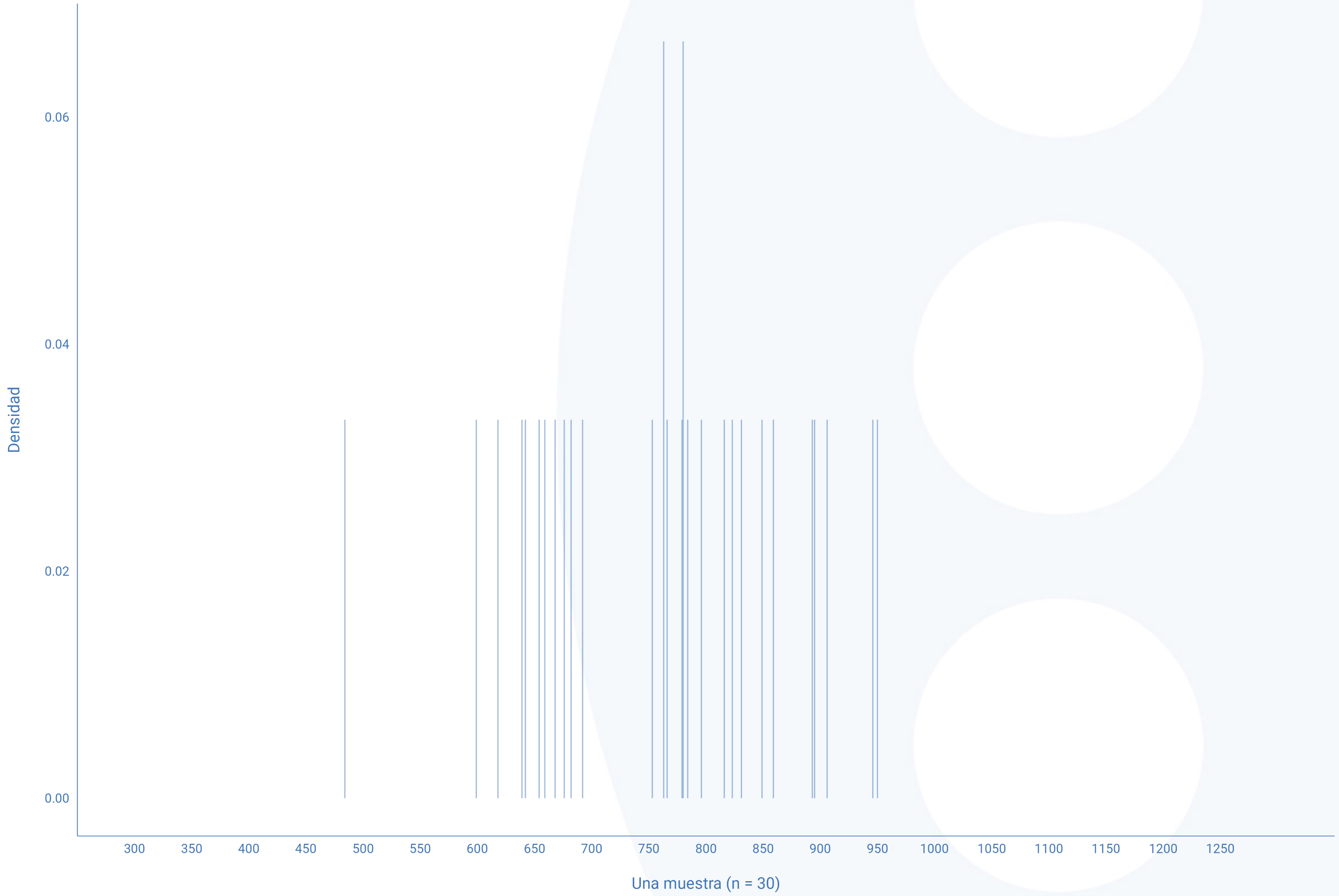
A medida que tenemos más muestras aleatorias, tenemos más chances de que haya mayor densidad entre los valores observados.

Ahora, veamos como se la distribución muestral de las muestras de 10 observaciones.

Para generar esta distribución muestral, lo que vamos a hacer es calcular la media de cada muestra generada.

Vamos a realizar este ejercicio para 500 muestras. **Veamos como se ve una colección de muestras de 30 casos.**

Una muestra aleatorias de 30 casos



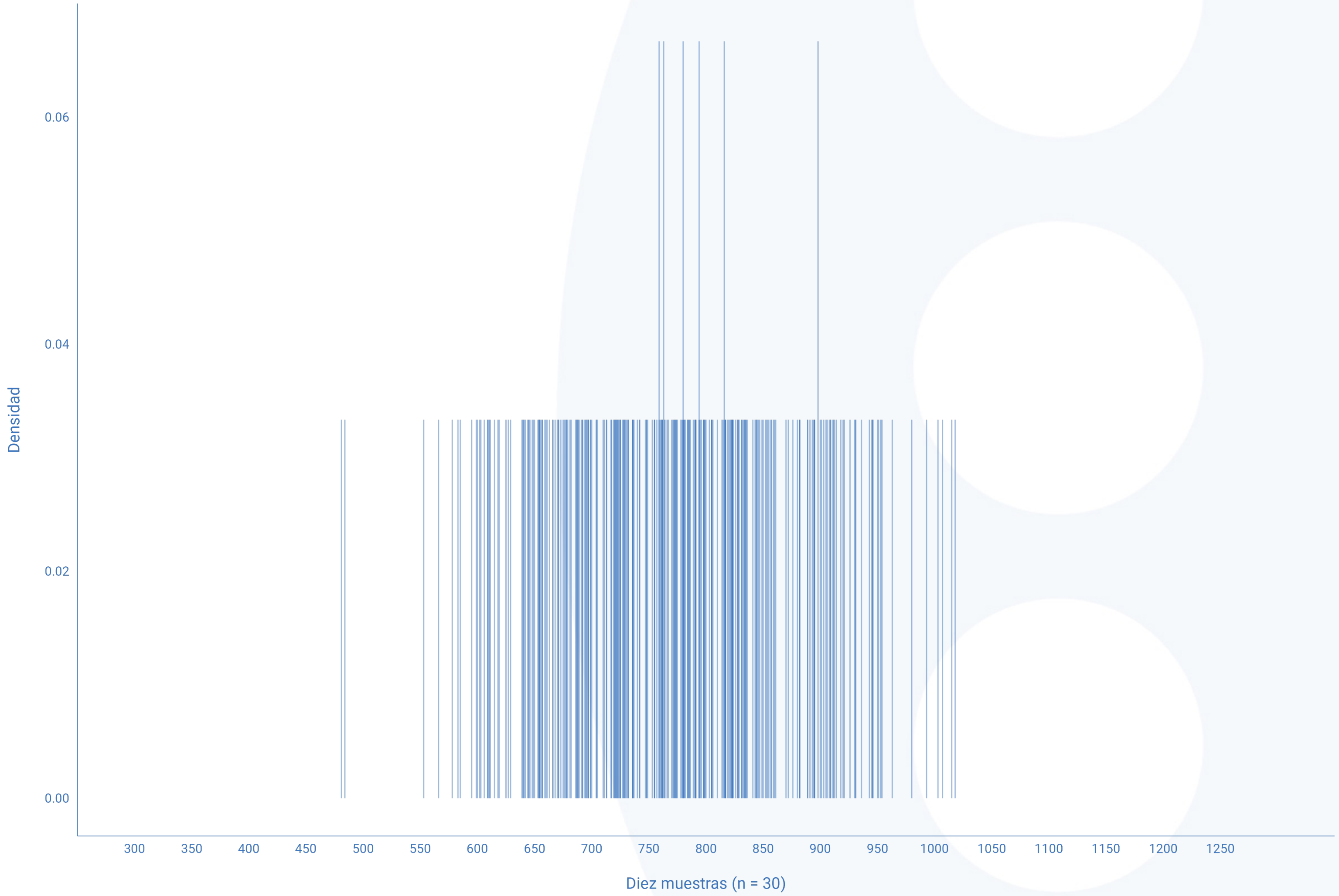
A medida que tenemos más muestras aleatorias, tenemos más chances de que haya mayor densidad entre los valores observados.

Ahora, veamos como se la distribución muestral de las muestras de 10 observaciones.

Para generar esta distribución muestral, lo que vamos a hacer es calcular la media de cada muestra generada.

Vamos a realizar este ejercicio para 500 muestras. **Veamos como se ve una colección de muestras de 30 casos.**

Diez muestras aleatorias de 30 casos



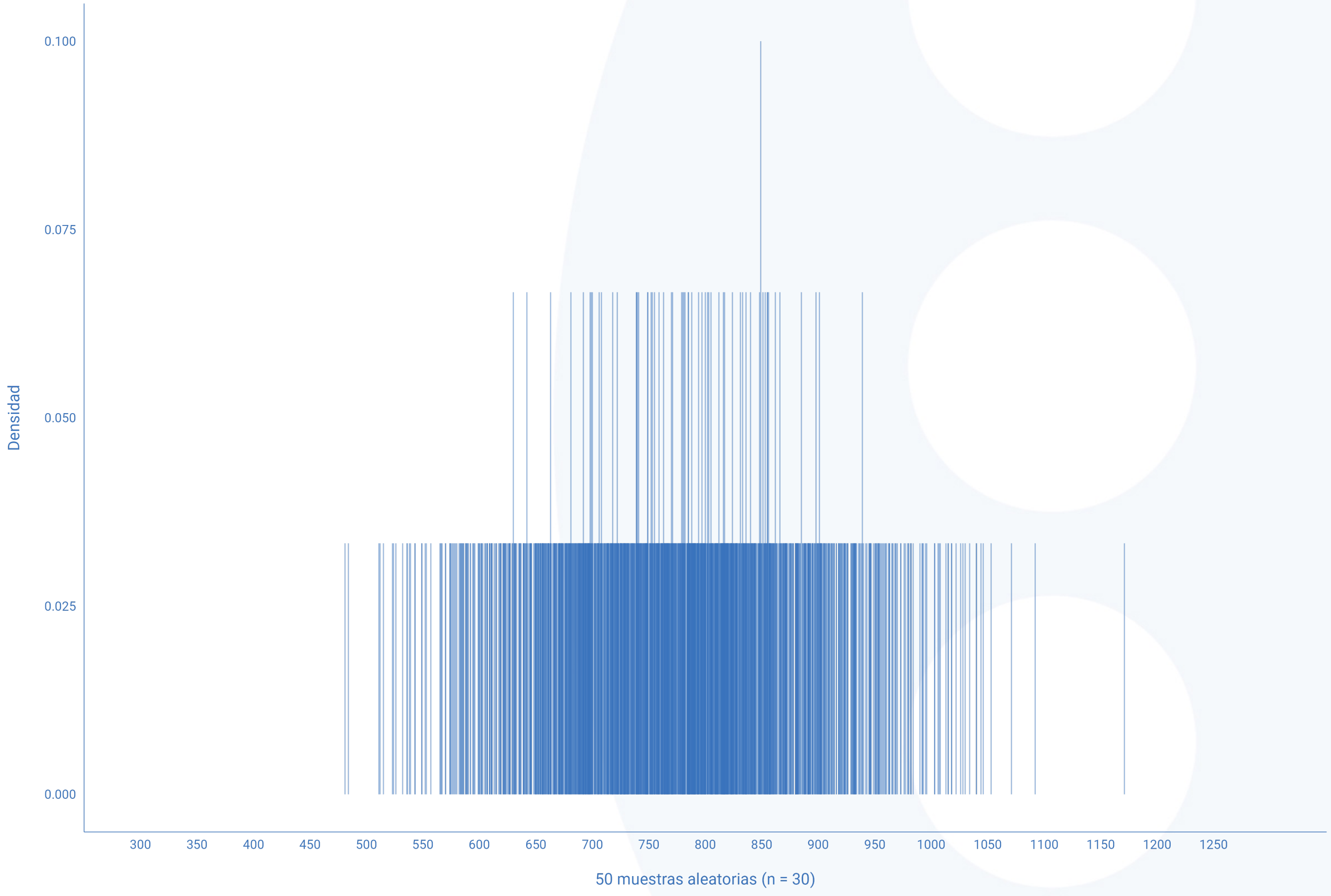
A medida que tenemos más muestras aleatorias, tenemos más chances de que haya mayor densidad entre los valores observados.

Ahora, veamos como se la distribución muestral de las muestras de 10 observaciones.

Para generar esta distribución muestral, lo que vamos a hacer es calcular la media de cada muestra generada.

Vamos a realizar este ejercicio para 500 muestras. **Veamos como se ve una colección de muestras de 30 casos.**

50 muestras aleatorias de 30 casos



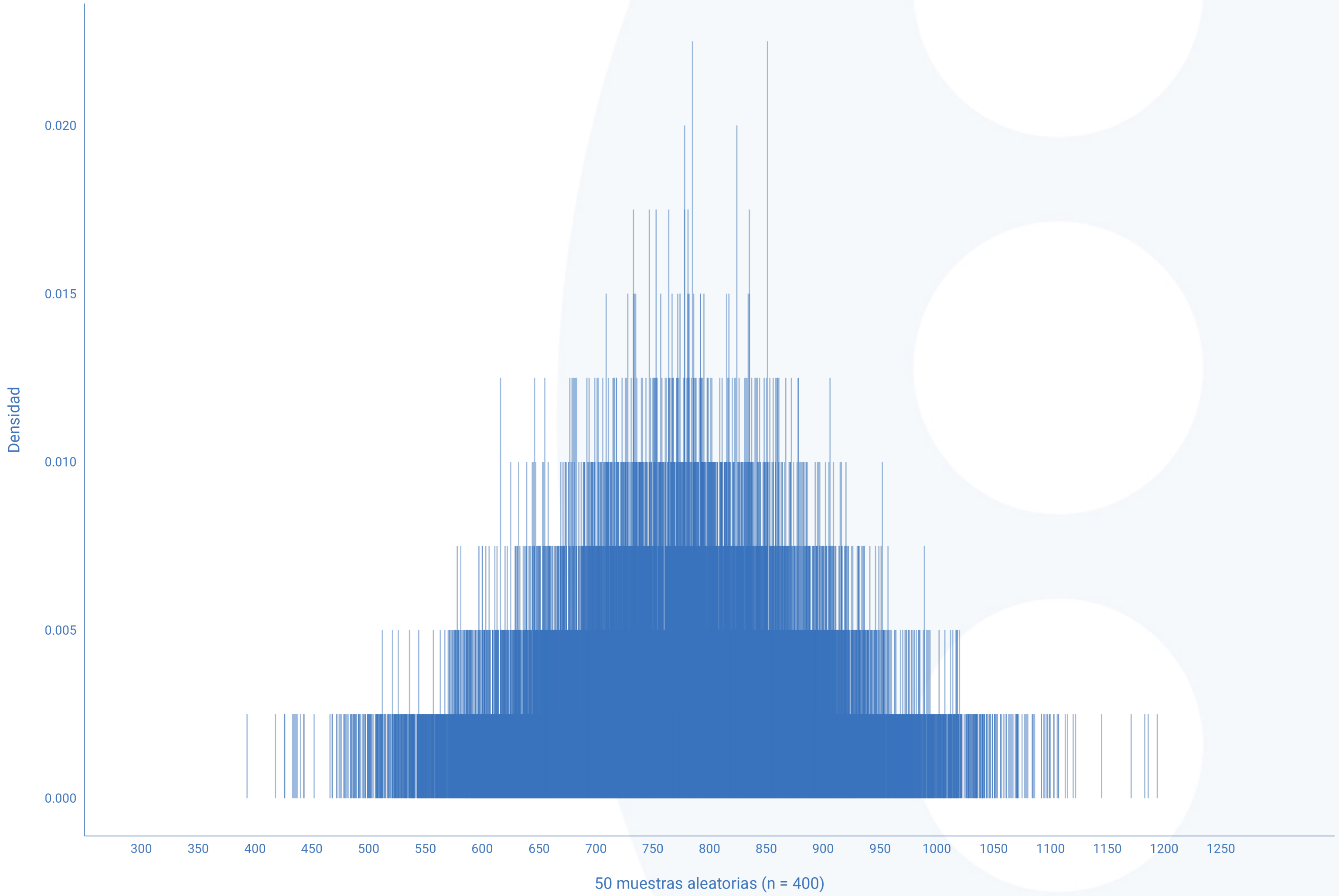
A medida que tenemos más muestras aleatorias, tenemos más chances de que haya mayor densidad entre los valores observados.

Ahora, veamos como se la distribución muestral de las muestras de 10 observaciones.

Para generar esta distribución muestral, lo que vamos a hacer es calcular la media de cada muestra generada.

Vamos a realizar este ejercicio para 500 muestras. **Veamos como se ve una colección de muestras de 30 casos.**

50 muestras aleatorias de 400 casos



A medida que tenemos más muestras aleatorias, tenemos más chances de que haya mayor densidad entre los valores observados.

Ahora, veamos como se la distribución muestral de las muestras de 10 observaciones.

Para generar esta distribución muestral, lo que vamos a hacer es calcular la media de cada muestra generada.

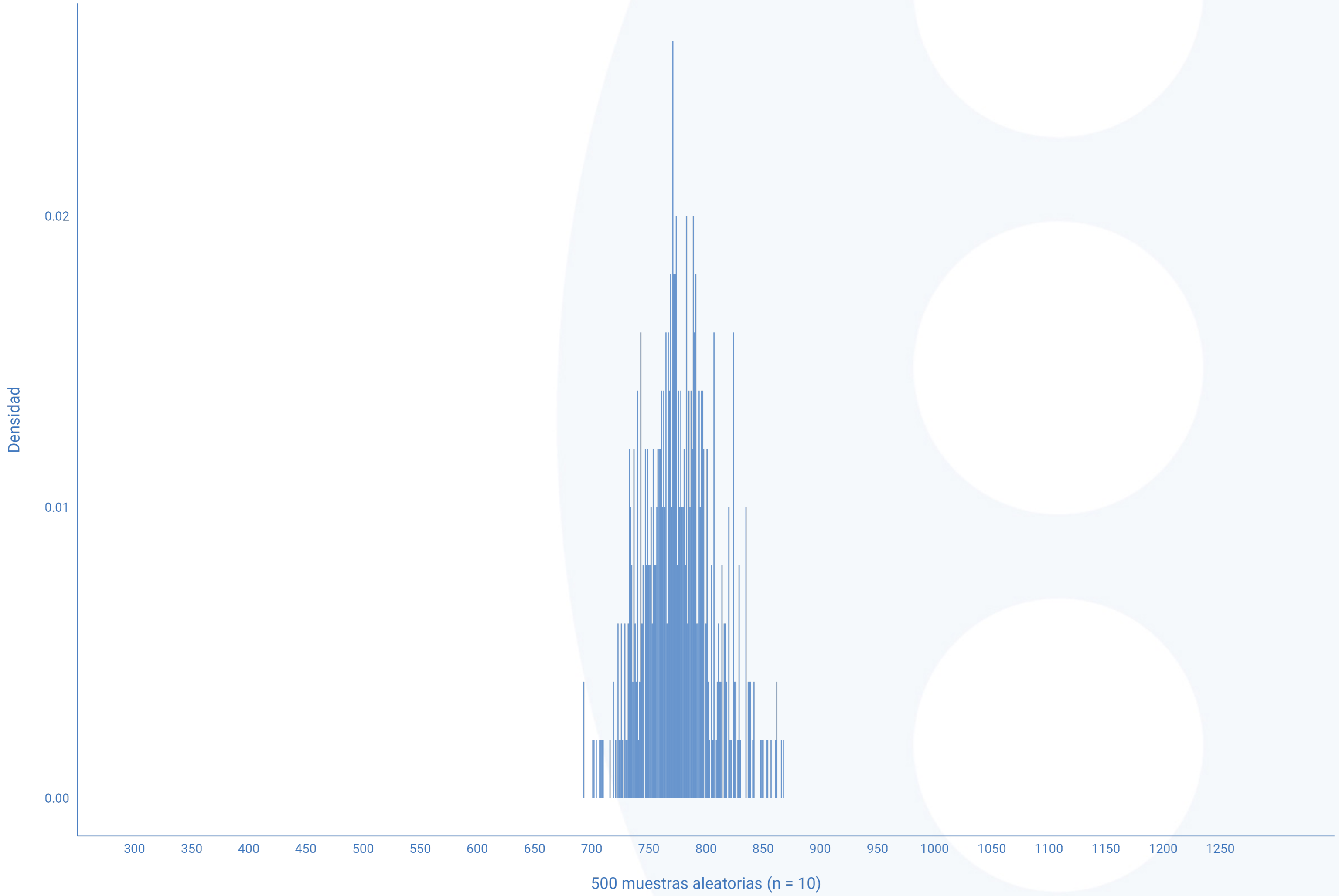
Vamos a realizar este ejercicio para 500 muestras. **Veamos como se ve una colección de muestras de 400 casos.**

Metodología Cuantitativa

Distribuciones muestrales

Medias de cada una de las muestras generadas

Distribución muestral de la media (n = 10)



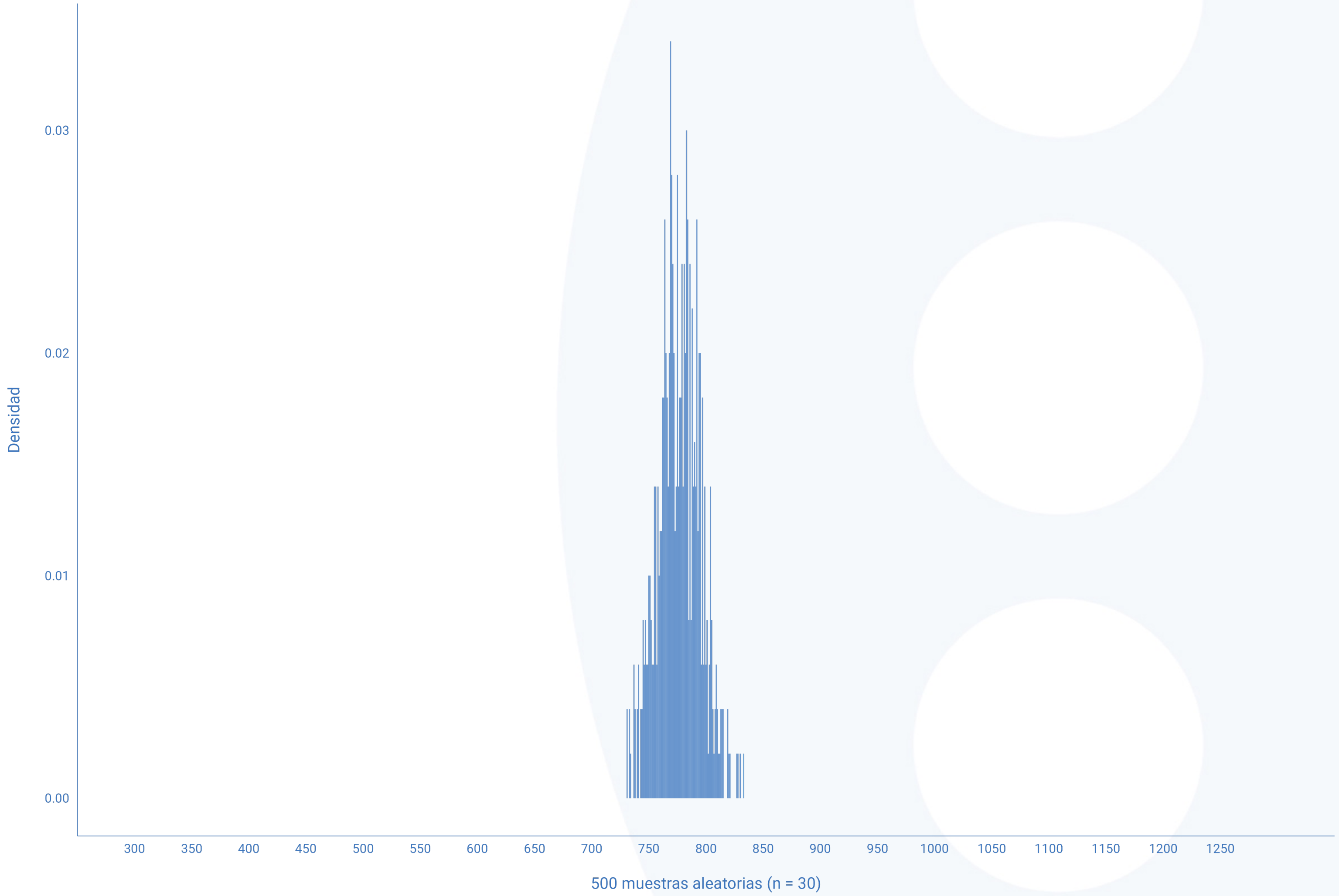
La distribución muestral de las medias para muestras de 10 casos, es mas estrecha que la variabilidad de todas las muestras generadas.

Y también, es menos variable o menos dispersa que la población. La expectativa es que es menos variable por un factor conocido.

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

¿Cómo se ve una distribución muestral de más casos? Veamos distribuciones de 30, 50, 100, y 500 casos.

Distribución muestral de la media (n = 30)



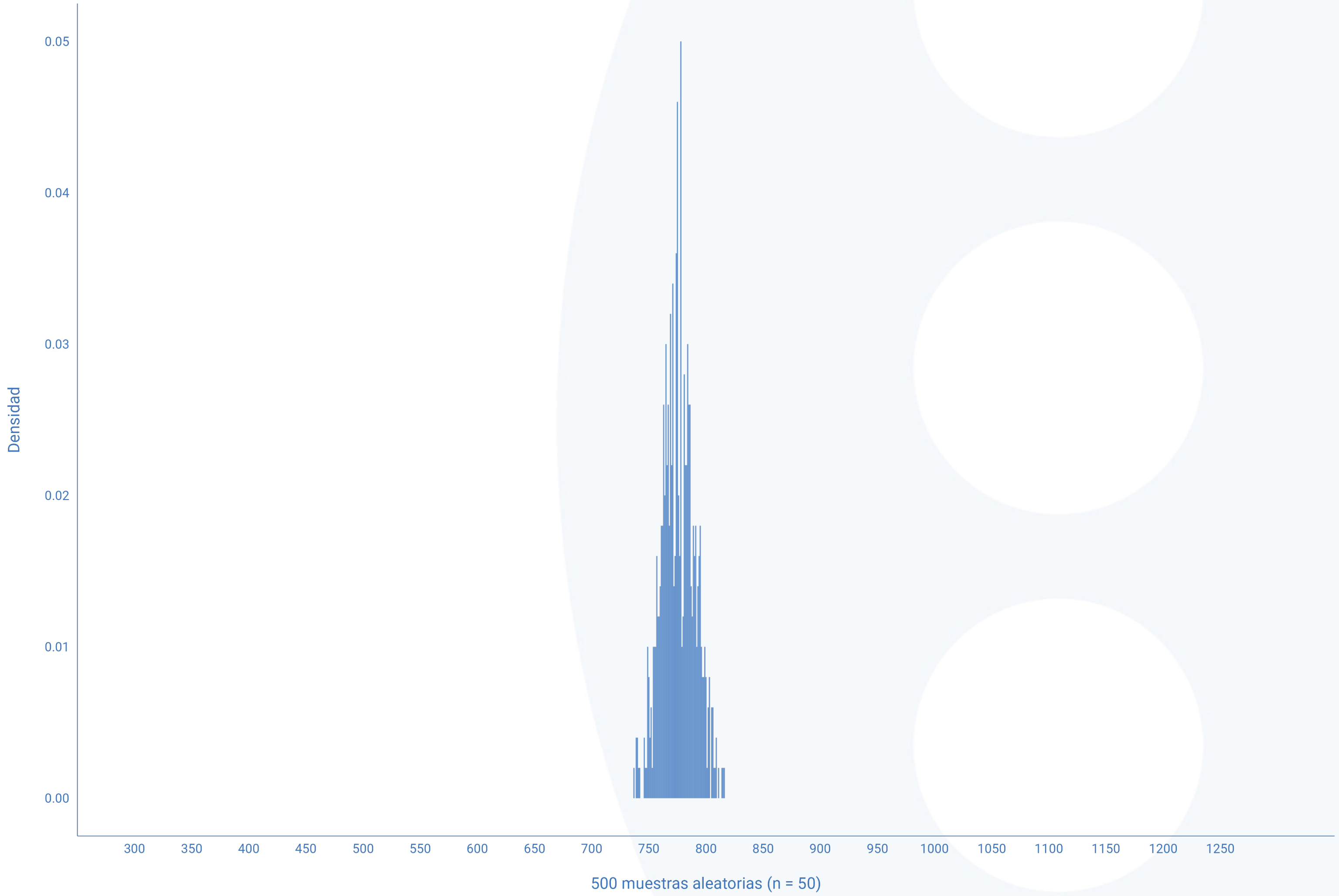
La distribución muestral de las medias para muestras de 10 casos, es mas estrecha que la variabilidad de todas las muestras generadas.

Y también, es menos variable o menos dispersa que la población. La expectativa es que es menos variable por un factor conocido.

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

¿Cómo se ve una distribución muestral de más casos? Veamos distribuciones de **30**, 50, 100, y 500 casos.

Distribución muestral de la media (n = 50)



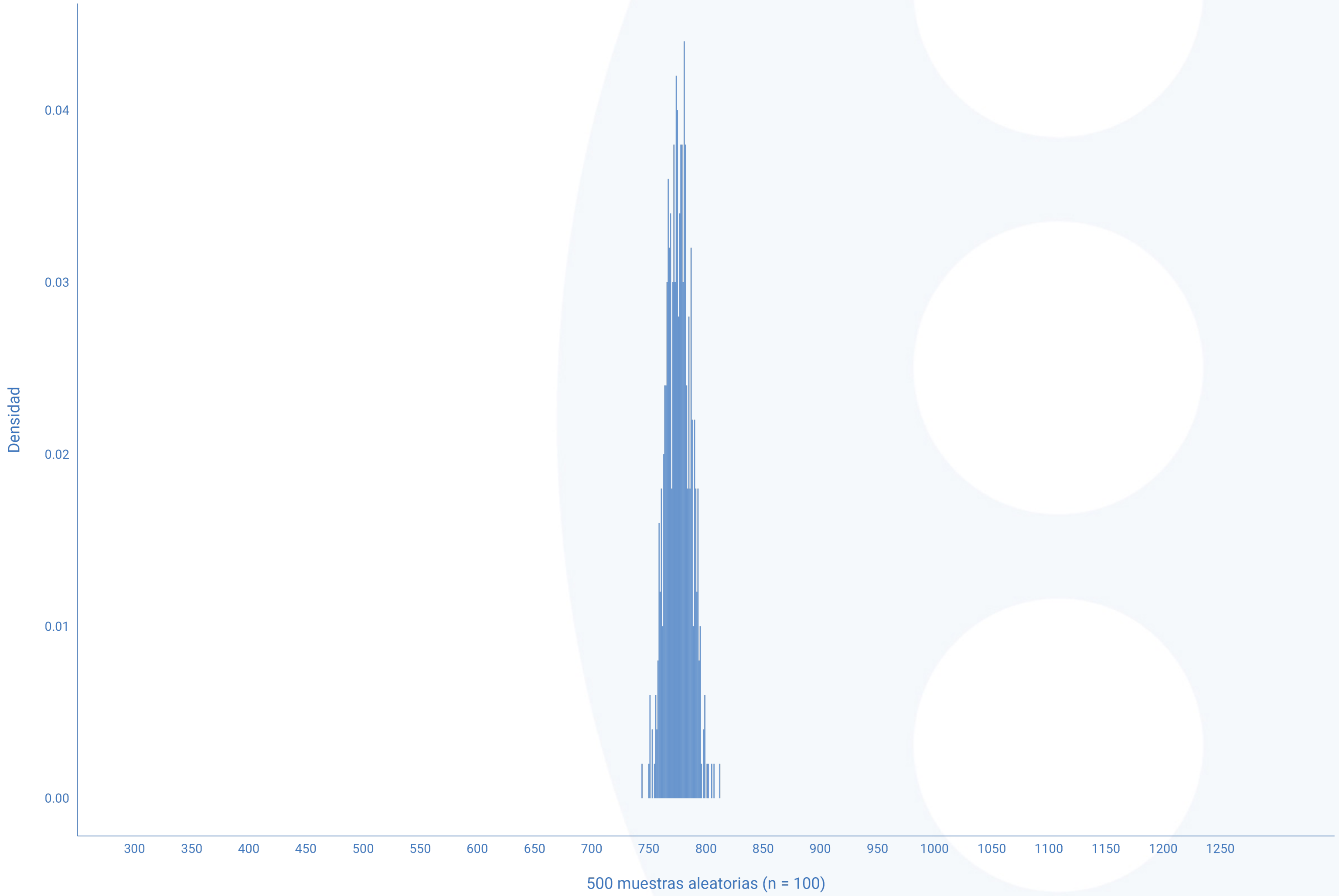
La distribución muestral de las medias para muestras de 10 casos, es mas estrecha que la variabilidad de todas las muestras generadas.

Y también, es menos variable o menos dispersa que la población. La expectativa es que es menos variable por un factor conocido.

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

¿Cómo se ve una distribución muestral de más casos? Veamos distribuciones de 30, **50**, 100, y 500 casos.

Distribución muestral de la media (n = 100)



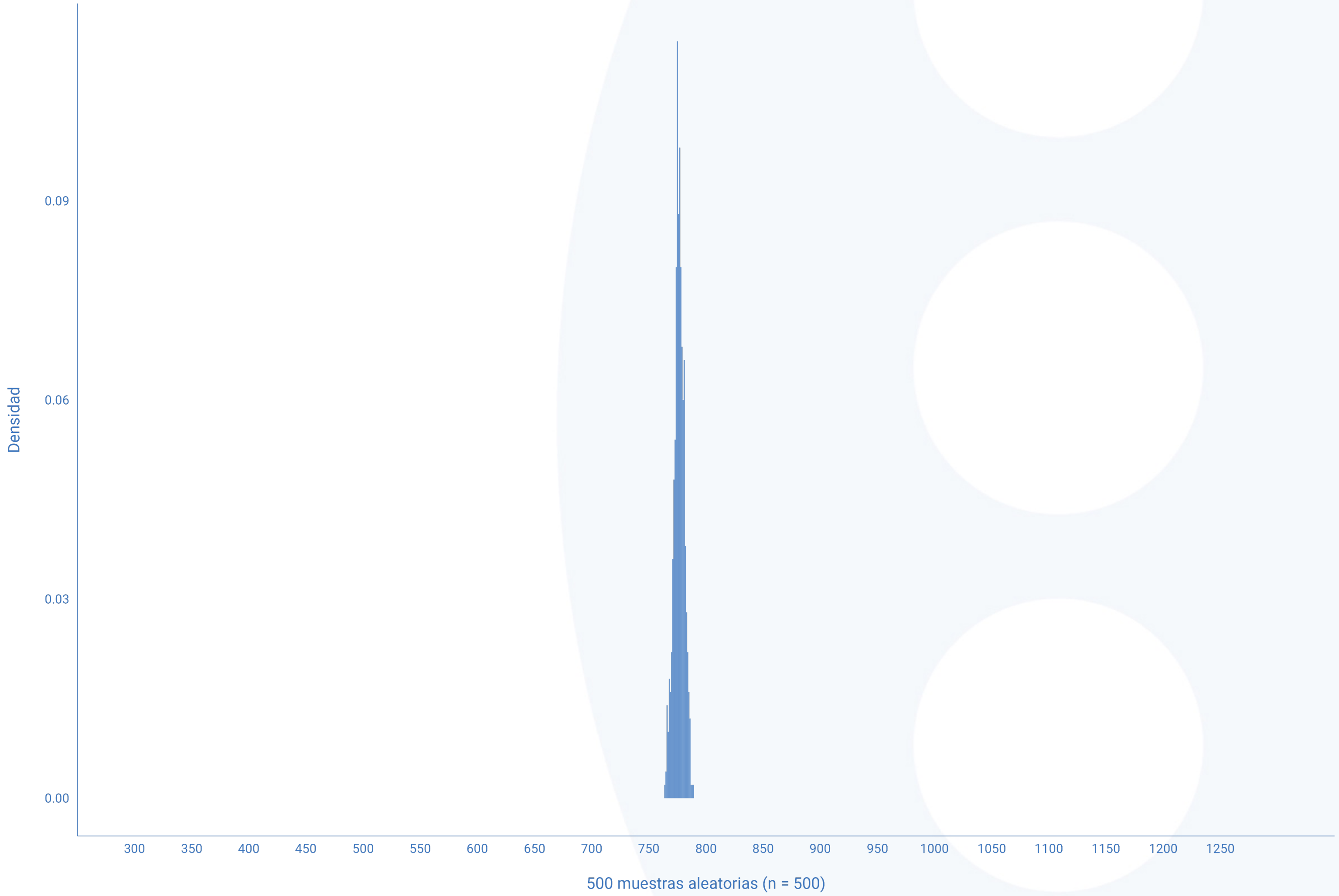
La distribución muestral de las medias para muestras de 10 casos, es mas estrecha que la variabilidad de todas las muestras generadas.

Y también, es menos variable o menos dispersa que la población. La expectativa es que es menos variable por un factor conocido.

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

¿Cómo se ve una distribución muestral de más casos? Veamos distribuciones de 30, 50, **100**, y 500 casos.

Distribución muestral de la media (n = 500)



A medida que el tamaño muestral aumenta, la dispersión de los valores posibles de las medias se hace más pequeño.

Otra forma de expresar esta idea, es que la desviación estándar, de la distribución muestral es cada vez más pequeña a medida de que el tamaño de las muestras es más grande.

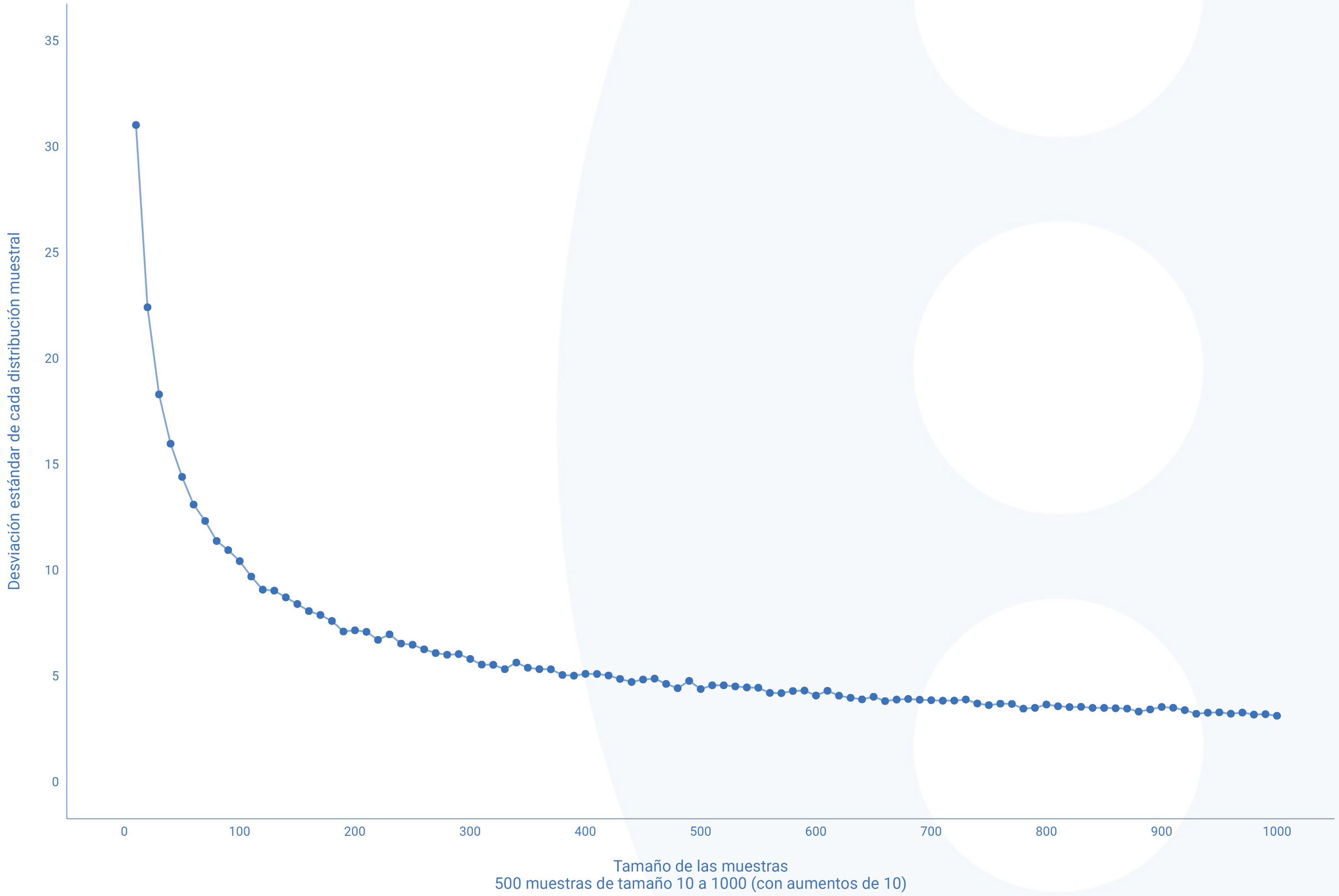
Ahora, vamos a visualizar esta misma idea, pero para un rango de muestras de 10 a 1000 observaciones.

Metodología Cuantitativa

Dispersión de la media

La dispersion de la media en cada distribución muestral

Dispersión de las medias de la distribución muestral, condicional al tamaño muestral



Entre muestras de 10 a 100, mientras mayor sea el tamaño muestral, la dispersión de las medias de cada distribución muestral decae rápidamente.

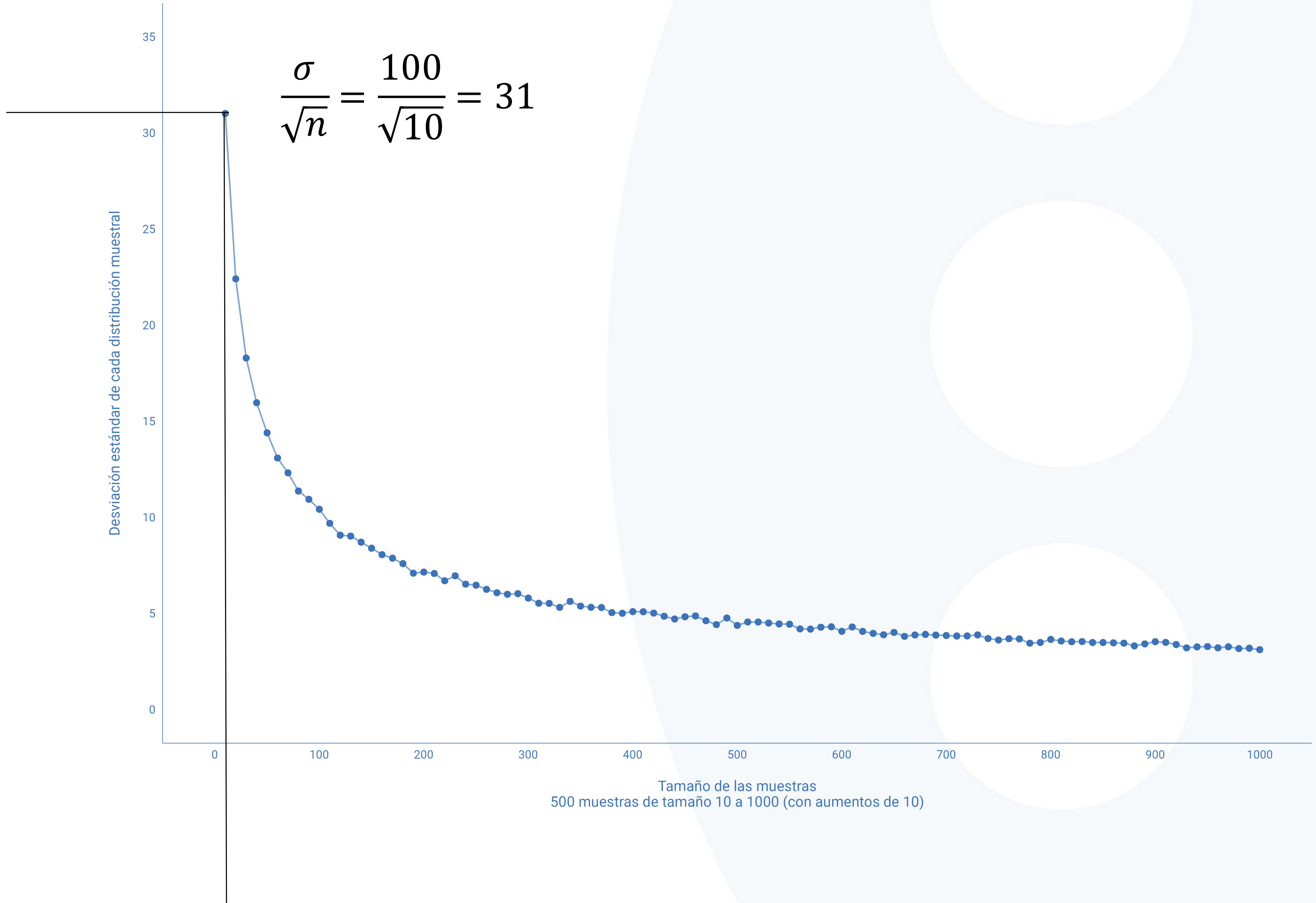
Entre 100 y 400 observaciones, sigue habiendo una menor dispersión, a medida de que aumenta el tamaño de las muestras.

De 400 caso en adelante, el aumento del tamaño es marginal.

Esta propiedad de las distribuciones muestrales es lo que expresa la ecuación del error estándar.

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

Dispersión de las medias de la distribución muestral, condicional al tamaño muestral



Entre muestras de 10 a 100, mientras mayor sea el tamaño muestral, la dispersión de las medias de cada distribución muestral decae rápidamente.

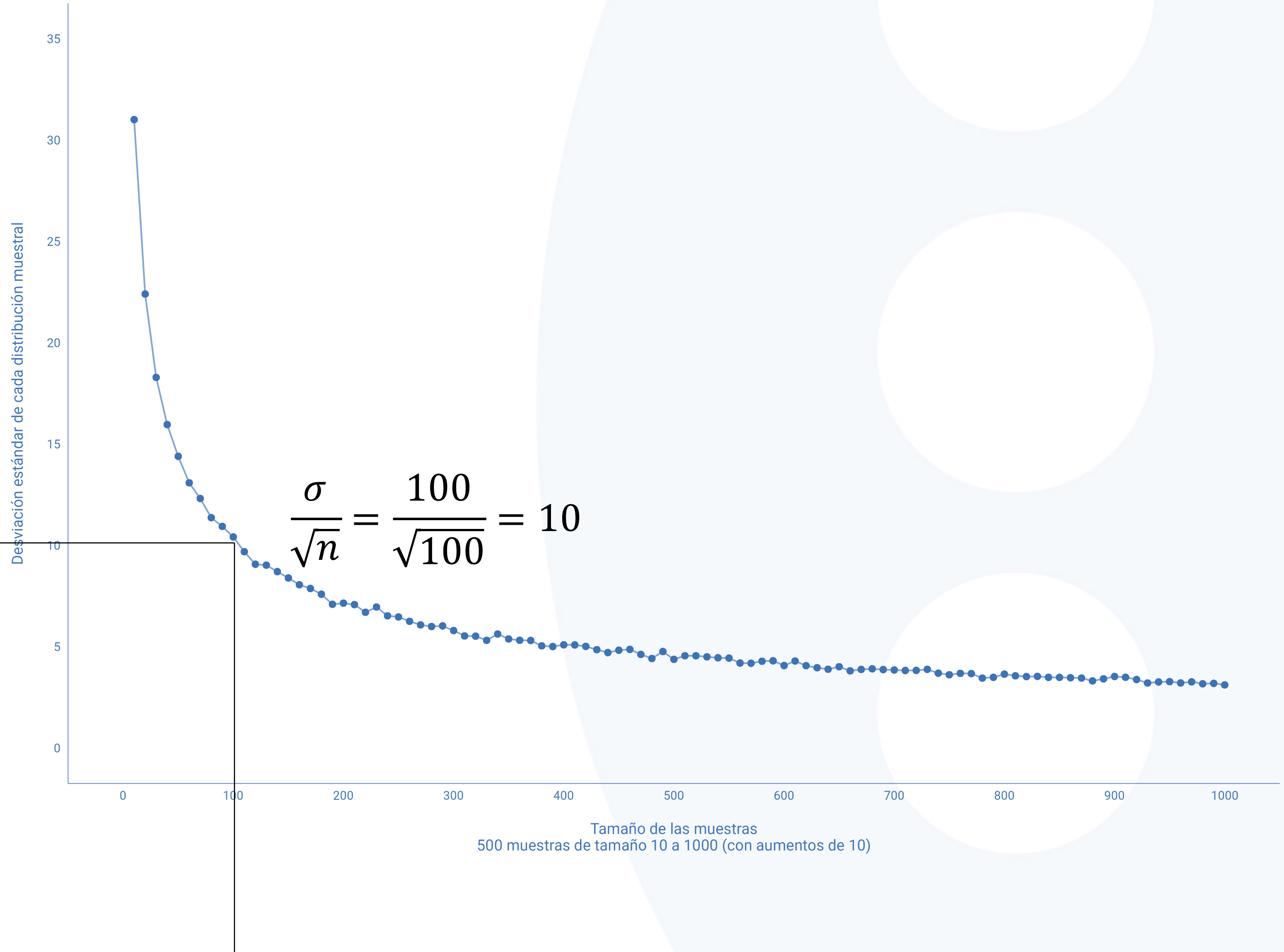
Entre 100 y 400 observaciones, sigue habiendo una menor dispersión, a medida de que aumenta el tamaño de las muestras.

De 400 caso en adelante, el aumento del tamaño es marginal.

Esta propiedad de las distribuciones muestrales es lo que expresa la ecuación del error estándar.

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

Dispersión de las medias de la distribución muestral, condicional al tamaño muestral



$$\frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{100}} = 10$$

Entre muestras de 10 a 100, mientras mayor sea el tamaño muestral, la dispersión de las medias de cada distribución muestral decae rápidamente.

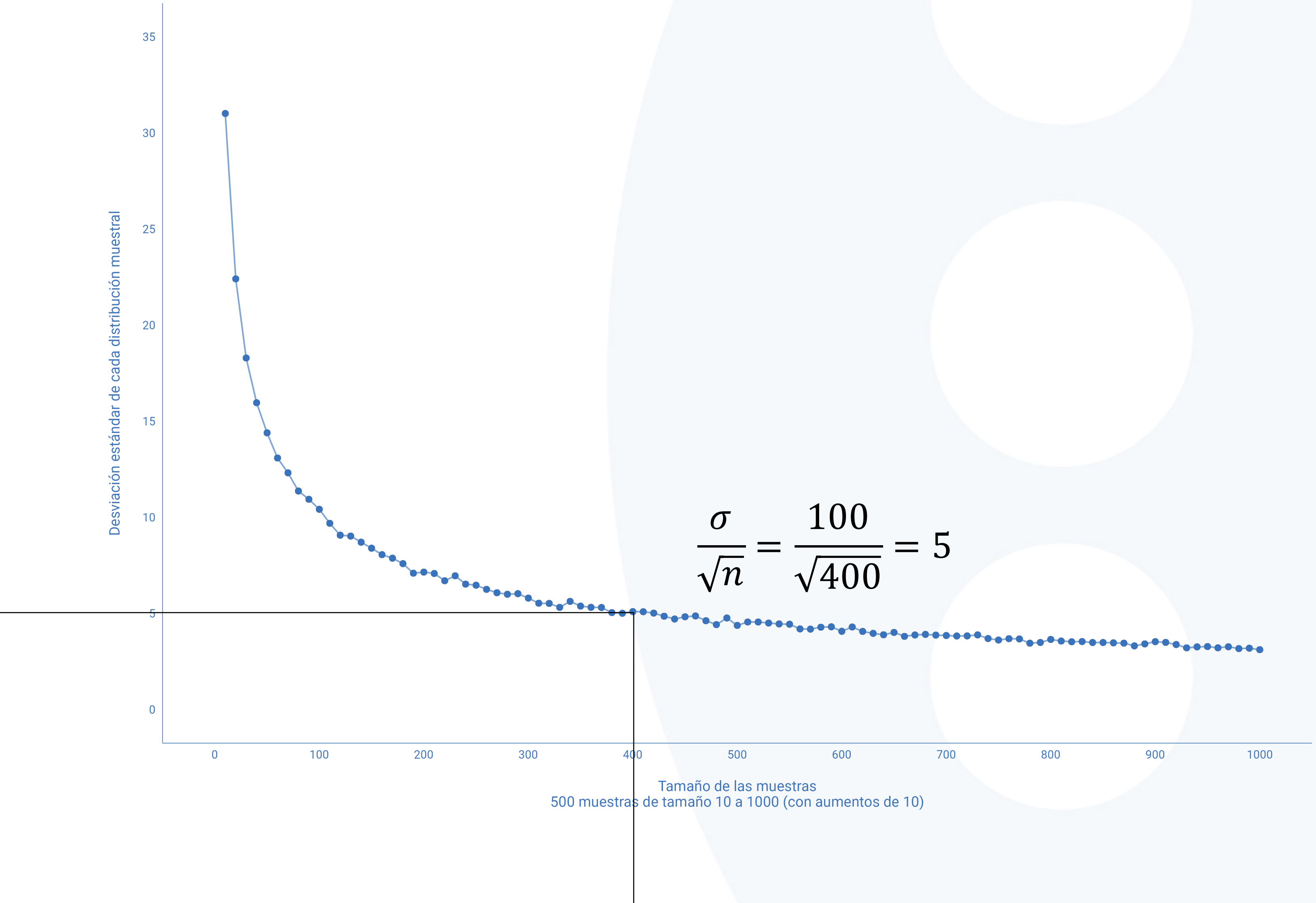
Entre 100 y 400 observaciones, sigue habiendo una menor dispersión, a medida de que aumenta el tamaño de las muestras.

De 400 caso en adelante, el aumento del tamaño es marginal.

Esta propiedad de las distribuciones muestrales es lo que expresa la ecuación del error estándar.

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

Dispersión de las medias de la distribución muestral, condicional al tamaño muestral



$$\frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{400}} = 5$$

Entre muestras de 10 a 100, mientras mayor sea el tamaño muestral, la dispersión de las medias de cada distribución muestral decae rápidamente.

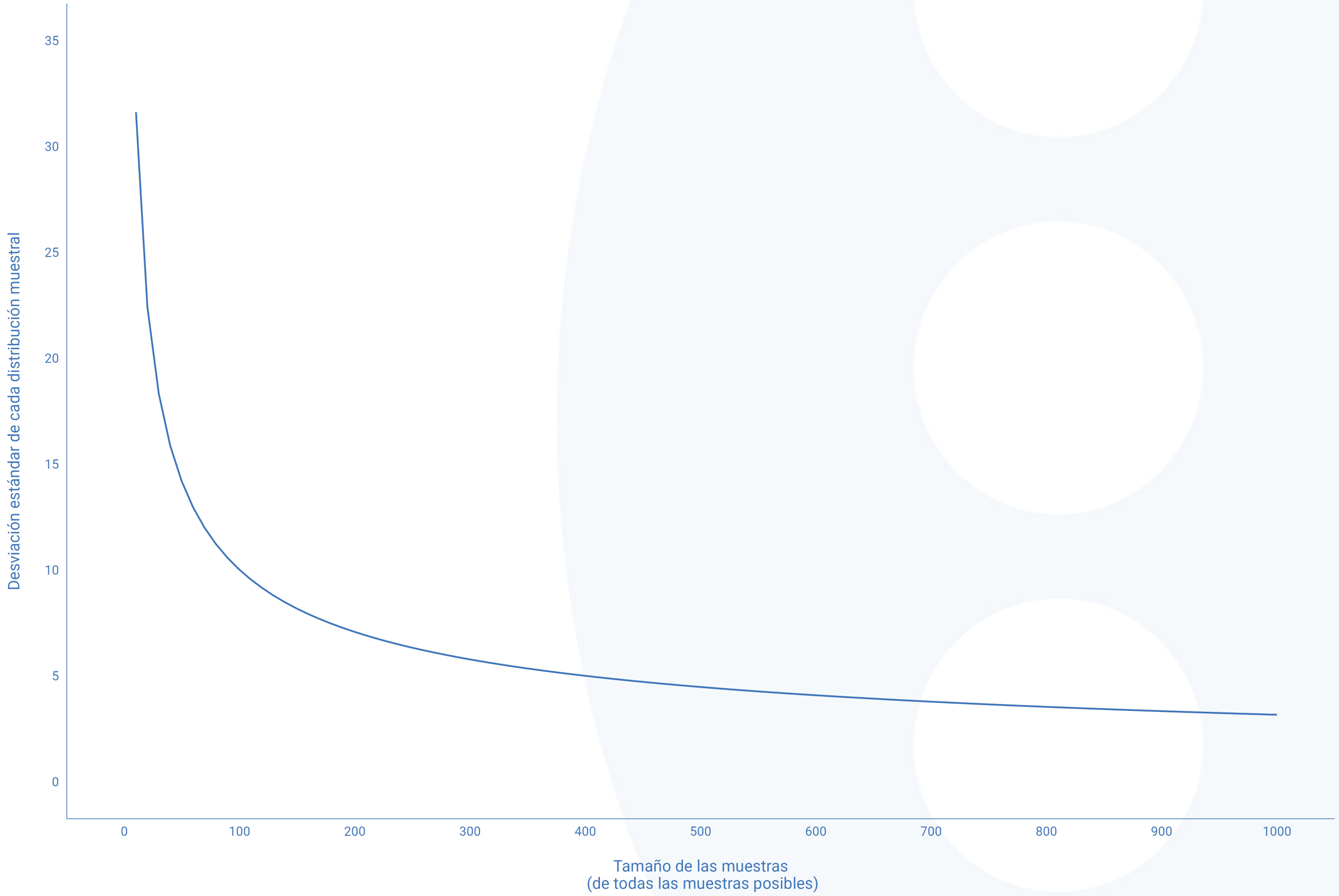
Entre 100 y 400 observaciones, sigue habiendo una menor dispersión, a medida de que aumenta el tamaño de las muestras.

De 400 caso en adelante, el aumento del tamaño es marginal.

Esta propiedad de las distribuciones muestrales es lo que expresa la ecuación del error estándar.

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

Dispersión de las medias de la distribución muestral, condicional al tamaño muestral (teórica)



Entre muestras de 10 a 100, mientras mayor sea el tamaño muestral, la dispersión de las medias de cada distribución muestral decae rápidamente.

Entre 100 y 400 observaciones, sigue habiendo una menor dispersión, a medida de que aumenta el tamaño de las muestras.

De 400 caso en adelante, el aumento del tamaño es marginal.

Esta propiedad de las distribuciones muestrales es lo que expresa la ecuación del error estándar.

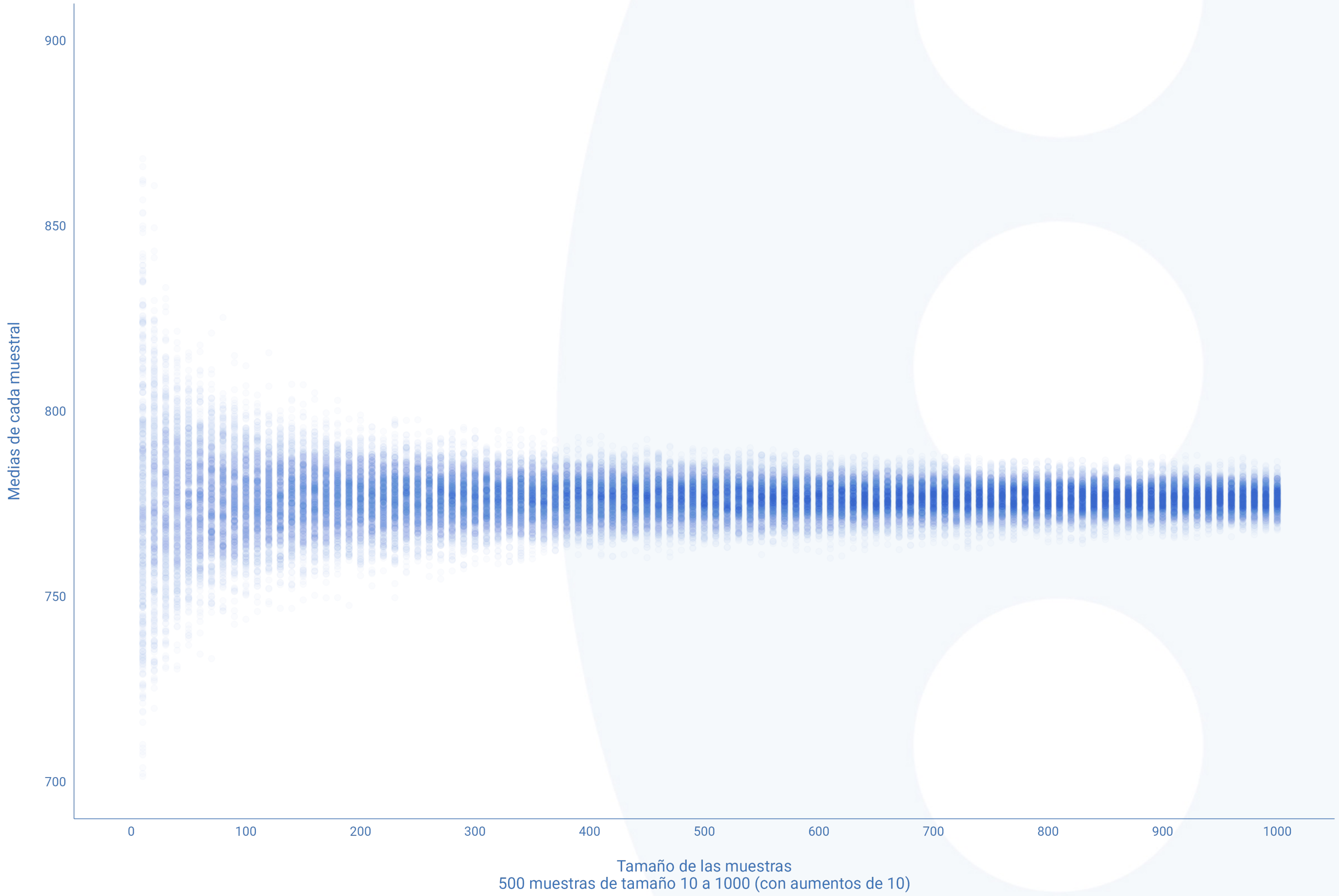
$$\frac{\sigma}{\sqrt{n}}$$

Metodología Cuantitativa

Ley de los grandes números

Mientras más grande sea la muestra, más se parece al promedio de la población

Ley de los grandes números

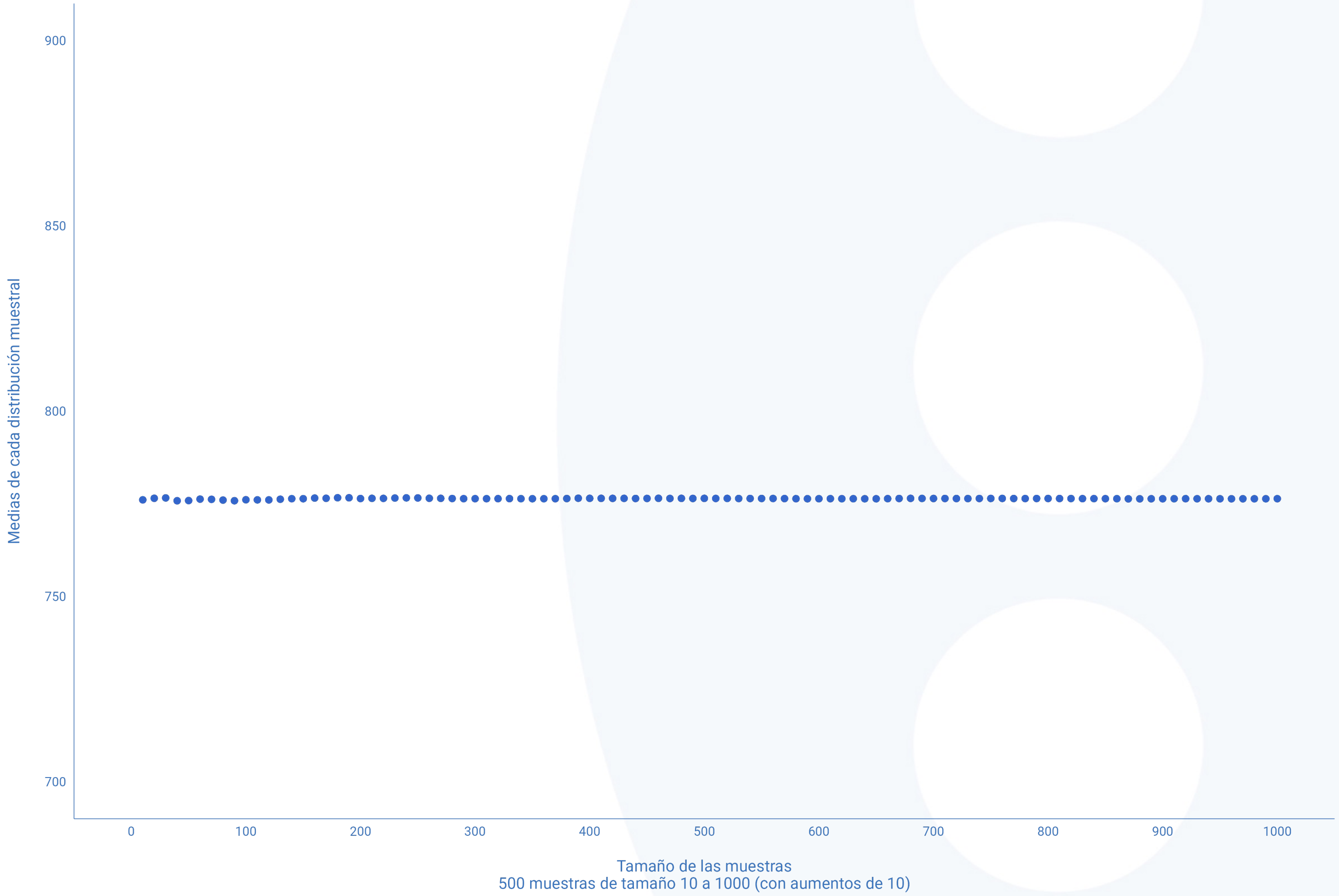


Mientras más grande sea una muestra aleatoria simple, menos se aleja del promedio de la población. Mientras más se acerque al infinito el tamaño muestral, la expectativa es que la media de cada muestra se aproxima al promedio de la población.

La versión más simple de la ley de los grandes números es que a medida que una muestra es más grande, su promedio se acerca al promedio de la población.

En la presente figura tenemos la media de cada muestra generada, condicionada por el tamaño muestral.

Ley de los grandes números y teorema del limite central



En la presente figura tenemos la media de cada distribución muestral, condicionada por el tamaño muestral.

El promedio de todos los promedio de cada muestra, es decir, el promedio de cada distribución muestral, converge al promedio de la población.

Mientras la ley de los grandes números versa sobre el promedio de una muestra; el teorema del limite central versa sobre la distribución muestral.

Metodología Cuantitativa

Distribución normal

La distribución muestral es una distribución normal

Teorema del limite central y distribución muestral

Tenemos cuatro ideas centrales.

Mientras más grande una muestra, más chances tiene de acercarse al promedio de la población.

El promedio de la distribución muestral, de seguro se acerca al promedio de la población.

La media de una muestra, se aleja del promedio de la distribución muestral, con una dispersión conocida (i.e., el error estándar de la media, o desviación estándar de las medias en la distribución muestral).

Además, sin importar la forma de la distribución de la población, las tres ideas anteriores se sostienen para las medias.

No importa la forma de la distribución de la población, la distribución muestral **nos permite recuperar parámetros de la población**.

Distribución muestral y distribución normal

Uno de los apoyos centrales de la inferencia basada en diseño, es que la **distribución muestral toma una distribución normal**.

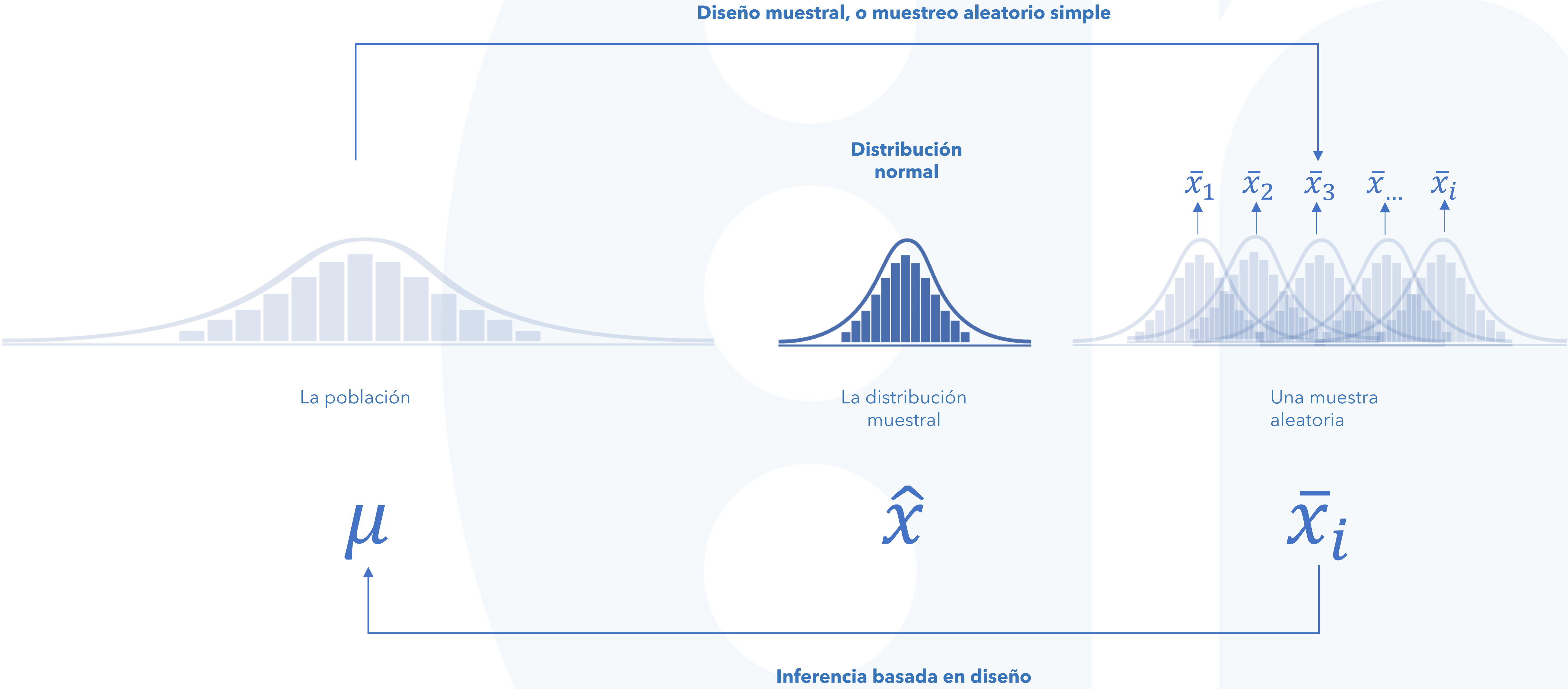
Y por tanto, podemos emplear una **distribución probabilística** para realizar inferencias. Las distribuciones probabilísticas las podemos pensar como histogramas de distribuciones ideales, que tienen una área bajo la curva que es igual a uno. Y con esta característica podemos saber cuales la proporción de casos esperados bajo o sobre cada valor de la distribución.

La **distribución normal estandarizada**, es una distribución de este tipo. Podemos emplear lo que sabemos de esta distribución para realizar inferencias con una muestral aleatoria.

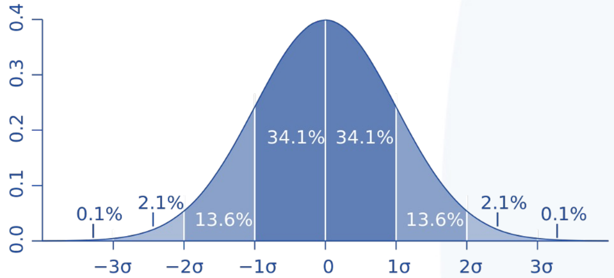
Y podemos proceder de una forma muy parecida, con las distribuciones t.

No importa la forma de la distribución de la población, la distribución muestral **nos permite recuperar parámetros de la población**.

La distribución muestral es una distribución normal

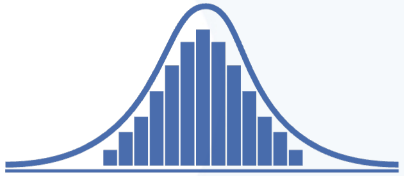


La distribución muestral como distribución normal estandarizada

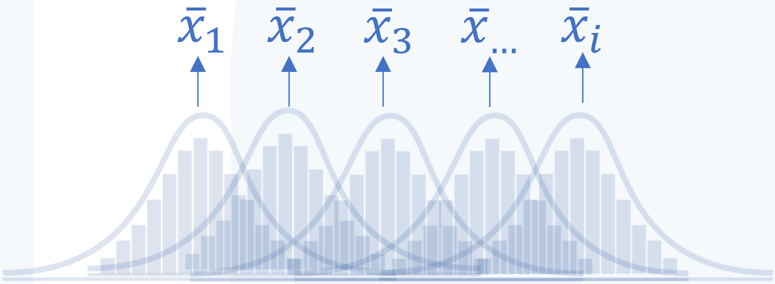


Distribución normal estandarizada

Distribución normal



La distribución muestral



Una muestra aleatoria

μ

$\hat{x}$

$\bar{x}_i$

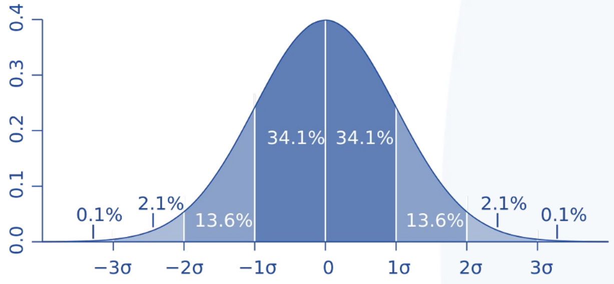
El parámetro poblacional, toma la posición de la media en la distribución normal estandarizada, donde esta es un valor cero. Recordemos que el promedio de las distribuciones muestrales converge a μ

Inferencia basada en diseño

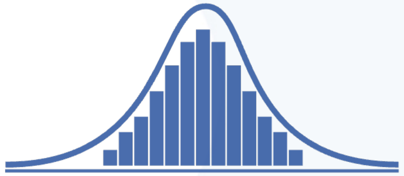
La distribución muestral como distribución normal estandarizada

Podemos transformar a la distribución muestral, a una distribución normal estandarizada

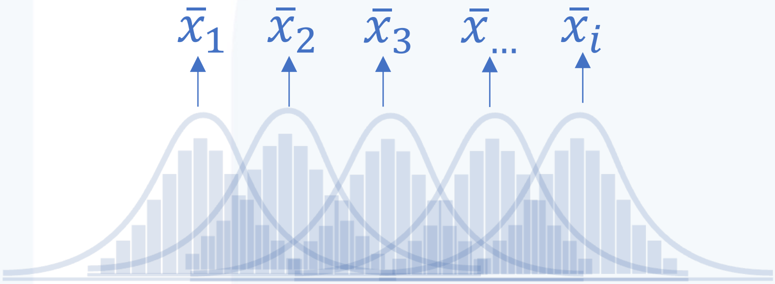
Distribución normal



Distribución normal estandarizada



La distribución muestral



Una muestra aleatoria

Podemos transformar los valores de una distribución muestral, a puntajes z.

μ

$\hat{x}$

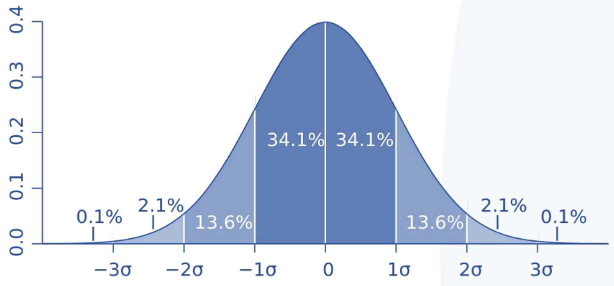
$\bar{x}_i$

Inferencia basada en diseño

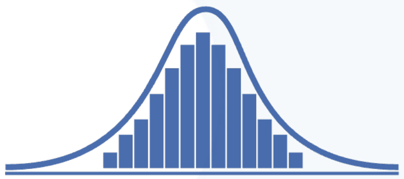
La distribución muestral como distribución normal estandarizada

```
# r code
mutate(z_score = (raw_score – mean(raw_score))/sd(raw_score)
```

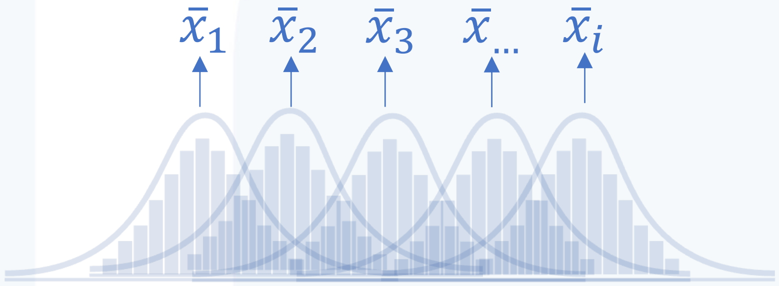
Distribución normal



Distribución normal estandarizada



La distribución muestral



Una muestra aleatoria

μ

$\hat{x}$

$\bar{x}_i$

Inferencia basada en diseño

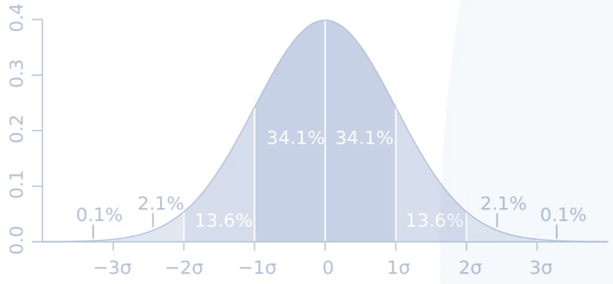
Podemos transformar los valores de una distribución muestral, a puntajes z.

Así como podemos transformar los valores de la distribución muestral a puntajes z, entonces por extensión, podemos hacer lo mismo con una de las muestras aleatorias.

La distribución muestral como distribución normal estandarizada

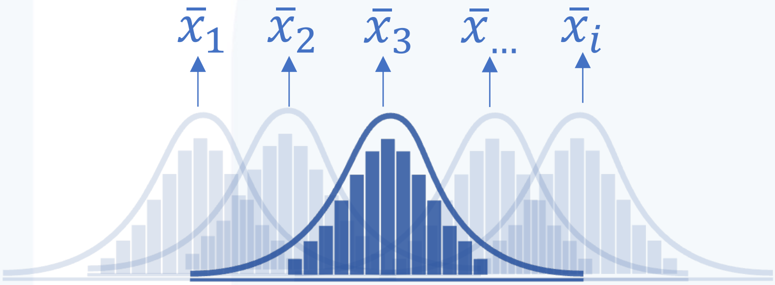
```
# r code
mutate(z_score = (raw_score – mean(raw_score))/sd(raw_score)
```

Distribución normal



Distribución normal estandarizada

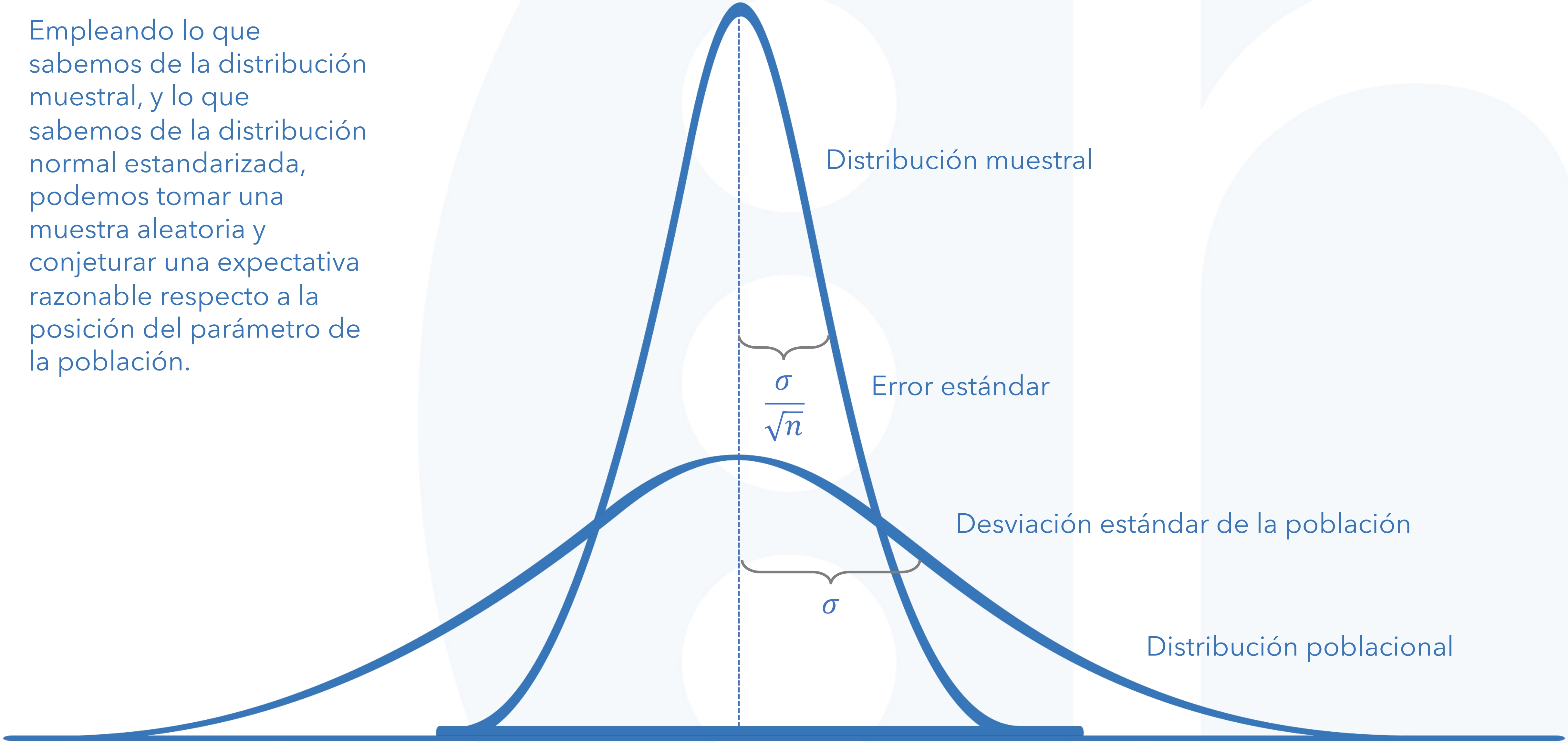
La distribución muestral



Una muestra aleatoria

Si transformamos los valores de una muestra aleatoria a puntajes z, entonces sabemos sabríamos su posición relativa con respecto al parámetro de interés.

Empleando lo que sabemos de la distribución muestral, y lo que sabemos de la distribución normal estandarizada, podemos tomar una muestra aleatoria y conjeturar una expectativa razonable respecto a la posición del parámetro de la población.

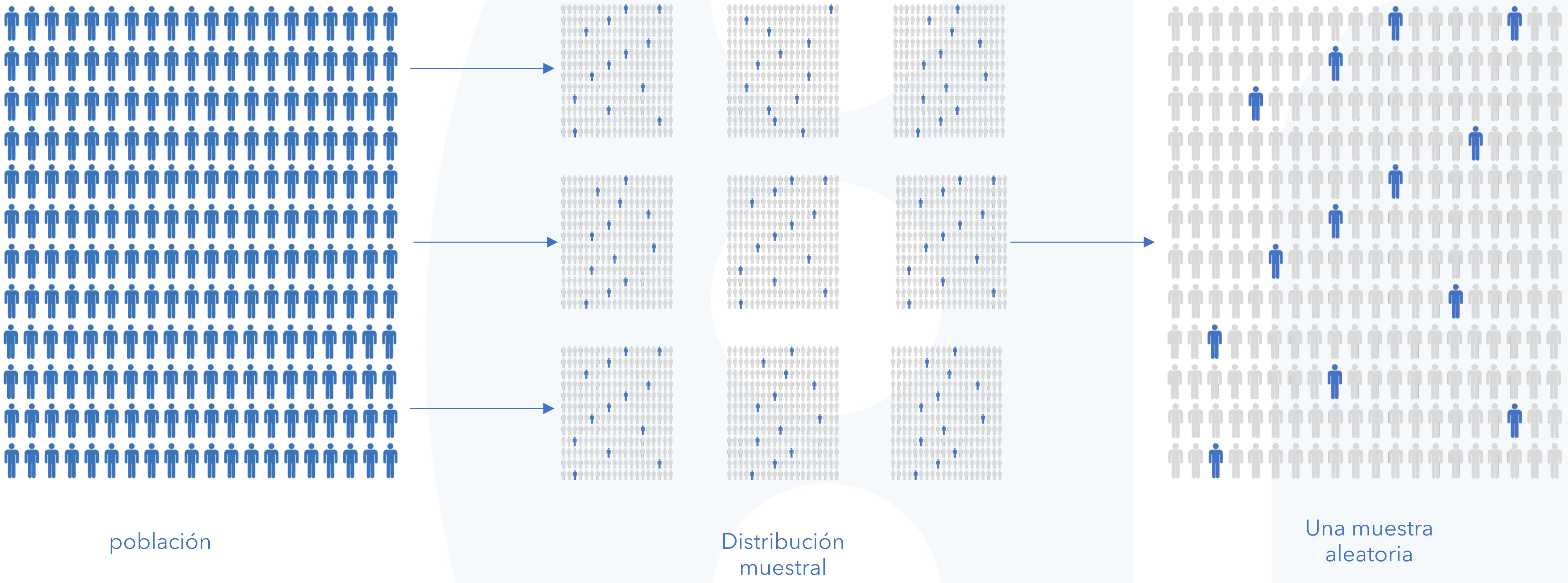


Metodología Cuantitativa

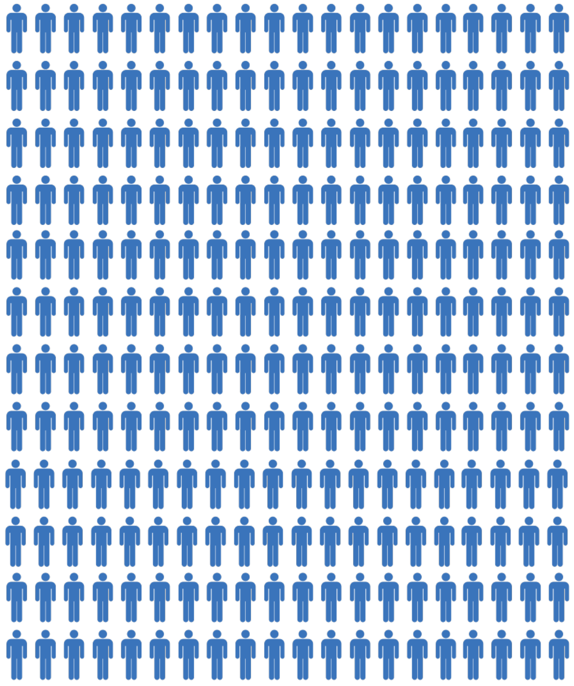
Muestras aleatorias

Posición de las muestras aleatorias con respecto al parámetro poblacional

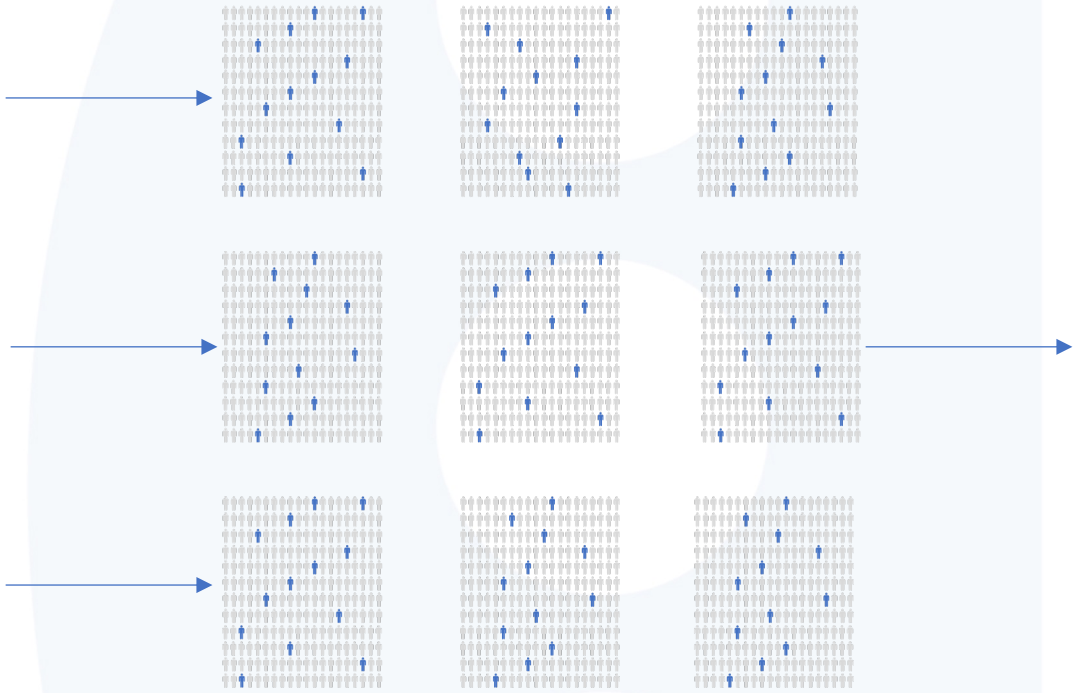
De la población, a la distribución muestral, a una muestra aleatoria



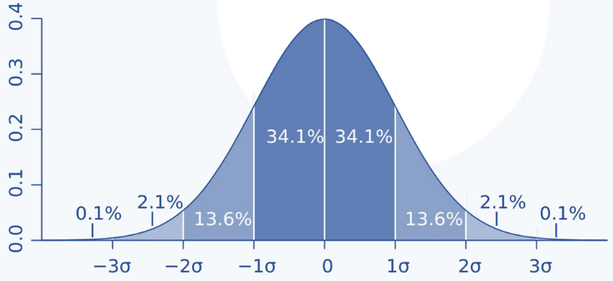
De la población, a la distribución muestral, a una muestra aleatoria

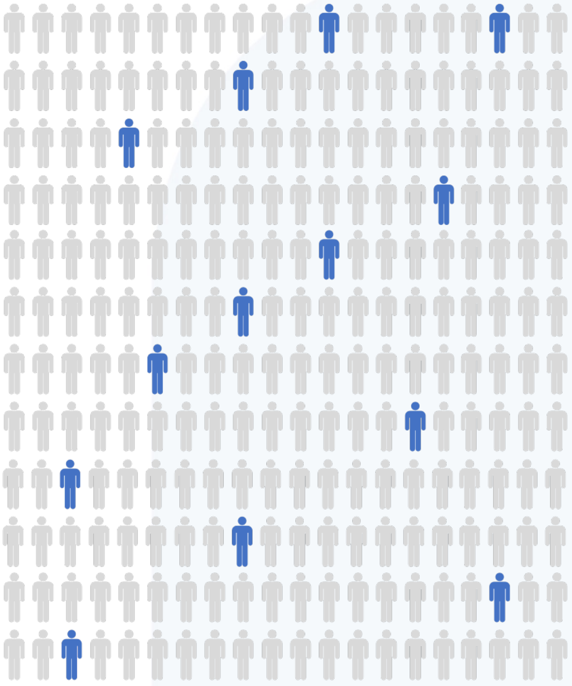


población

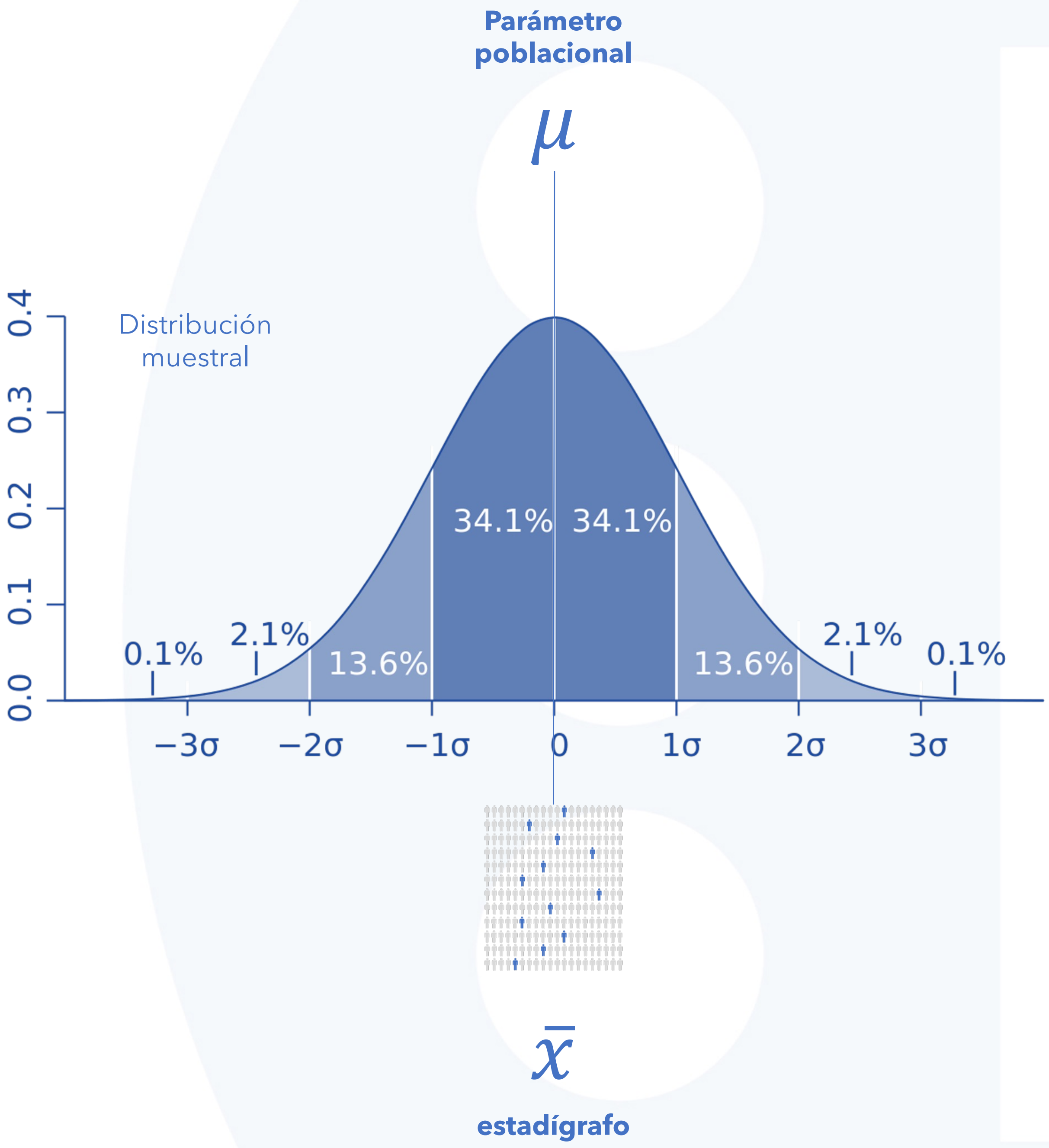


Distribución
muestral





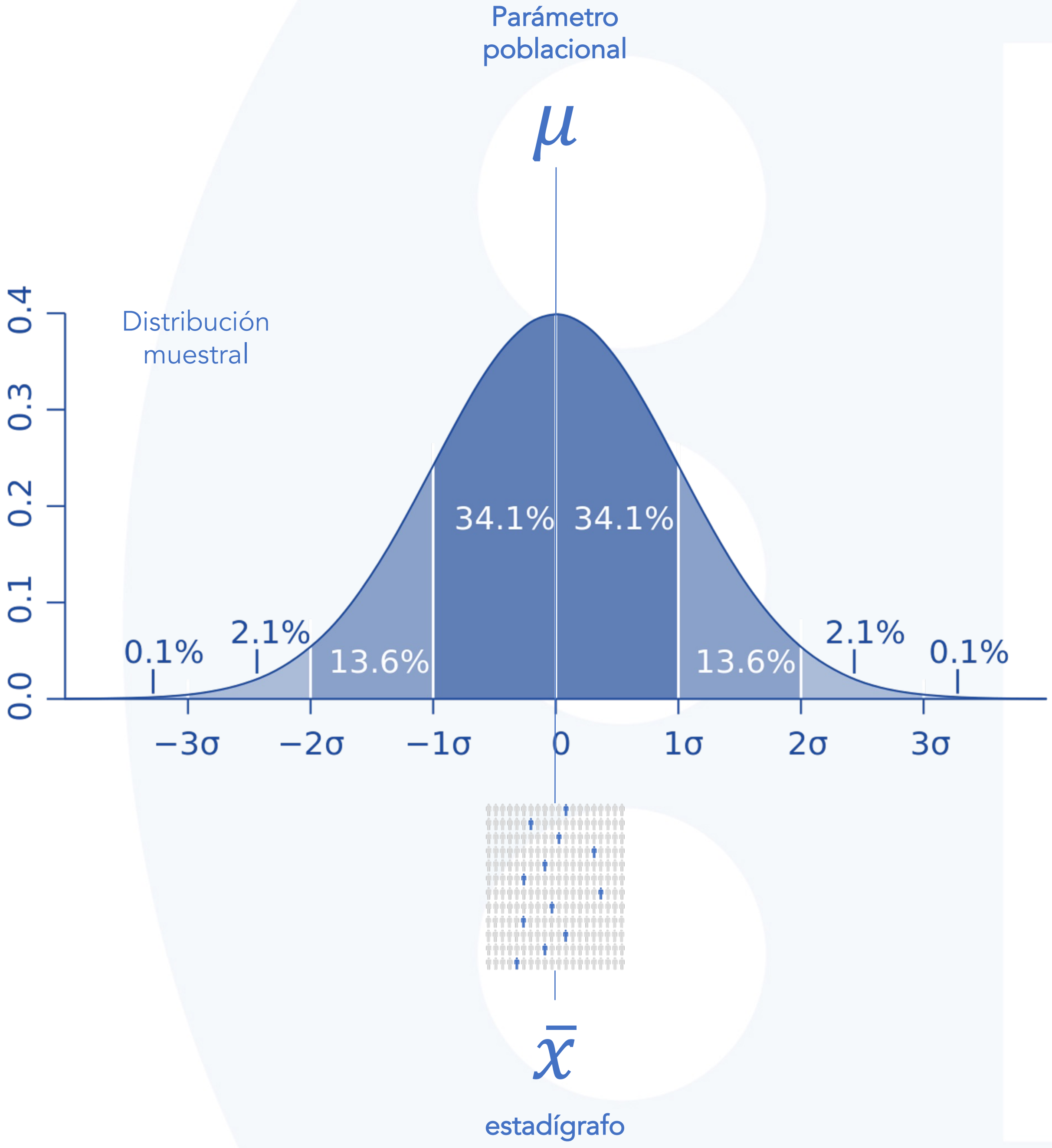
Una muestra
aleatoria

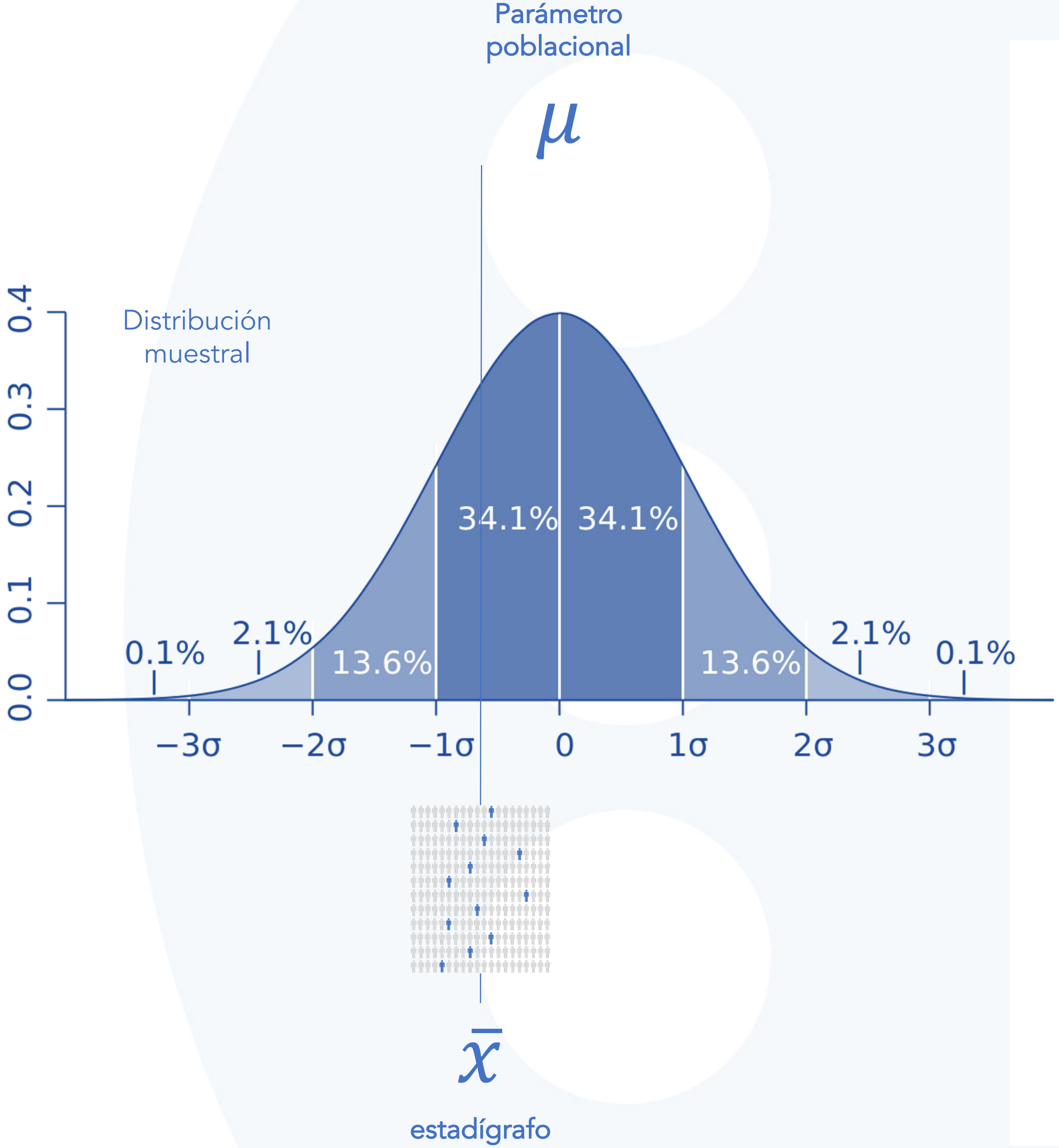


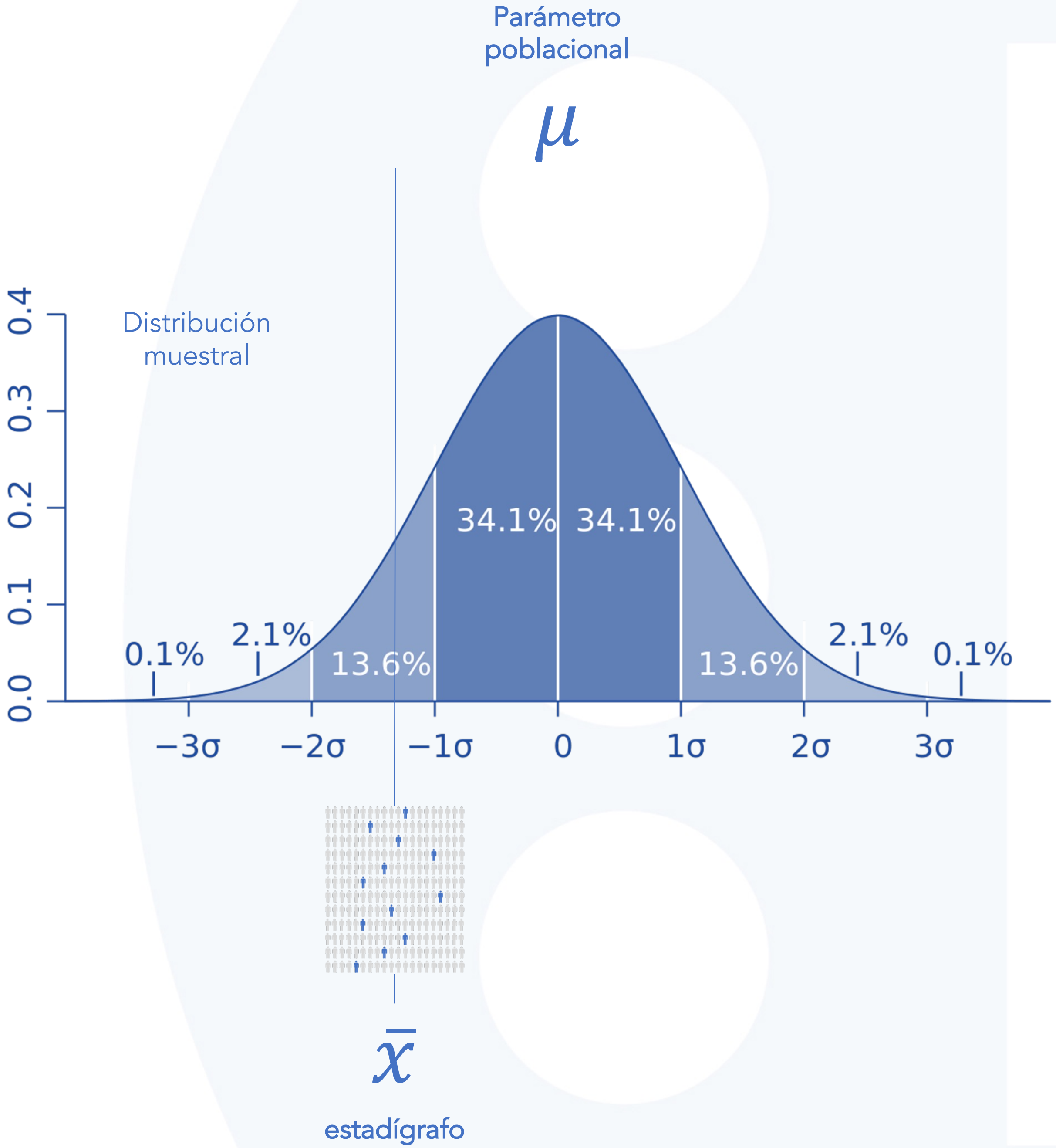
Sabemos que la media de la distribución muestral, converge con el parámetro de la población.

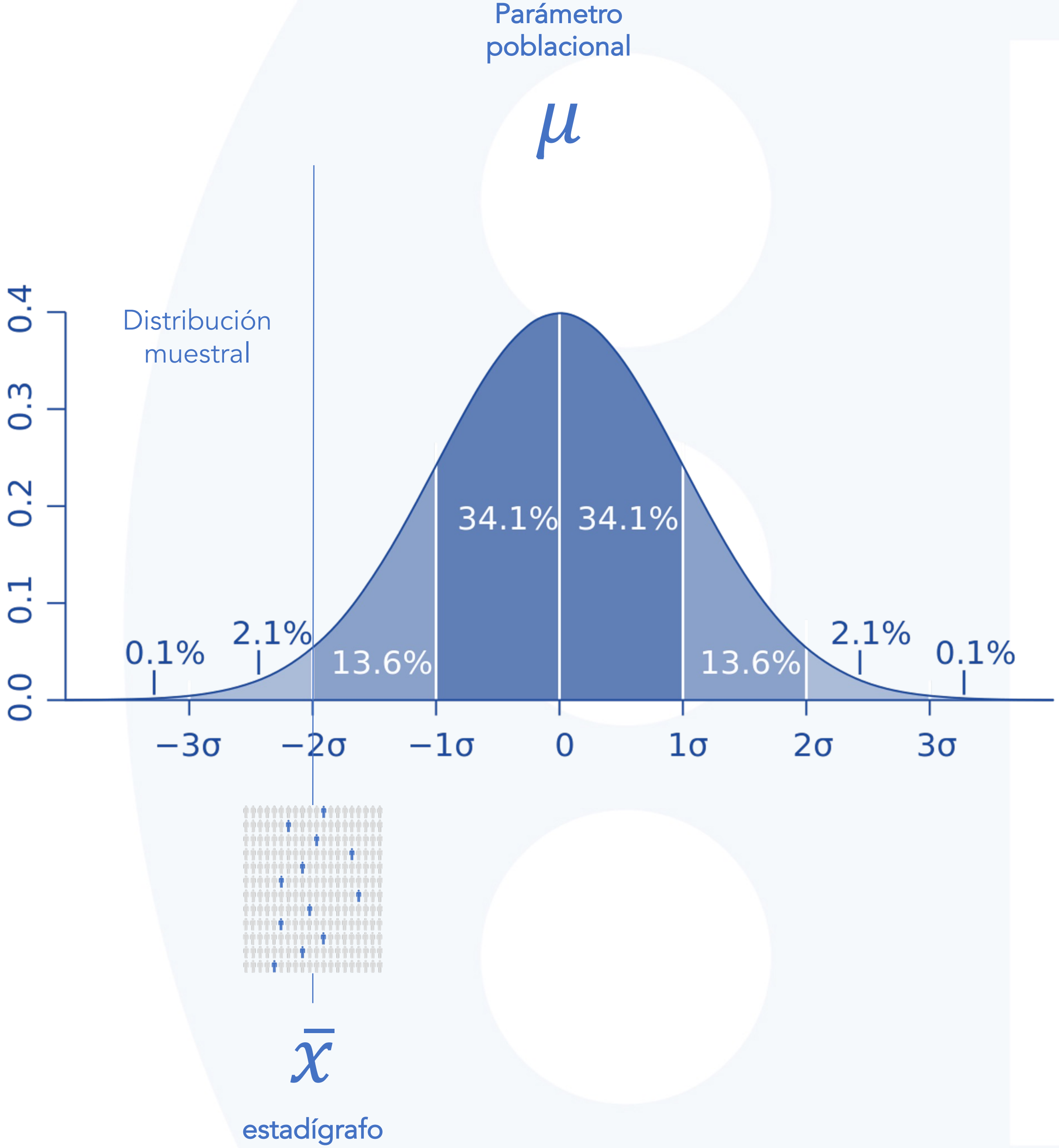
Y que la media de una muestra aleatoria, podría tomar cualquier posición con respecto a la distribución muestral.

Y que debido a que esta distribución es normal, entonces podemos expresar su desviación con respecto al parámetro normal empleando a la distribución normal estandarizada.

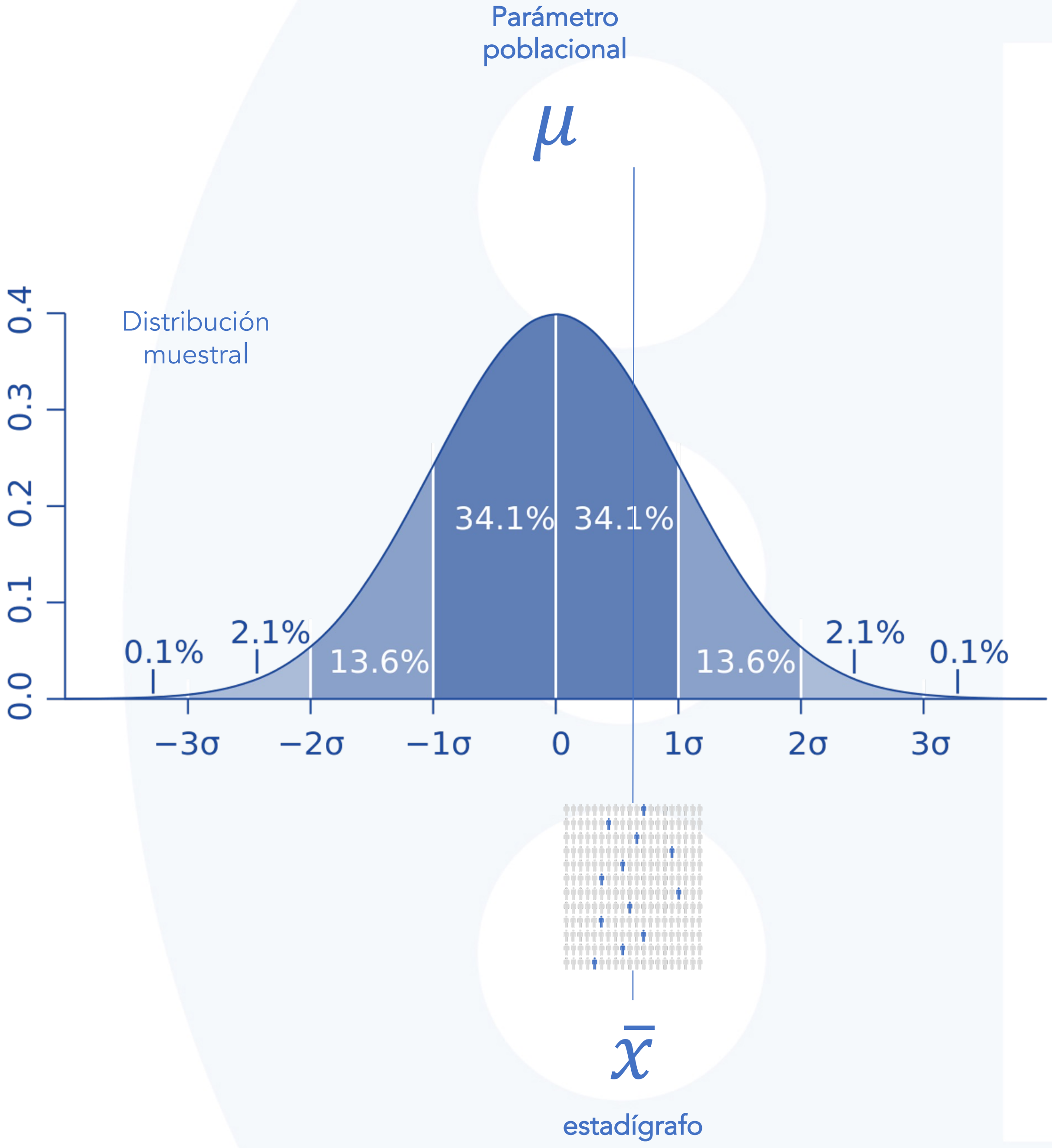

Parámetro poblacional
μ
Distribución muestral
0.4
0.3
0.2
0.1
0.0
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
0.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ
x̄
estadígrafo

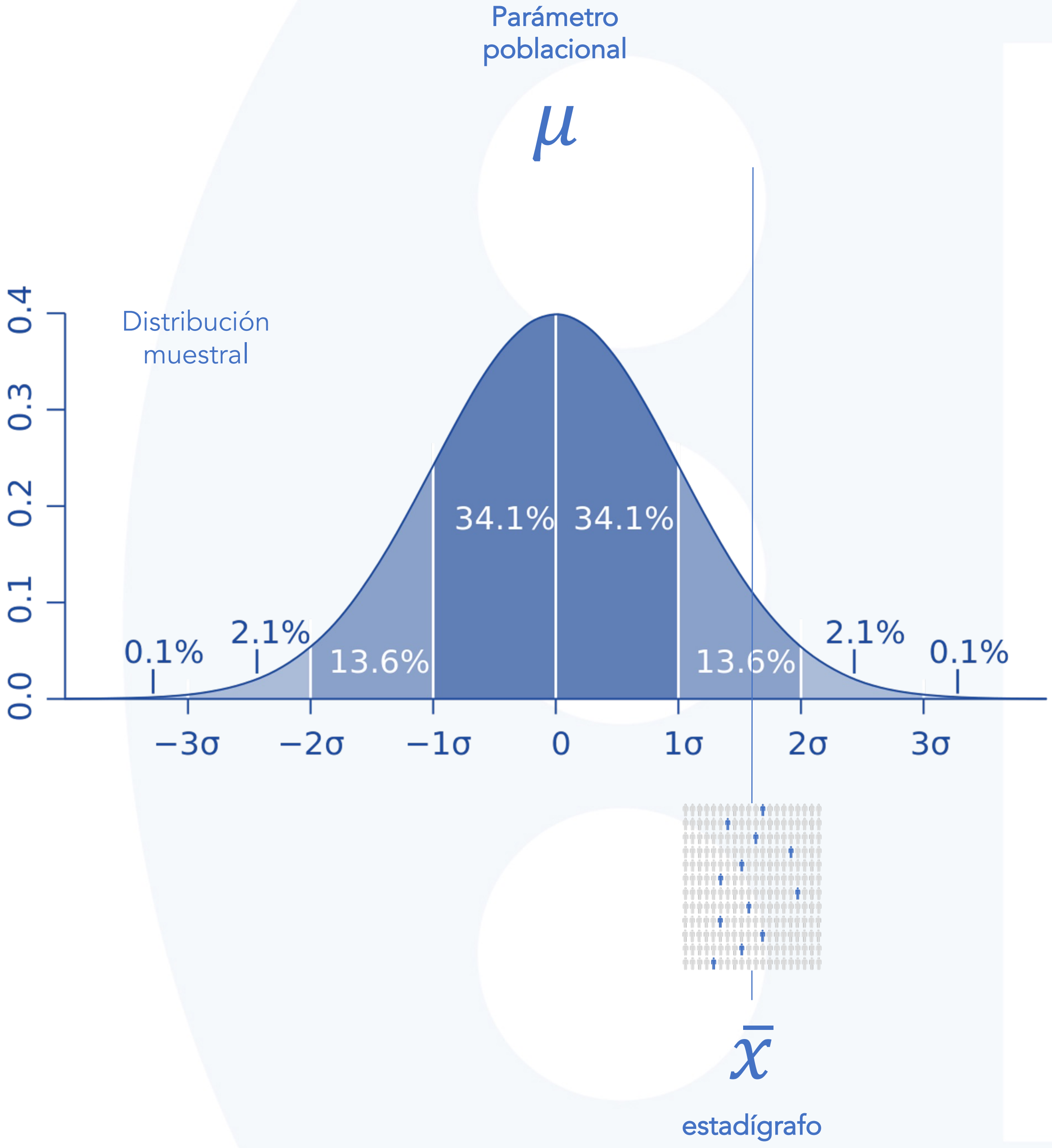

Parámetro poblacional
μ
Distribución muestral
0.4
0.3
0.2
0.1
0.0
0.1%
2.1%
13.6%
34.1%
34.1%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ
x̄
estadígrafo


Parámetro poblacional
μ
Distribución muestral
0.4
0.3
0.2
0.1
0.0
0.1%
2.1%
13.6%
34.1%
34.1%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ
x̄
estadígrafo

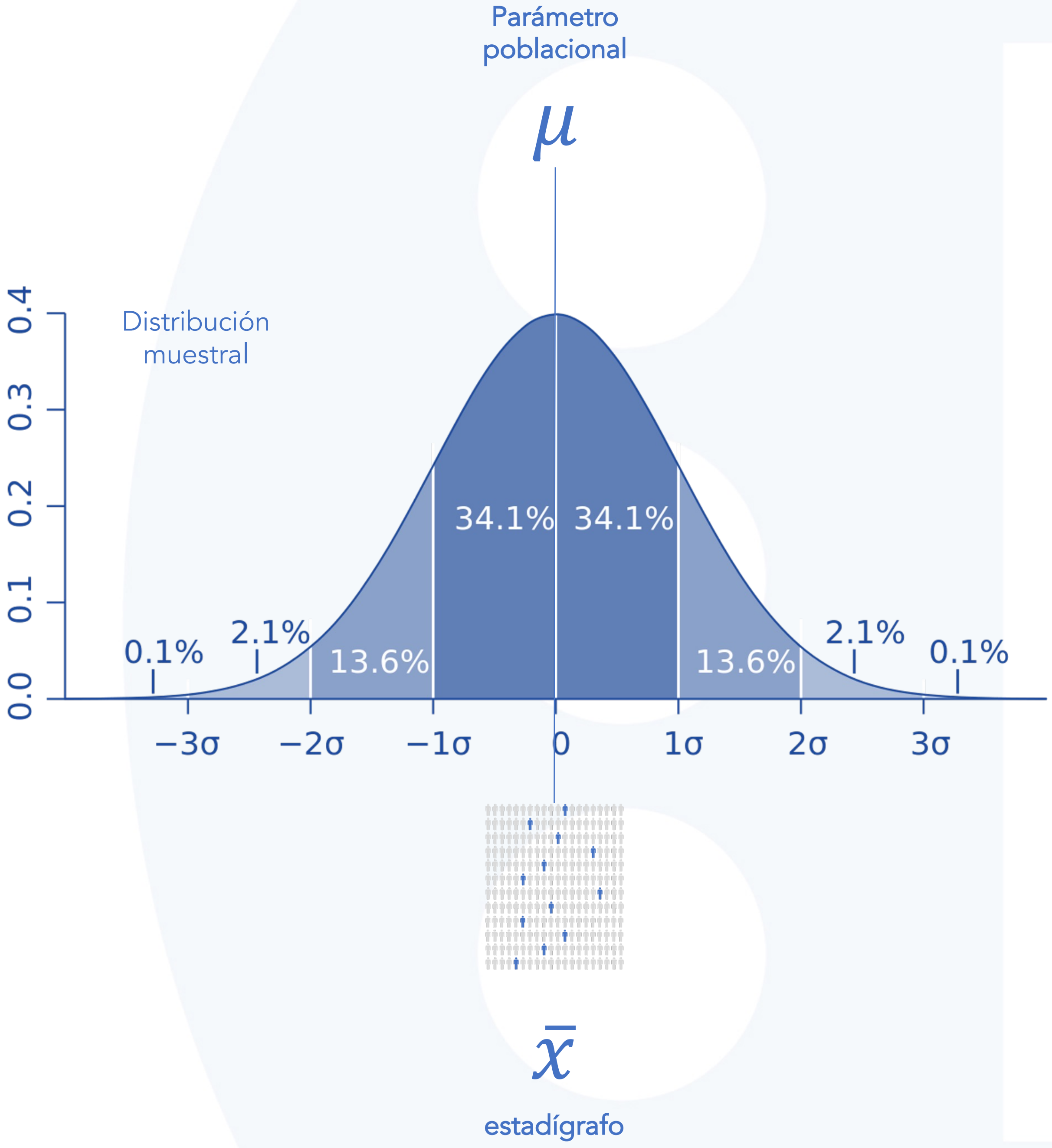


Parámetro
poblacional
μ
Distribución
muestral
0.0
0.1
0.2
0.3
0.4
0.1%
2.1%
13.6%
34.1%
34.1%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ
$\bar{x}$
estadígrafo


Parámetro
poblacional
μ
Distribución
muestral
0.4
0.3
0.2
0.1
0.0
34.1%
34.1%
13.6%
13.6%
2.1%
2.1%
0.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ
x̄
estadígrafo



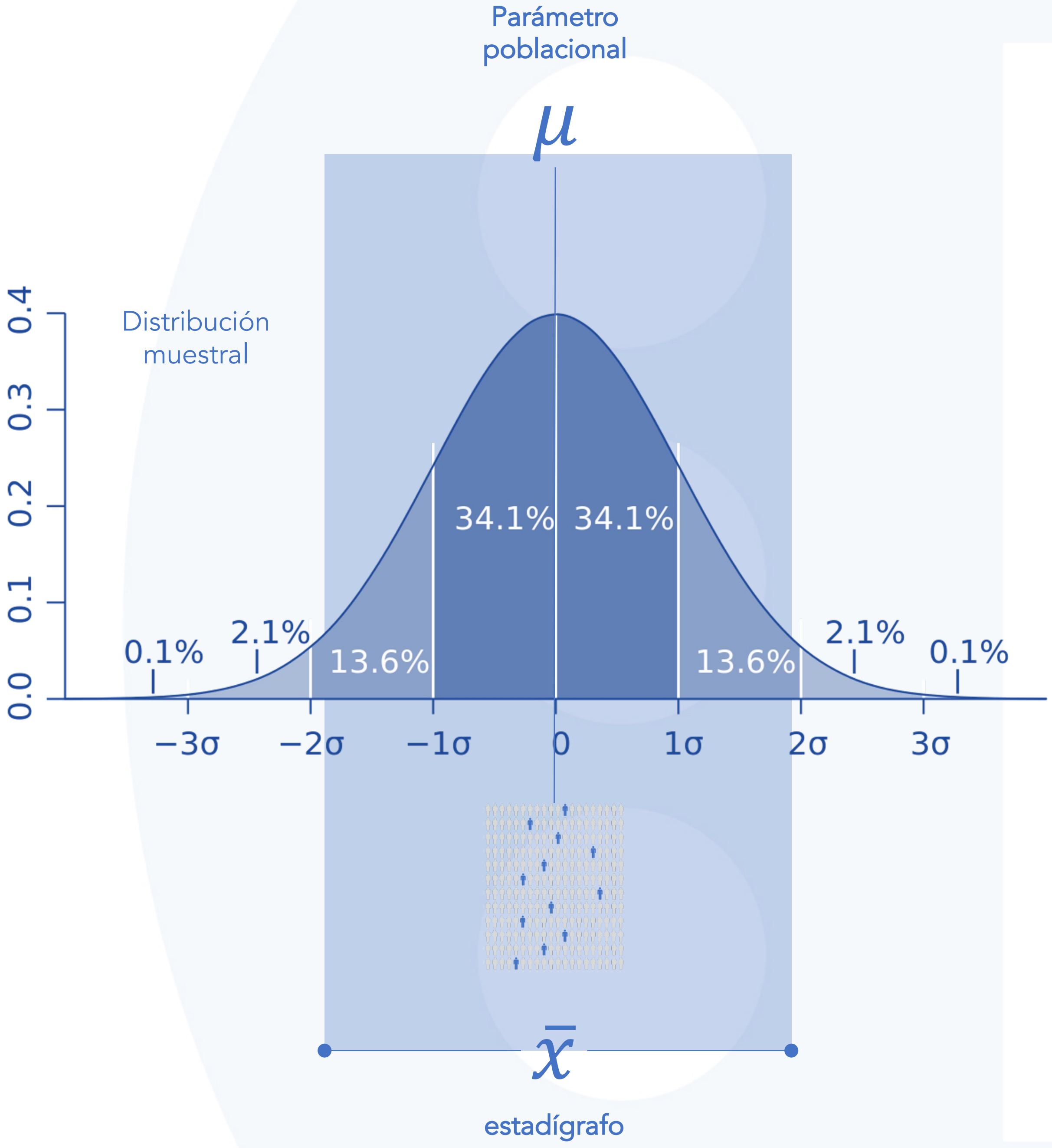
Parámetro poblacional
μ
Distribución muestral
0.0
0.1
0.2
0.3
0.4
0.1%
2.1%
13.6%
34.1%
34.1%
13.6%
2.1%
0.1%
−3σ
−2σ
−1σ
0
1σ
2σ
3σ
$\bar{x}$
estadígrafo



Tenemos incertidumbre respecto a cual es la muestra aleatoria que tenemos de la distribución de muestras aleatorias posibles.

Tampoco sabemos, cuando la muestra aleatoria que tenemos coincide con el centro de la distribución.

Lo único que sabemos, es que la muestra aleatoria que tenemos, es una de todas las muestras aleatorias posibles.

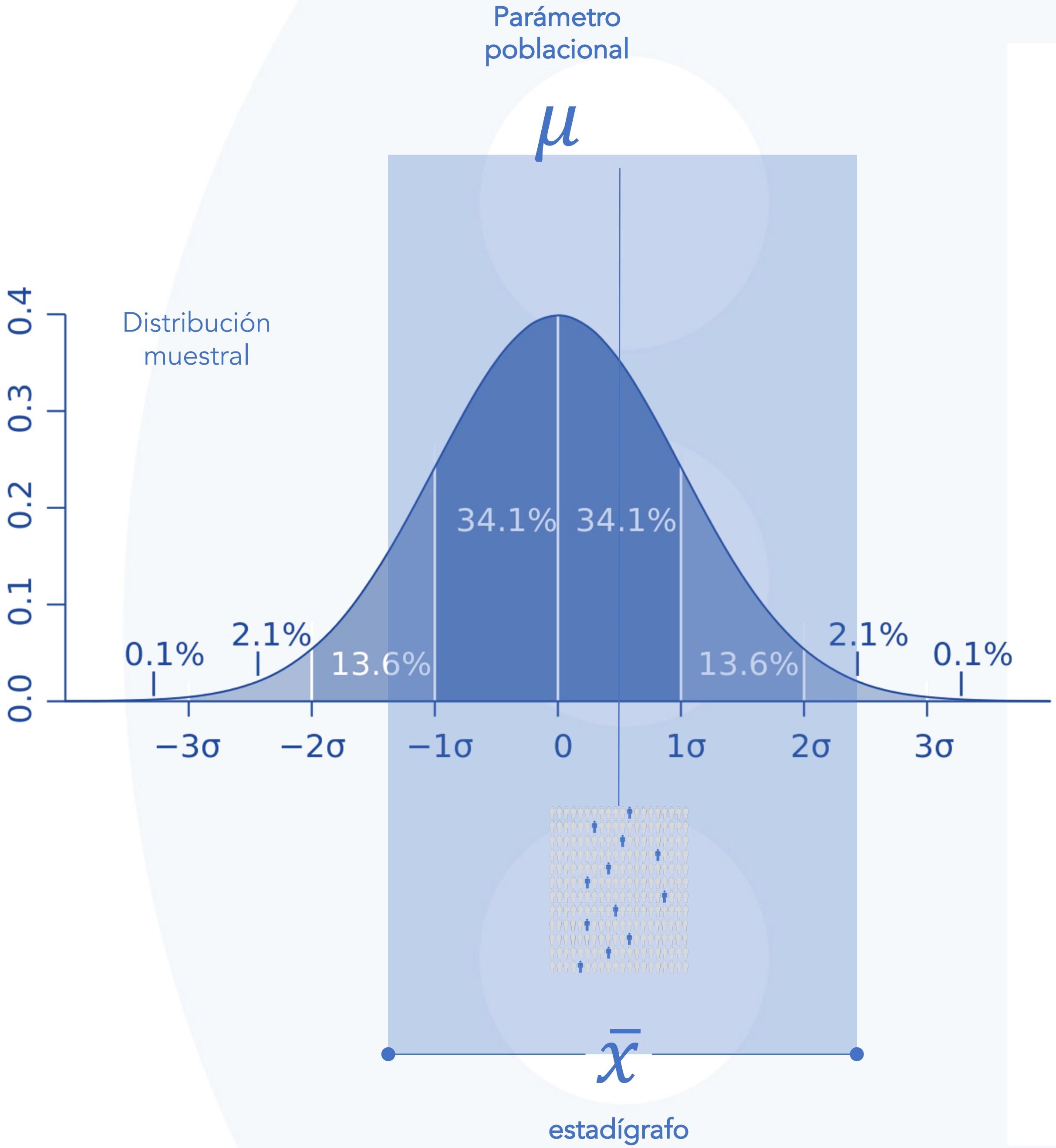


Tenemos incertidumbre respecto a cual es la muestra aleatoria que tenemos de la distribución de muestras aleatorias posibles.

Tampoco sabemos, cuando la muestra aleatoria que tenemos coincide con el centro de la distribución.

Lo único que sabemos, es que la muestra aleatoria que tenemos, es una de todas las muestras aleatorias posibles.

Lo que podemos hacer es establecer un rango de valores; un intervalo de confianza, el cual "capture" al parámetro de la población, una mayoría de las veces.

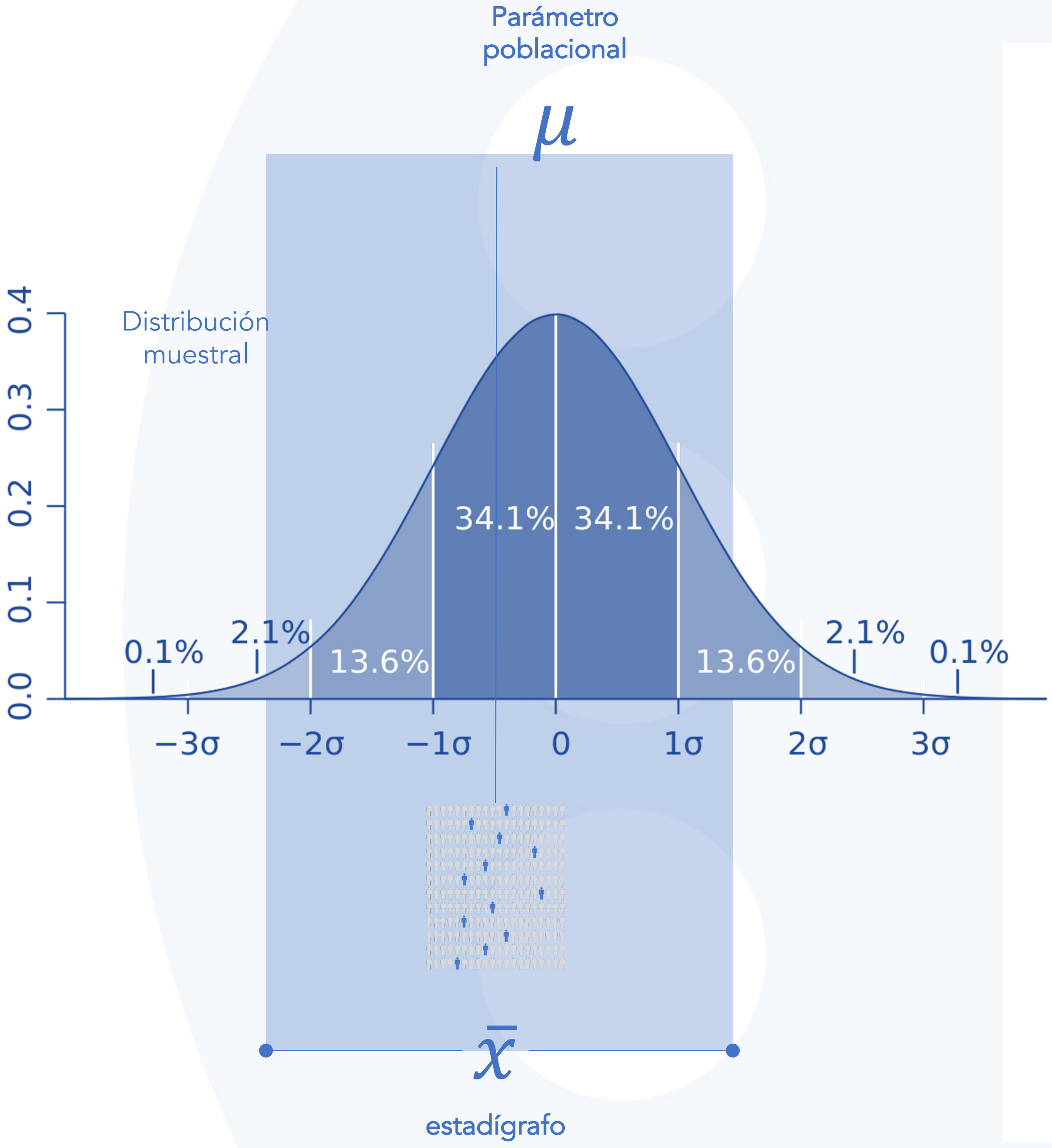


Tenemos incertidumbre respecto a cual es la muestra aleatoria que tenemos de la distribución de muestras aleatorias posibles.

Tampoco sabemos, cuando la muestra aleatoria que tenemos coincide con el centro de la distribución.

Lo único que sabemos, es que la muestra aleatoria que tenemos, es una de todas las muestras aleatorias posibles.

Lo que podemos hacer es establecer un rango de valores; un intervalo de confianza, el cual "capture" al parámetro de la población, una mayoría de las veces.

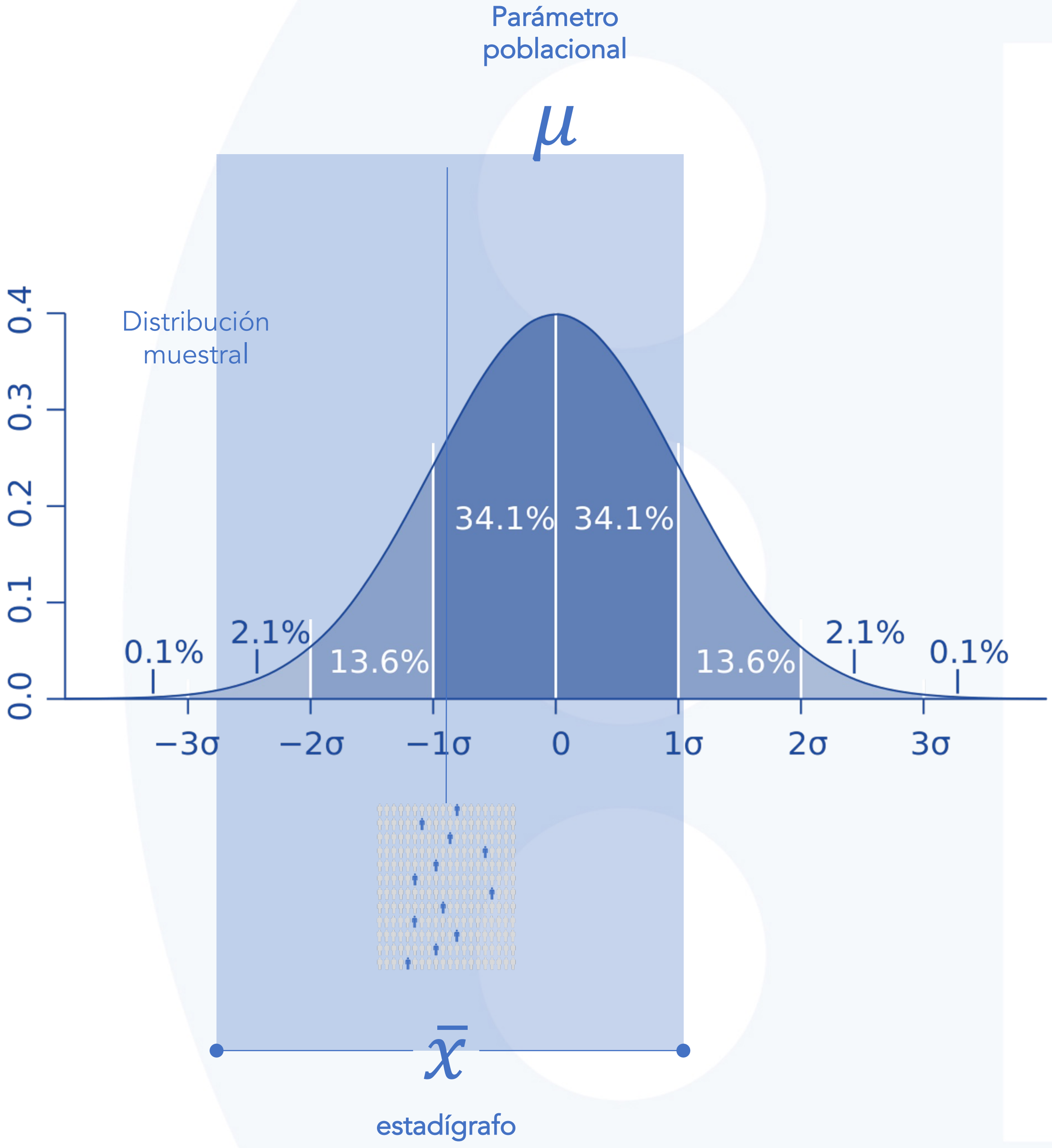


Tenemos incertidumbre respecto a cual es la muestra aleatoria que tenemos de la distribución de muestras aleatorias posibles.

Tampoco sabemos, cuando la muestra aleatoria que tenemos coincide con el centro de la distribución.

Lo único que sabemos, es que la muestra aleatoria que tenemos, es una de todas las muestras aleatorias posibles.

Lo que podemos hacer es establecer un rango de valores; un intervalo de confianza, el cual "capture" al parámetro de la población, una mayoría de las veces.

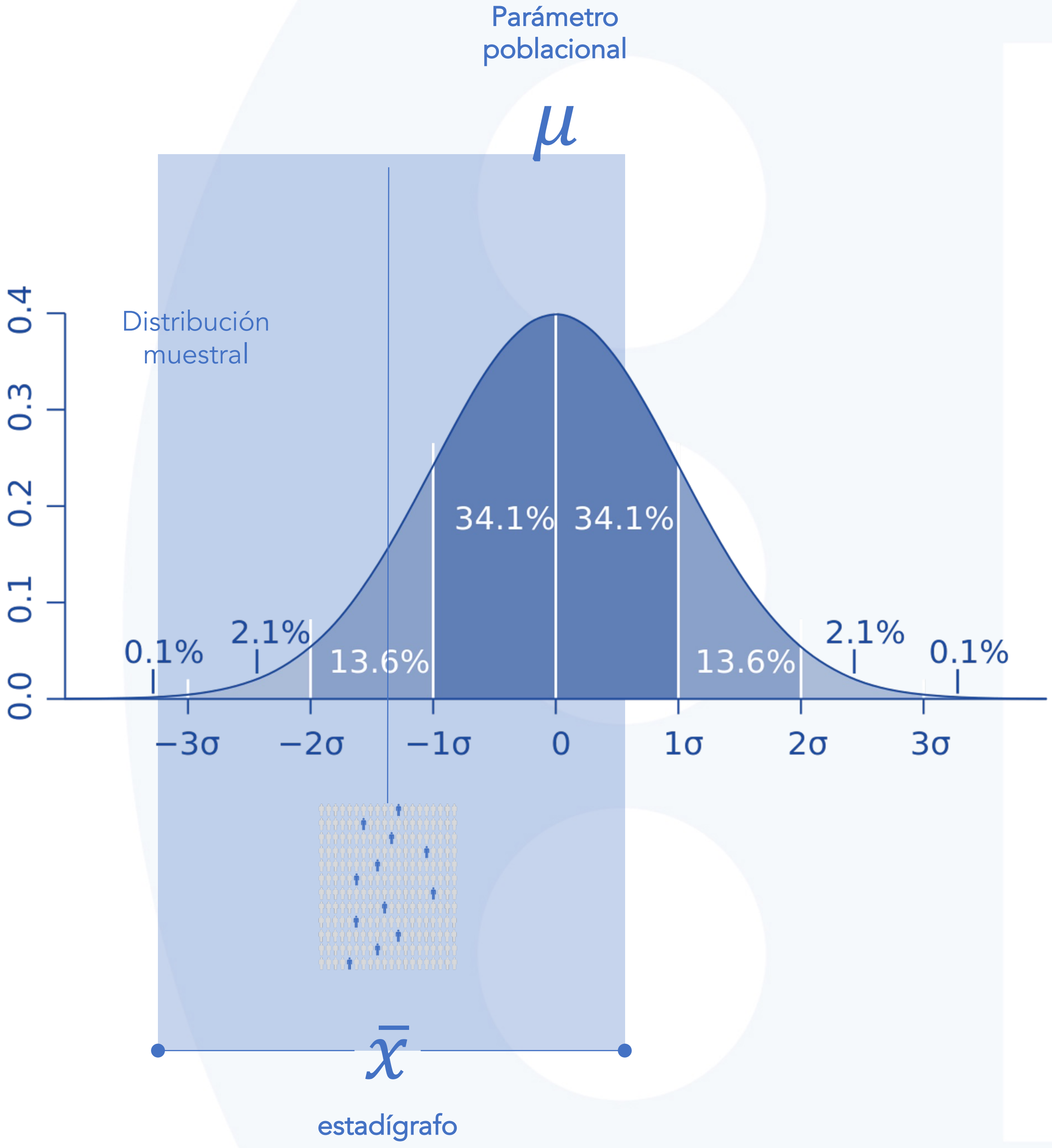


Tenemos incertidumbre respecto a cual es la muestra aleatoria que tenemos de la distribución de muestras aleatorias posibles.

Tampoco sabemos, cuando la muestra aleatoria que tenemos coincide con el centro de la distribución.

Lo único que sabemos, es que la muestra aleatoria que tenemos, es una de todas las muestras aleatorias posibles.

Lo que podemos hacer es establecer un rango de valores; un intervalo de confianza, el cual "capture" al parámetro de la población, una mayoría de las veces.

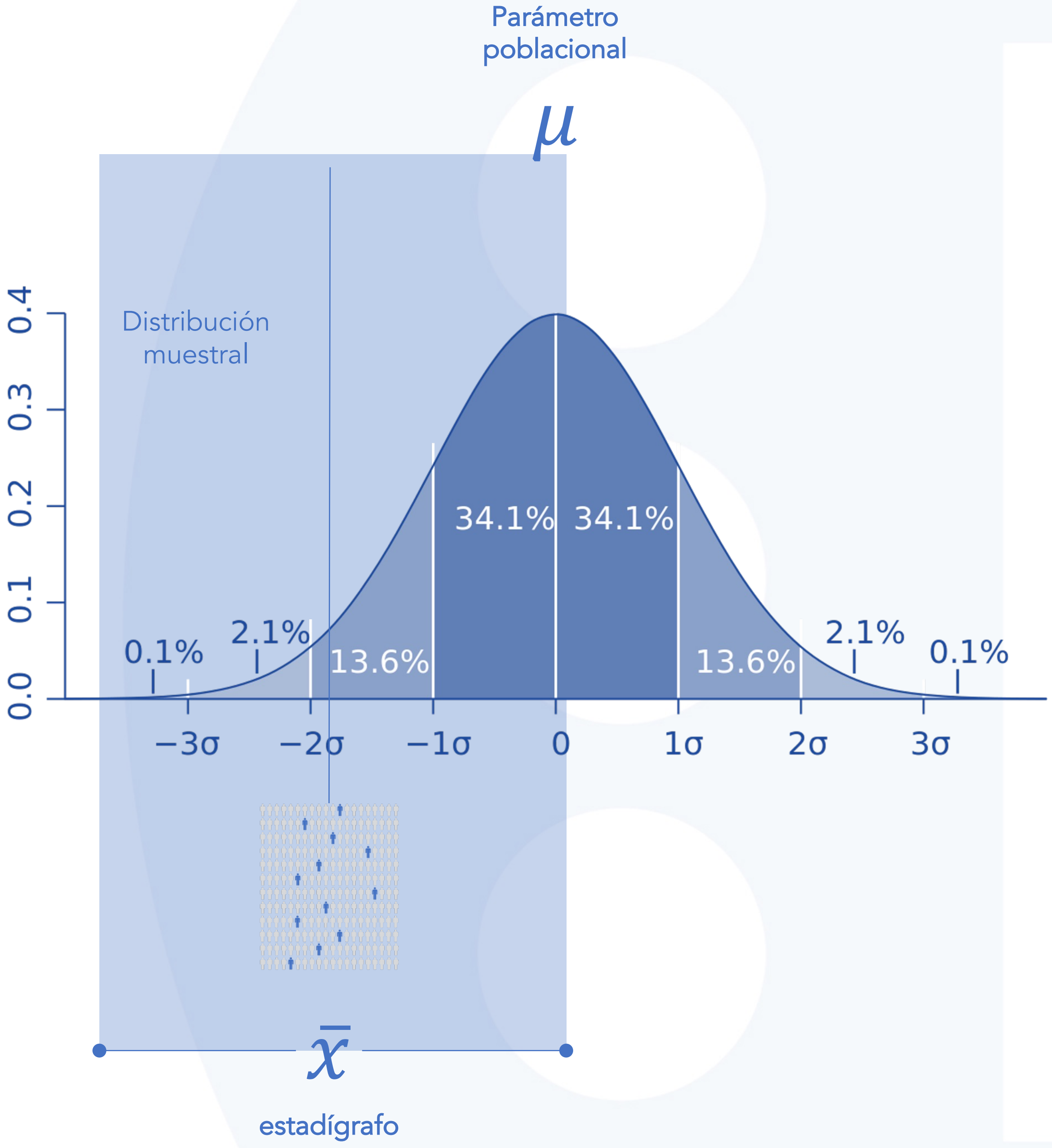


Tenemos incertidumbre respecto a cual es la muestra aleatoria que tenemos de la distribución de muestras aleatorias posibles.

Tampoco sabemos, cuando la muestra aleatoria que tenemos coincide con el centro de la distribución.

Lo único que sabemos, es que la muestra aleatoria que tenemos, es una de todas las muestras aleatorias posibles.

Lo que podemos hacer es establecer un rango de valores; un intervalo de confianza, el cual "capture" al parámetro de la población, una mayoría de las veces.

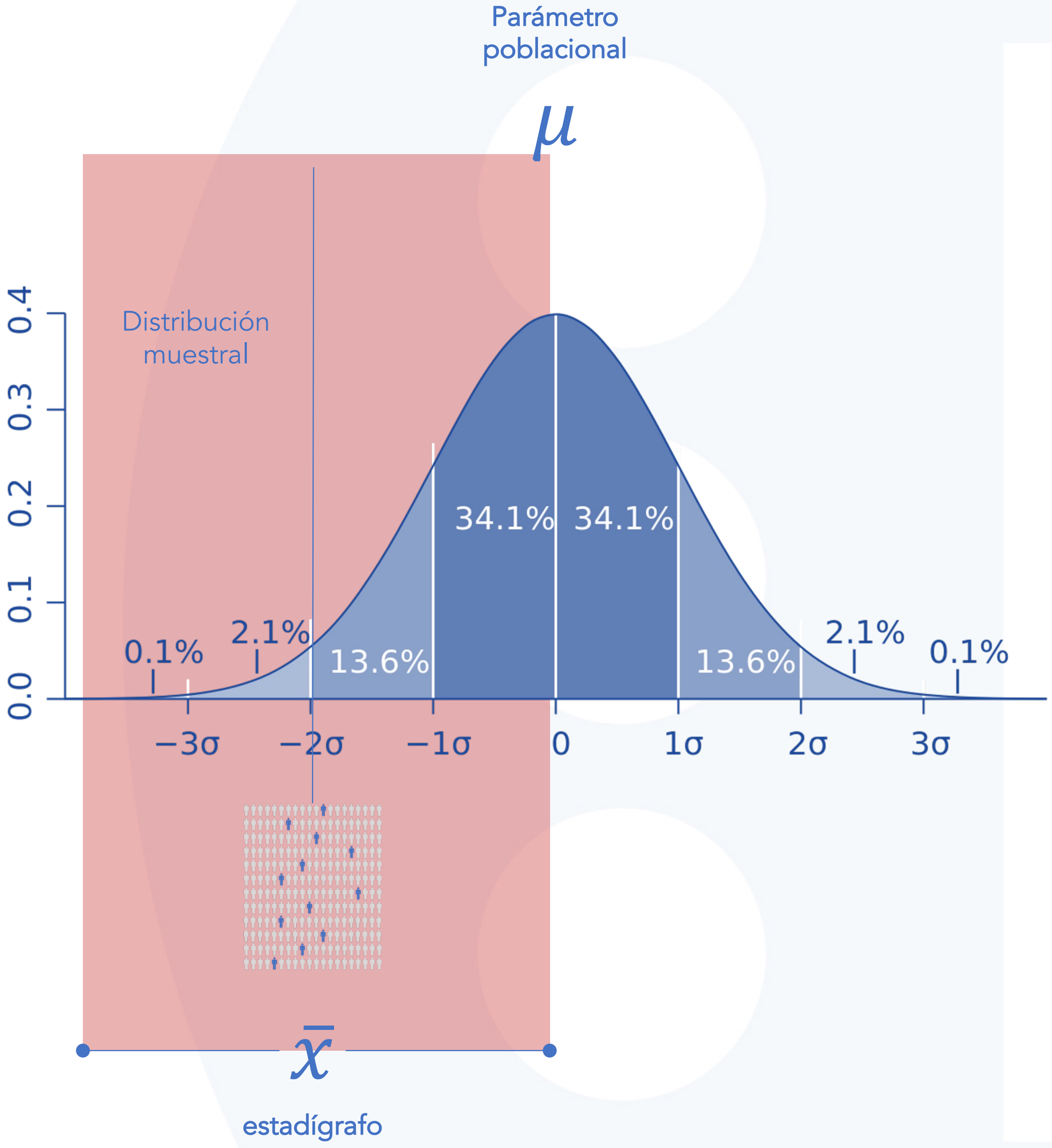


Tenemos incertidumbre respecto a cual es la muestra aleatoria que tenemos de la distribución de muestras aleatorias posibles.

Tampoco sabemos, cuando la muestra aleatoria que tenemos coincide con el centro de la distribución.

Lo único que sabemos, es que la muestra aleatoria que tenemos, es una de todas las muestras aleatorias posibles.

Lo que podemos hacer es establecer un rango de valores; un intervalo de confianza, el cual "capture" al parámetro de la población, una mayoría de las veces.



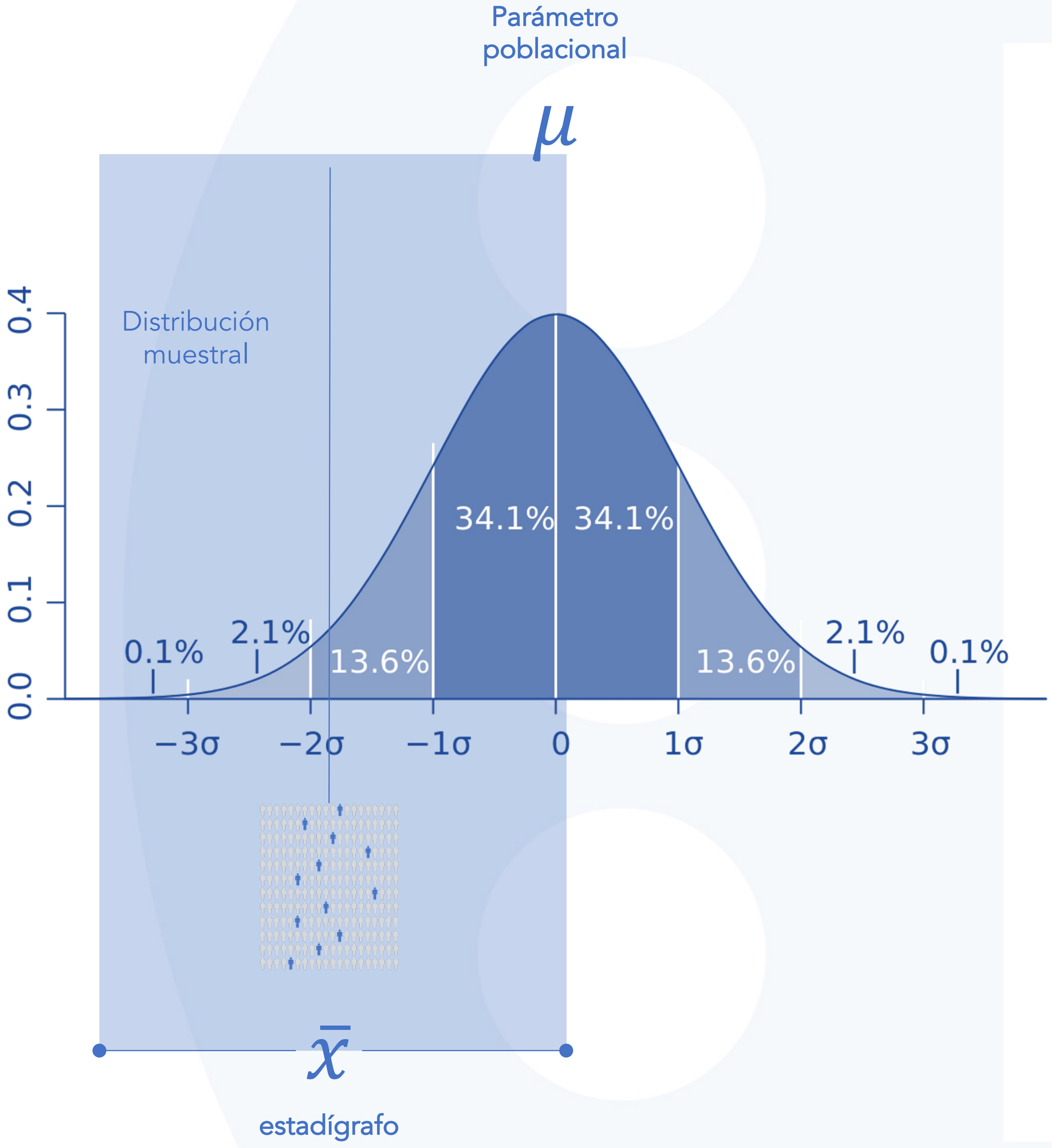
Tenemos incertidumbre respecto a cual es la muestra aleatoria que tenemos de la distribución de muestras aleatorias posibles.

Tampoco sabemos, cuando la muestra aleatoria que tenemos coincide con el centro de la distribución.

Lo único que sabemos, es que la muestra aleatoria que tenemos, es una de todas las muestras aleatorias posibles.

Lo que podemos hacer es establecer un rango de valores; un intervalo de confianza, el cual "capture" al parámetro de la población, una mayoría de las veces.

Si este rango es de 1.96 de Error estándar, entonces esperamos que solo un 5% de las veces, una muestra aleatoria no capture el parámetro poblacional.



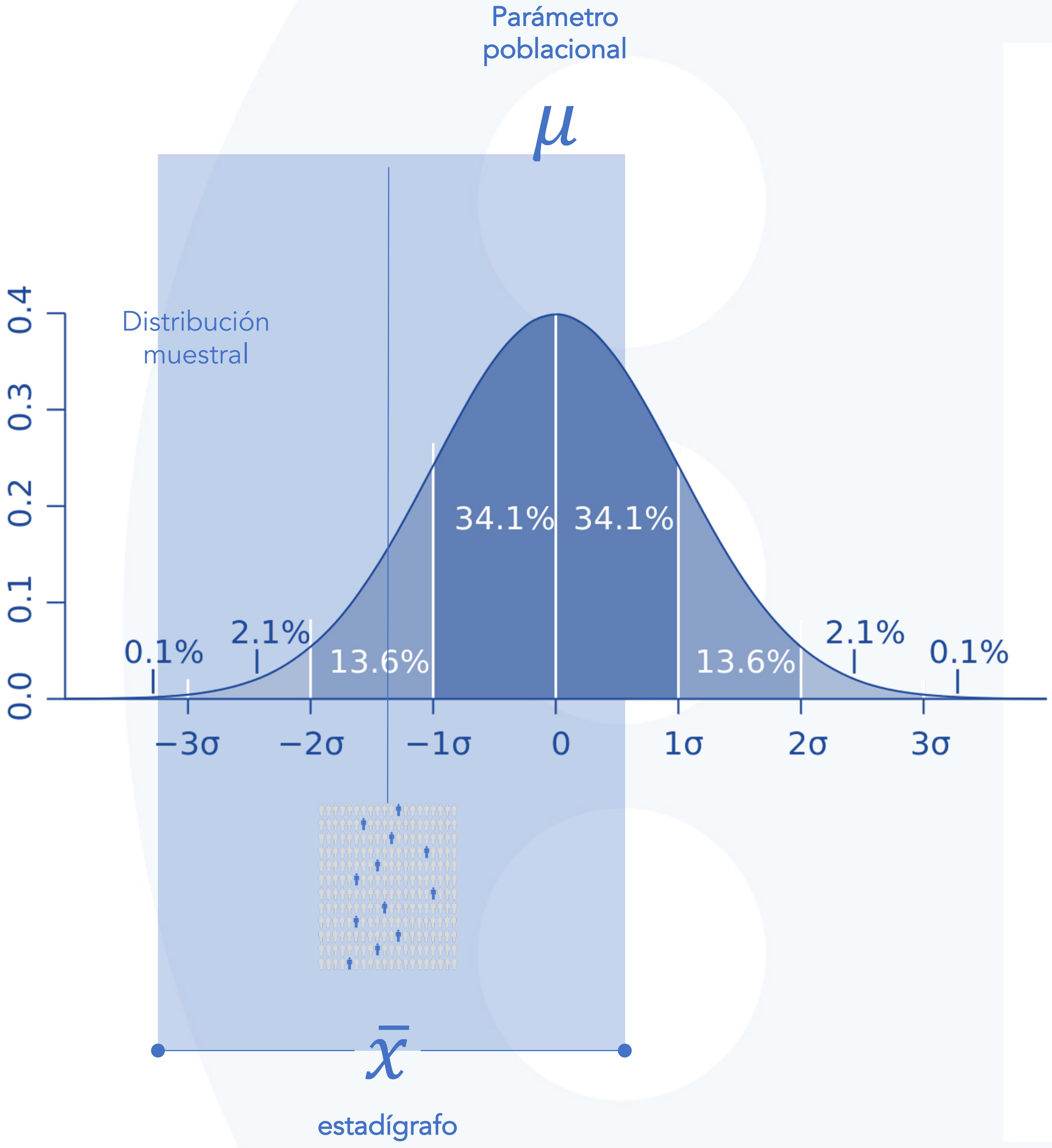
Tenemos incertidumbre respecto a cual es la muestra aleatoria que tenemos de la distribución de muestras aleatorias posibles.

Tampoco sabemos, cuando la muestra aleatoria que tenemos coincide con el centro de la distribución.

Lo único que sabemos, es que la muestra aleatoria que tenemos, es una de todas las muestras aleatorias posibles.

Lo que podemos hacer es establecer un rango de valores; un intervalo de confianza, el cual "capture" al parámetro de la población, una mayoría de las veces.

Si este rango es de 1.96 de Error estándar, entonces esperamos que solo un 5% de las veces, una muestra aleatoria no capture el parámetro poblacional.



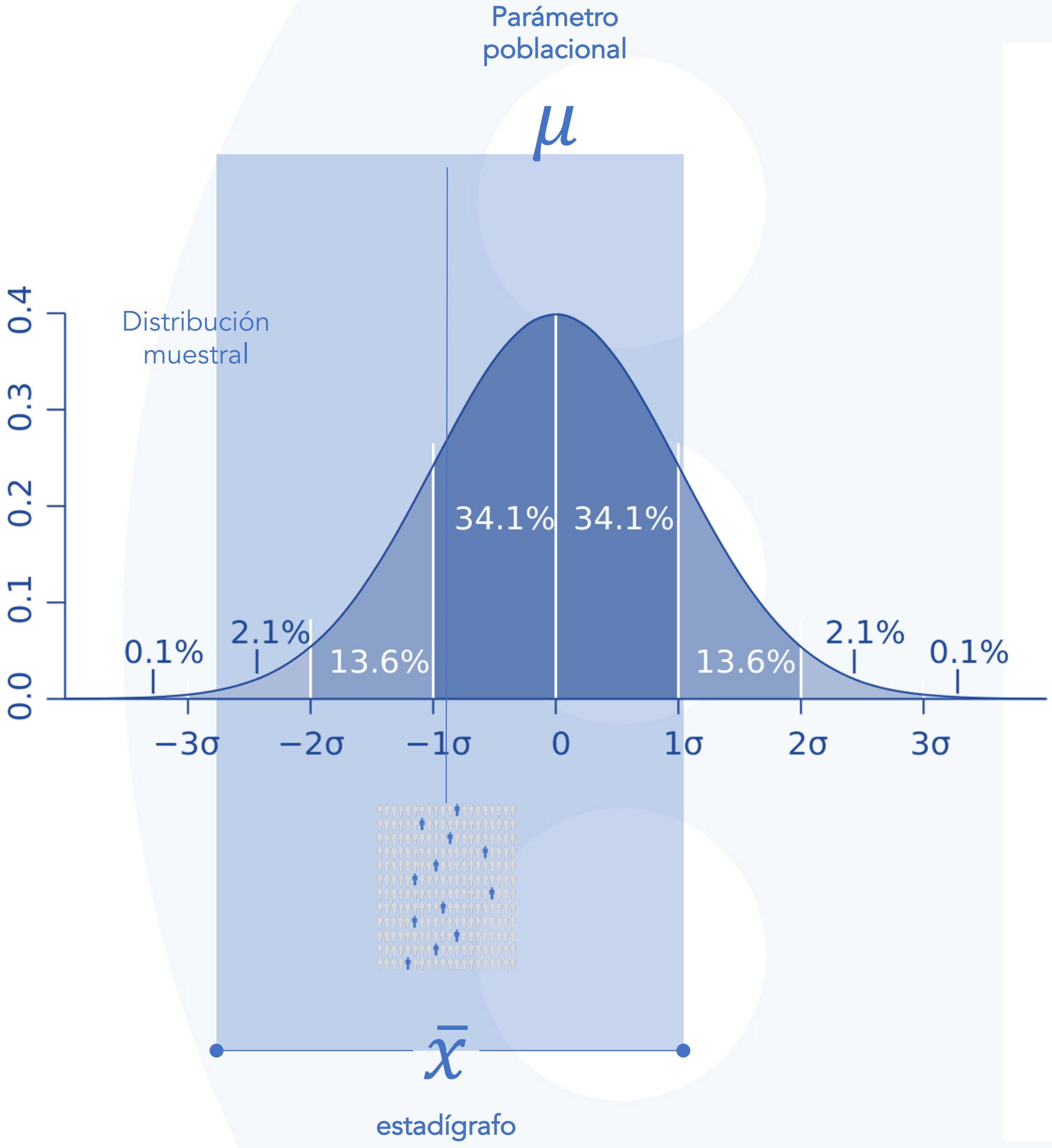
Tenemos incertidumbre respecto a cual es la muestra aleatoria que tenemos de la distribución de muestras aleatorias posibles.

Tampoco sabemos, cuando la muestra aleatoria que tenemos coincide con el centro de la distribución.

Lo único que sabemos, es que la muestra aleatoria que tenemos, es una de todas las muestras aleatorias posibles.

Lo que podemos hacer es establecer un rango de valores; un intervalo de confianza, el cual "capture" al parámetro de la población, una mayoría de las veces.

Si este rango es de 1.96 de Error estándar, entonces esperamos que solo un 5% de las veces, una muestra aleatoria no capture el parámetro poblacional.



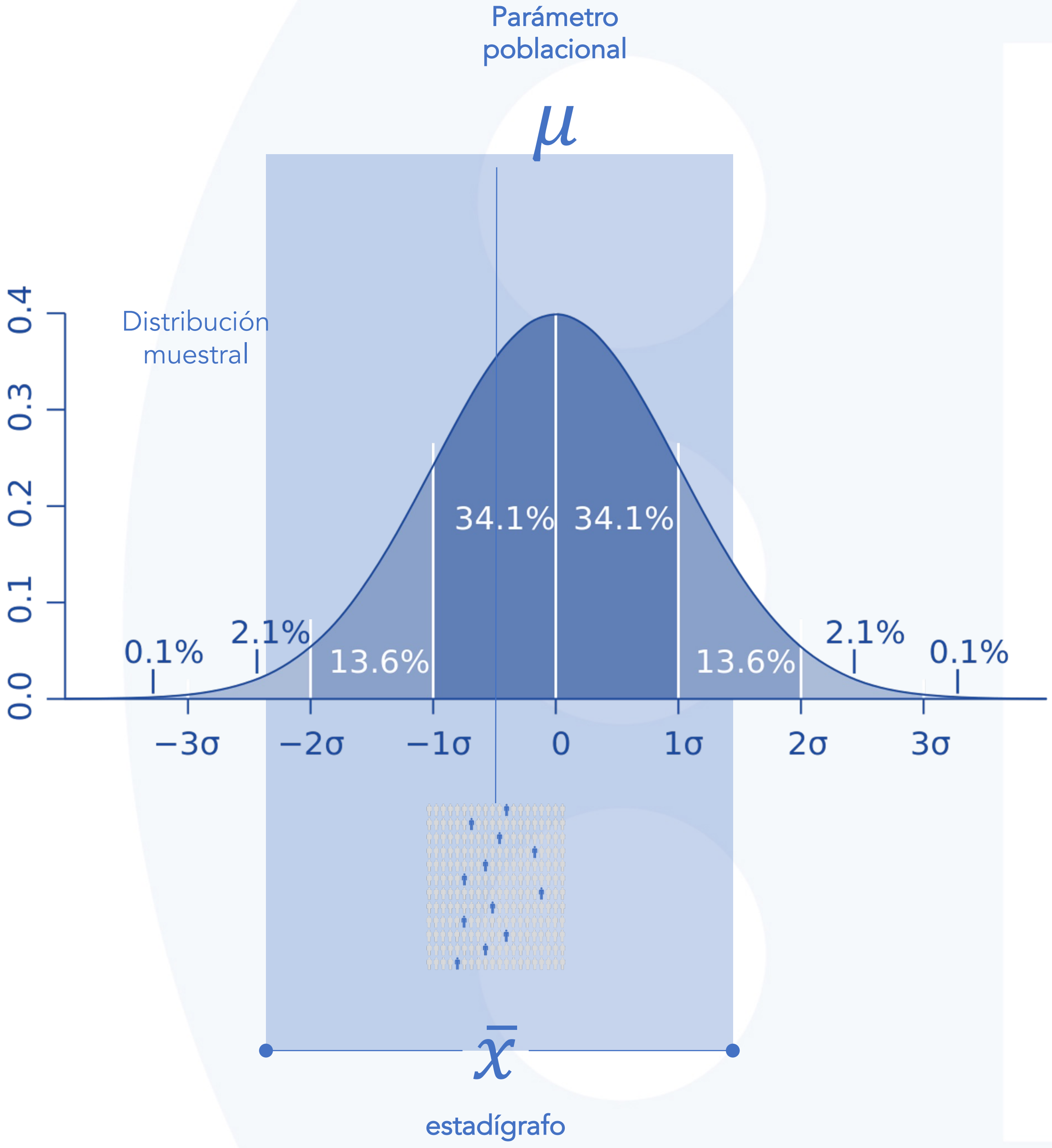
Tenemos incertidumbre respecto a cual es la muestra aleatoria que tenemos de la distribución de muestras aleatorias posibles.

Tampoco sabemos, cuando la muestra aleatoria que tenemos coincide con el centro de la distribución.

Lo único que sabemos, es que la muestra aleatoria que tenemos, es una de todas las muestras aleatorias posibles.

Lo que podemos hacer es establecer un rango de valores; un intervalo de confianza, el cual "capture" al parámetro de la población, una mayoría de las veces.

Si este rango es de 1.96 de Error estándar, entonces esperamos que solo un 5% de las veces, una muestra aleatoria no capture el parámetro poblacional.



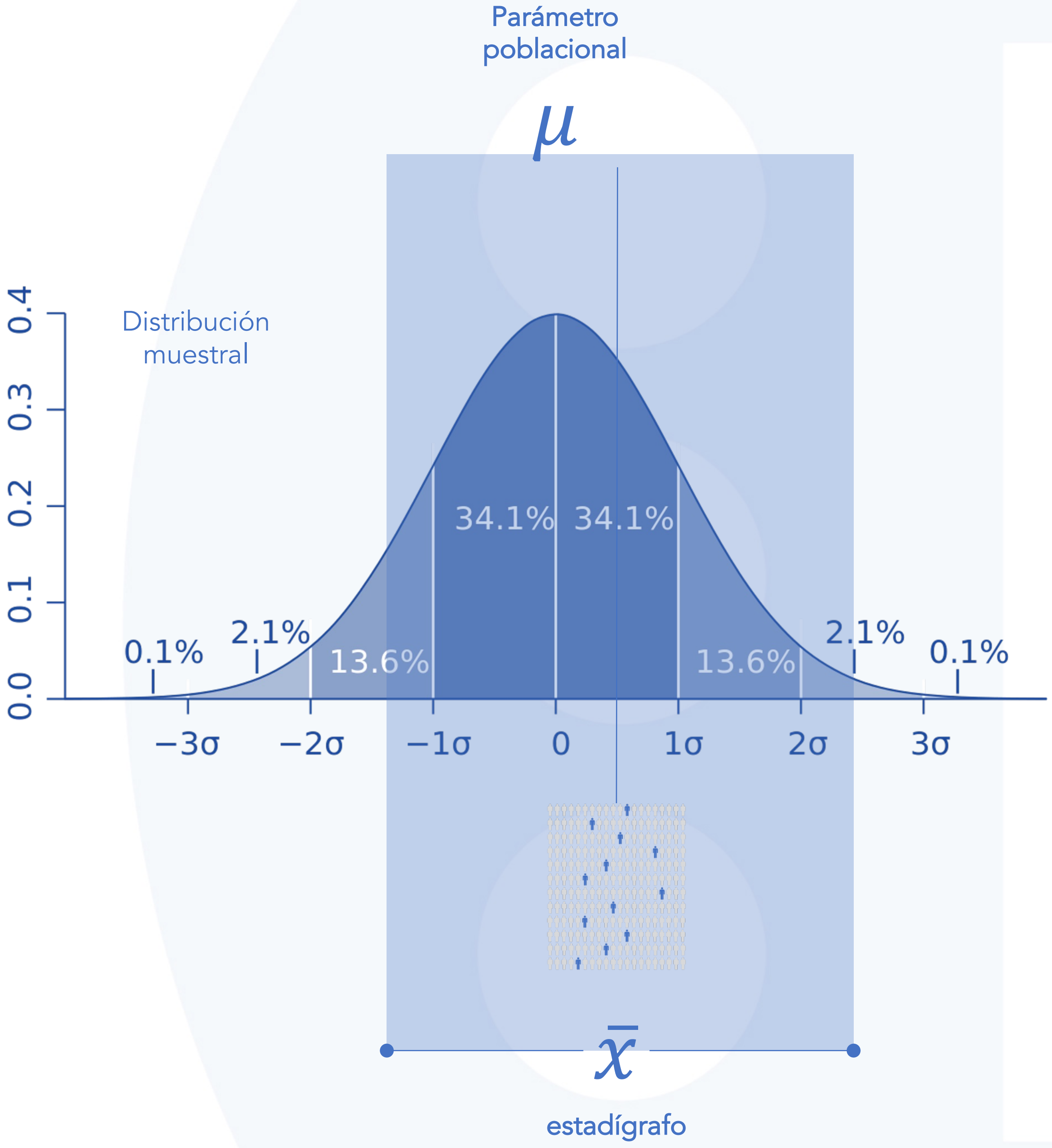
Tenemos incertidumbre respecto a cual es la muestra aleatoria que tenemos de la distribución de muestras aleatorias posibles.

Tampoco sabemos, cuando la muestra aleatoria que tenemos coincide con el centro de la distribución.

Lo único que sabemos, es que la muestra aleatoria que tenemos, es una de todas las muestras aleatorias posibles.

Lo que podemos hacer es establecer un rango de valores; un intervalo de confianza, el cual "capture" al parámetro de la población, una mayoría de las veces.

Si este rango es de 1.96 de Error estándar, entonces esperamos que solo un 5% de las veces, una muestra aleatoria no capture el parámetro poblacional.



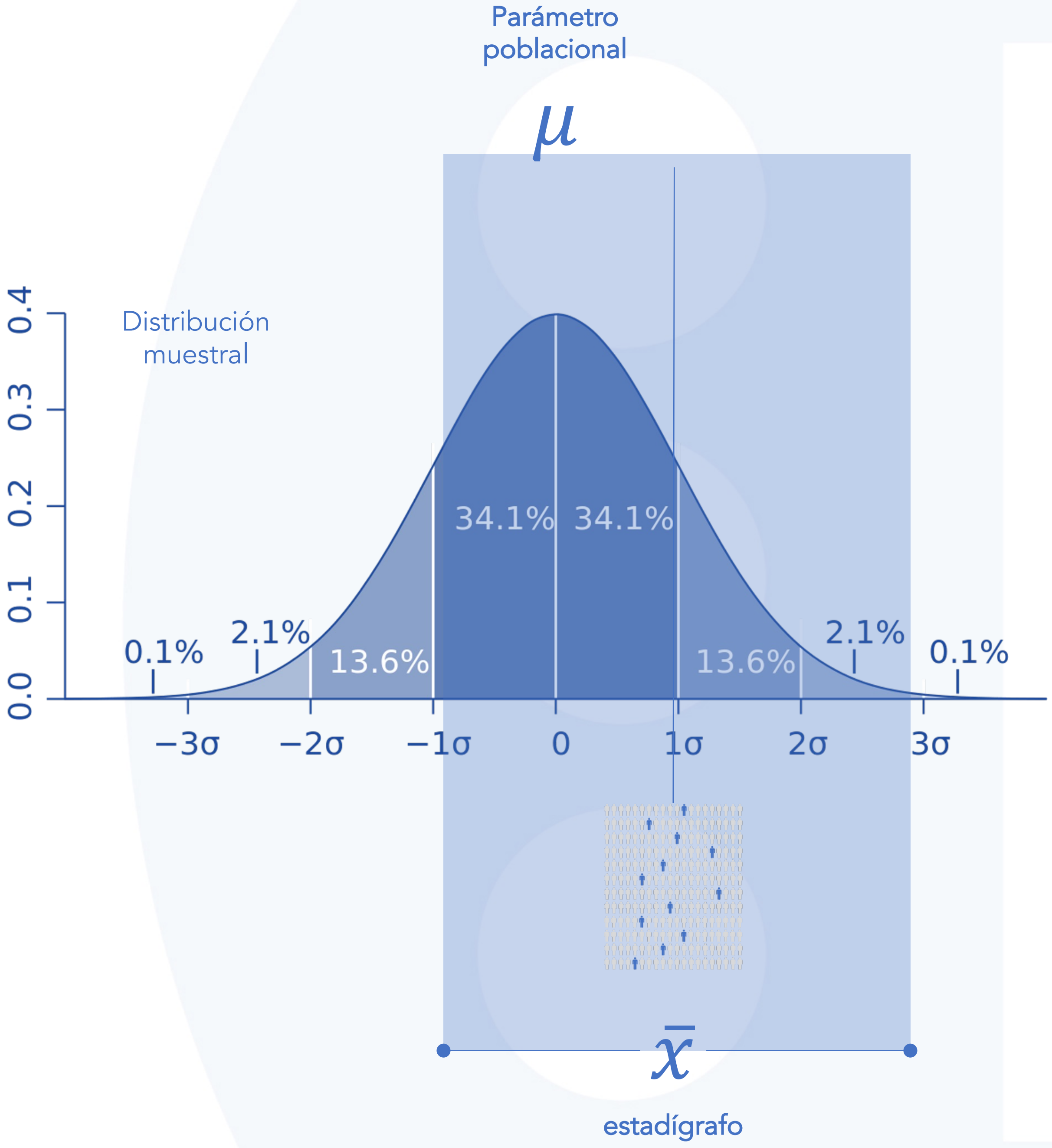
Tenemos incertidumbre respecto a cual es la muestra aleatoria que tenemos de la distribución de muestras aleatorias posibles.

Tampoco sabemos, cuando la muestra aleatoria que tenemos coincide con el centro de la distribución.

Lo único que sabemos, es que la muestra aleatoria que tenemos, es una de todas las muestras aleatorias posibles.

Lo que podemos hacer es establecer un rango de valores; un intervalo de confianza, el cual "capture" al parámetro de la población, una mayoría de las veces.

Si este rango es de 1.96 de Error estándar, entonces esperamos que solo un 5% de las veces, una muestra aleatoria no capture el parámetro poblacional.



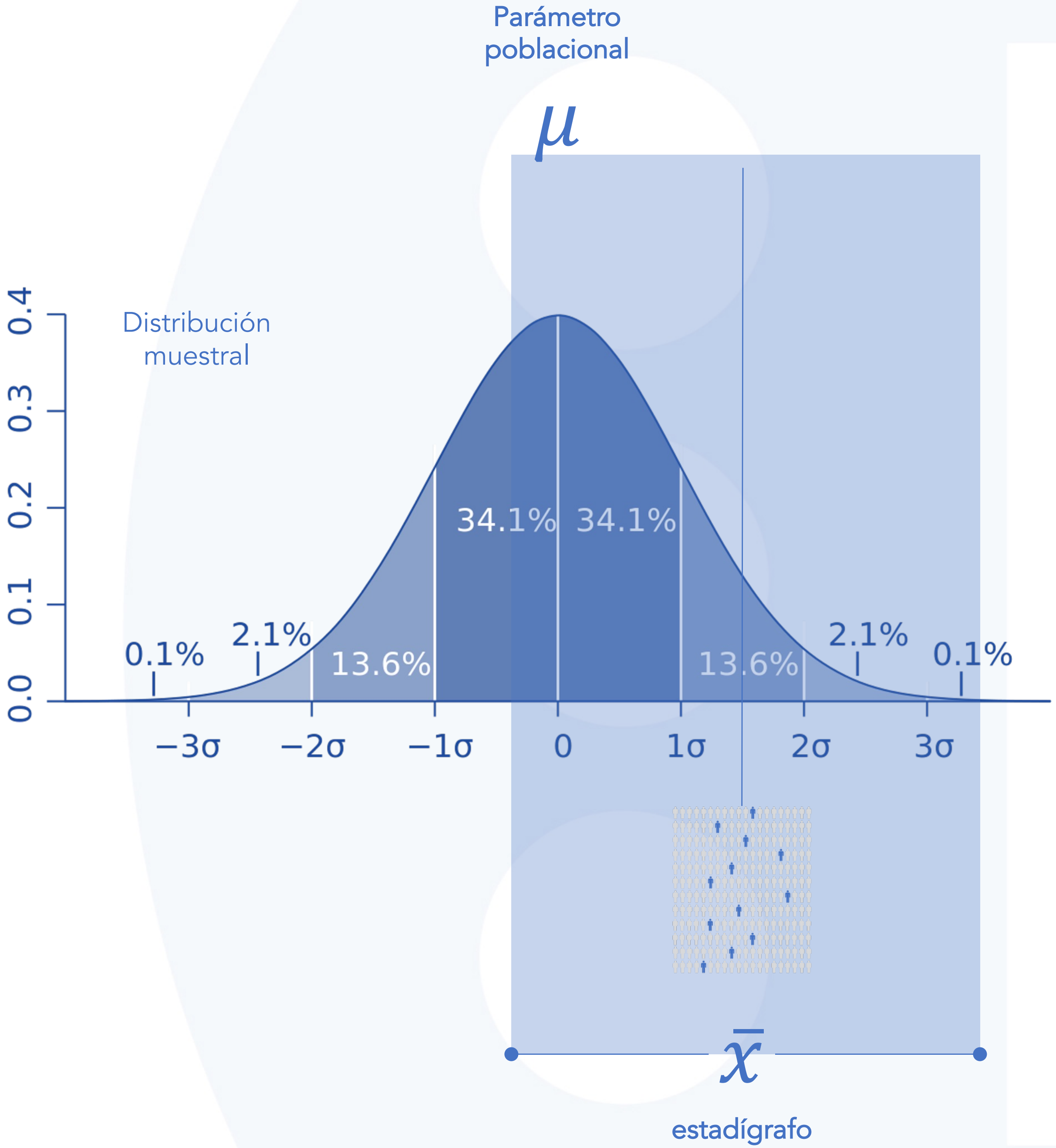
Tenemos incertidumbre respecto a cual es la muestra aleatoria que tenemos de la distribución de muestras aleatorias posibles.

Tampoco sabemos, cuando la muestra aleatoria que tenemos coincide con el centro de la distribución.

Lo único que sabemos, es que la muestra aleatoria que tenemos, es una de todas las muestras aleatorias posibles.

Lo que podemos hacer es establecer un rango de valores; un intervalo de confianza, el cual "capture" al parámetro de la población, una mayoría de las veces.

Si este rango es de 1.96 de Error estándar, entonces esperamos que solo un 5% de las veces, una muestra aleatoria no capture el parámetro poblacional.



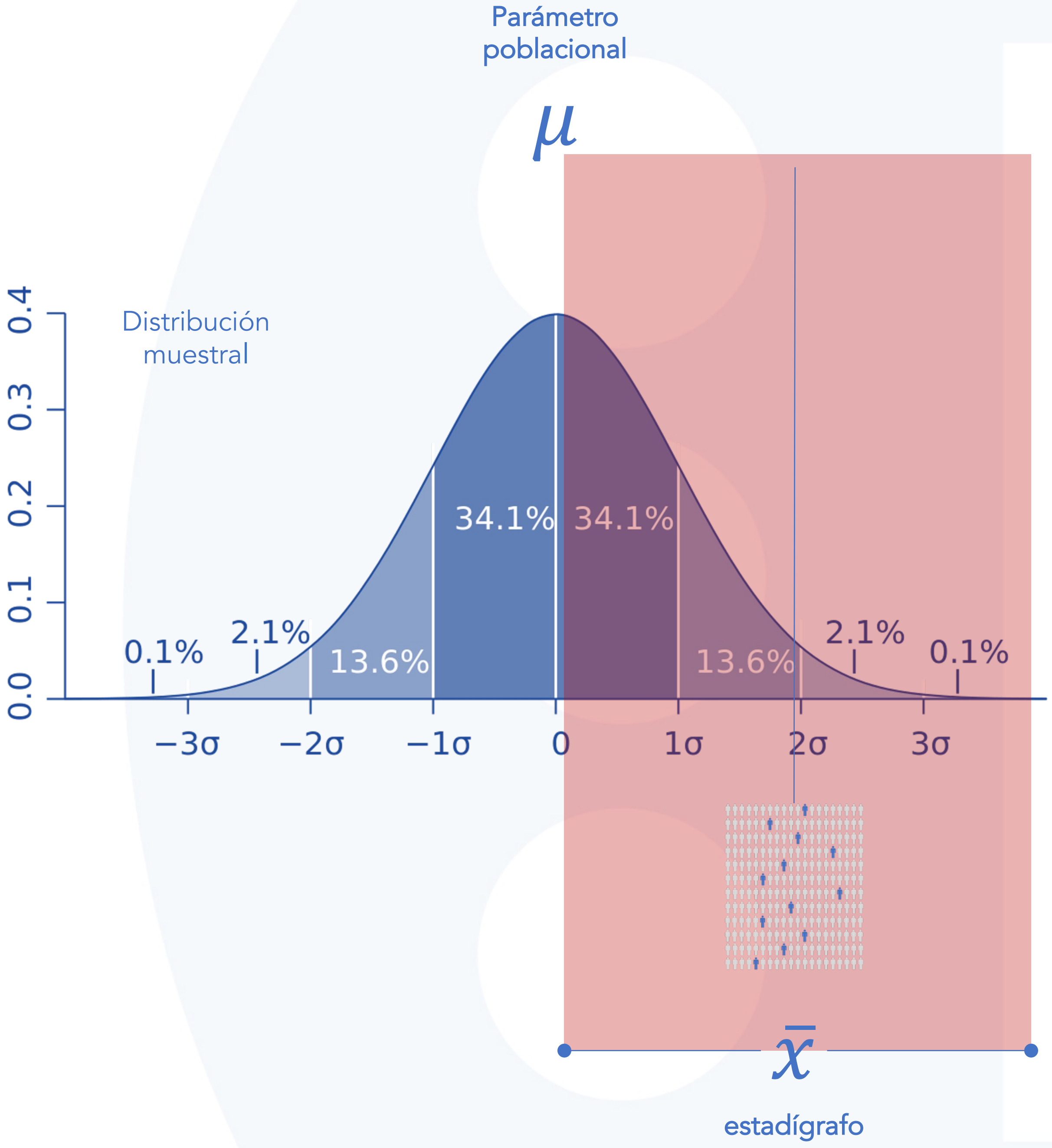
Tenemos incertidumbre respecto a cual es la muestra aleatoria que tenemos de la distribución de muestras aleatorias posibles.

Tampoco sabemos, cuando la muestra aleatoria que tenemos coincide con el centro de la distribución.

Lo único que sabemos, es que la muestra aleatoria que tenemos, es una de todas las muestras aleatorias posibles.

Lo que podemos hacer es establecer un rango de valores; un intervalo de confianza, el cual "capture" al parámetro de la población, una mayoría de las veces.

Si este rango es de 1.96 de Error estándar, entonces esperamos que solo un 5% de las veces, una muestra aleatoria no capture el parámetro poblacional.



Tenemos incertidumbre respecto a cual es la muestra aleatoria que tenemos de la distribución de muestras aleatorias posibles.

Tampoco sabemos, cuando la muestra aleatoria que tenemos coincide con el centro de la distribución.

Lo único que sabemos, es que la muestra aleatoria que tenemos, es una de todas las muestras aleatorias posibles.

Lo que podemos hacer es establecer un rango de valores; un intervalo de confianza, el cual "capture" al parámetro de la población, una mayoría de las veces.

Si este rango es de 1.96 de Error estándar, entonces **esperamos que solo un 5% de las veces, una muestra aleatoria no capture el parámetro poblacional.**

Metodología Cuantitativa

Inferencia interval

Intervalos de confianza

Inferencia por medio de Intervalos de confianza

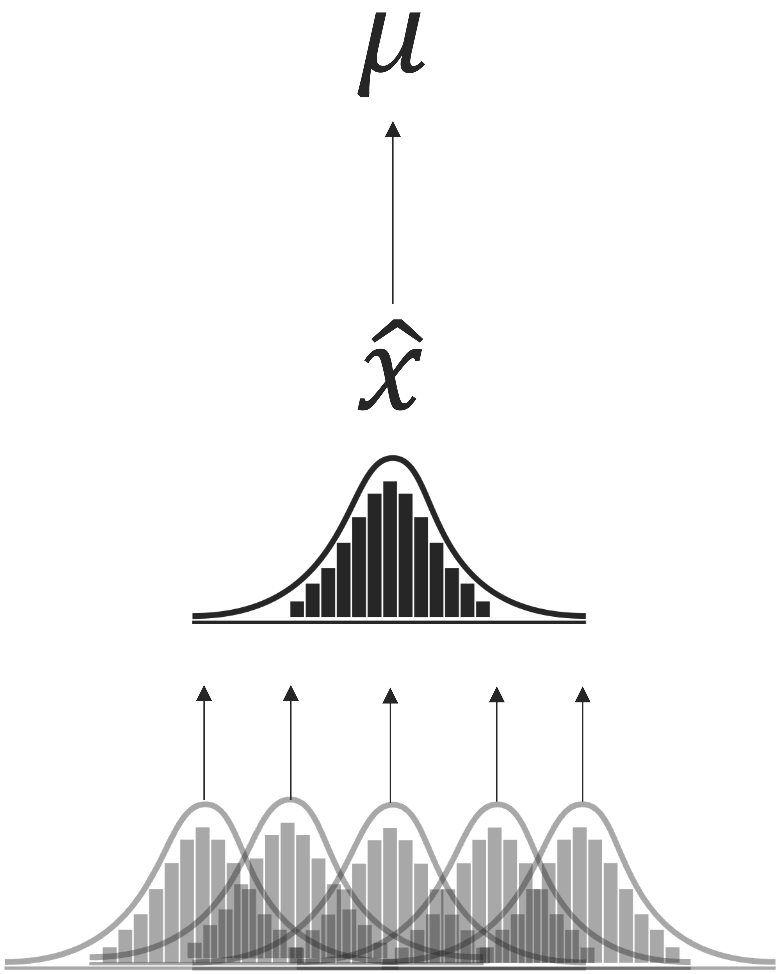
Una muestra aleatoria, puede contener el parámetro de interés, o puede no contenerlo

Dado que la distribución muestral es normal, entonces, podemos emplear a la distribución normal estandarizada, una **distribución probabilística,** para realizar inferencias.

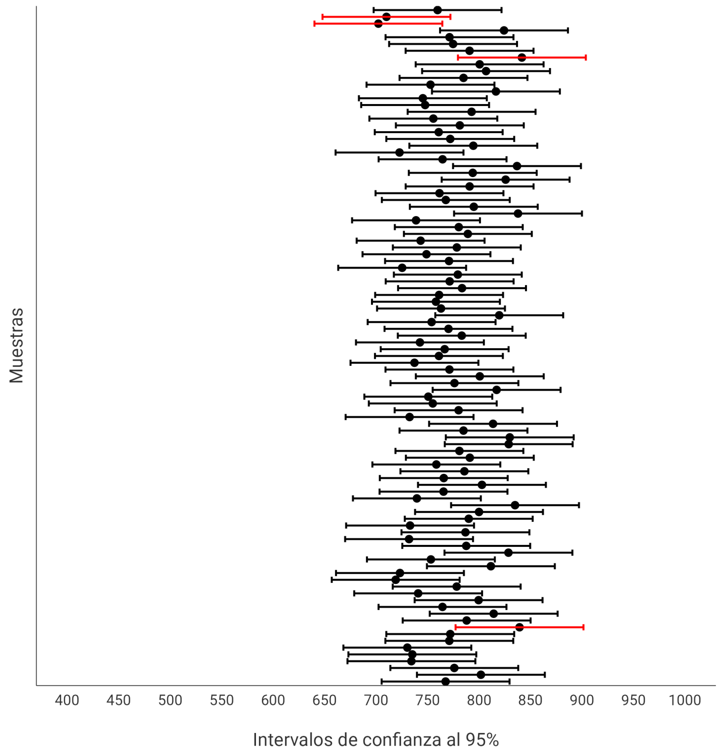
Empleando la información de la distribución normal estandarizada, podemos crear intervalos, que un 95% de las veces contengan al parámetro poblacional.

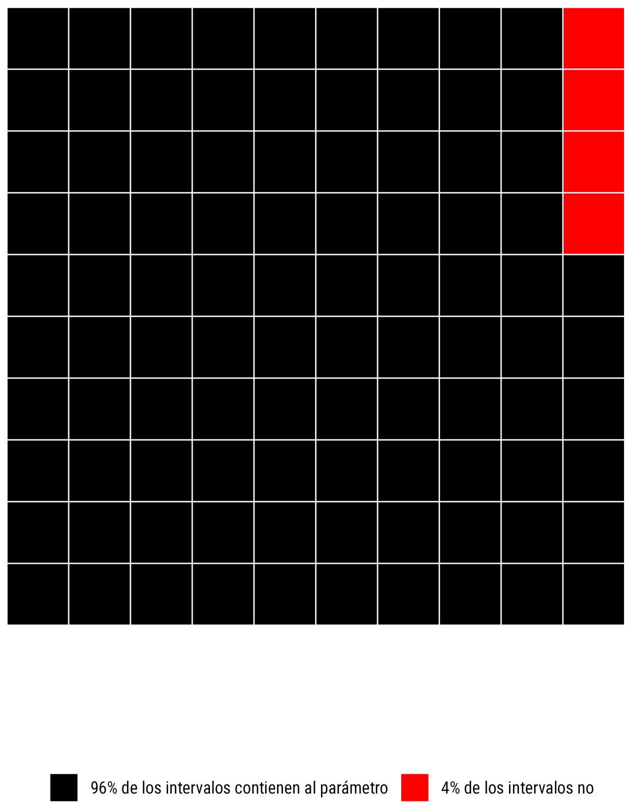
$$CI95\% \ [\bar{x} \ - 1.96\frac{\sigma}{\sqrt{n}} \ ; \bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \ ]$$

Realizar inferencias poblacionales con este método, consiste en indicar que el parámetro de la población (e.g., medias, proporciones), podría encontrarse en un rango de valores.



Distribución Muestral de 30 observaciones



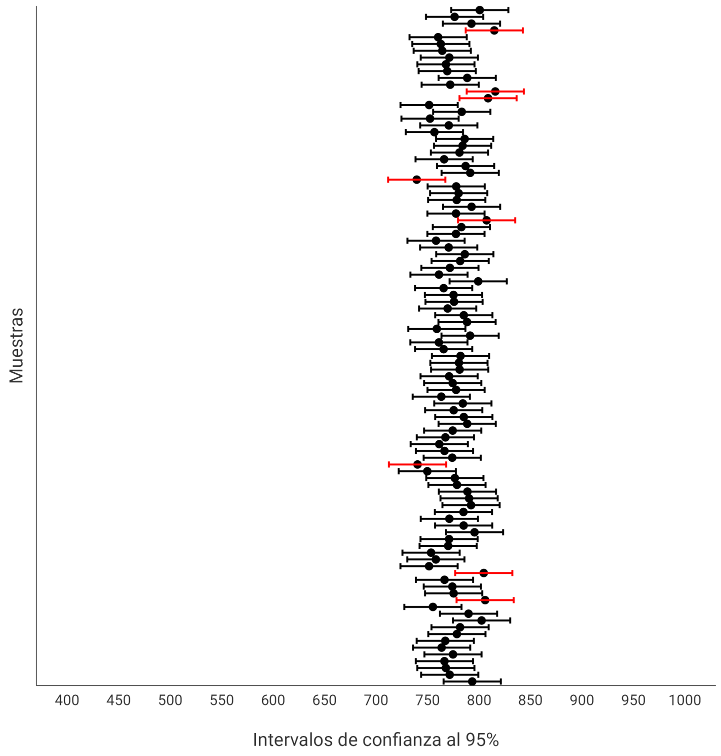


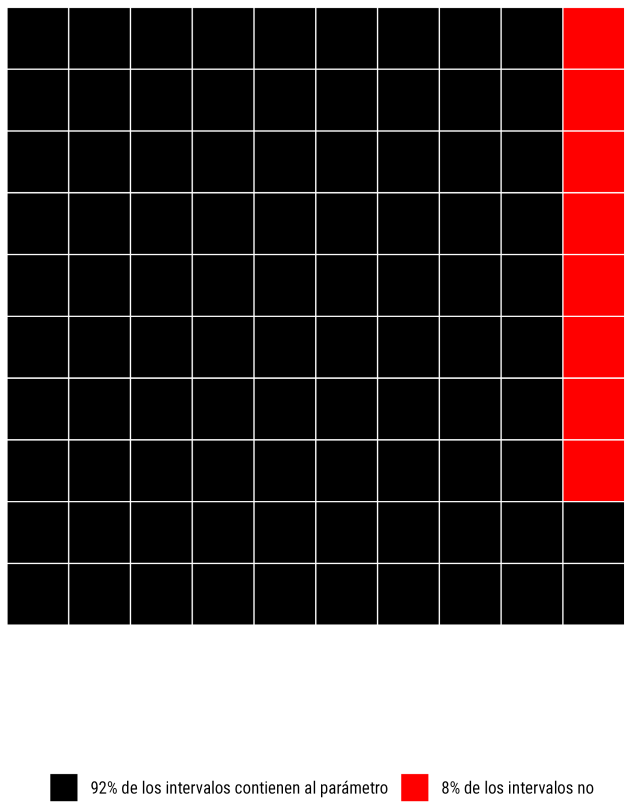
Generamos una distribución muestral de 30 observaciones, con 100 replicas. Calculamos el límite inferior y superior de un intervalo de 95% de confianza al rededor de la media de cada muestra.

Luego, clasificamos cada uno de estos intervalos según si contiene o no, al parámetro de la población.

Un intervalo, contiene al parámetro de la población, si el límite inferior es menor al parámetro de la población, y si el límite superior es mayor al parámetro poblacional.

Distribución Muestral de 50 observaciones

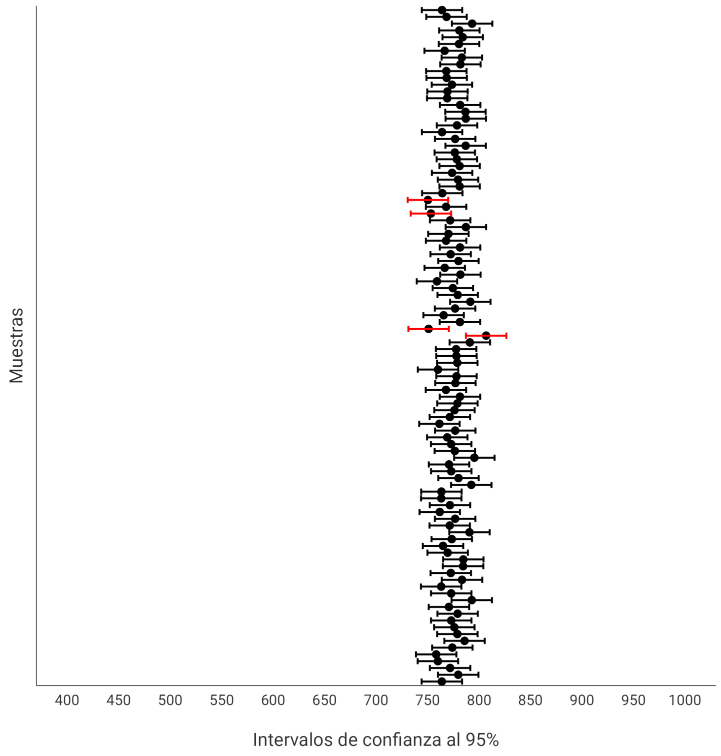


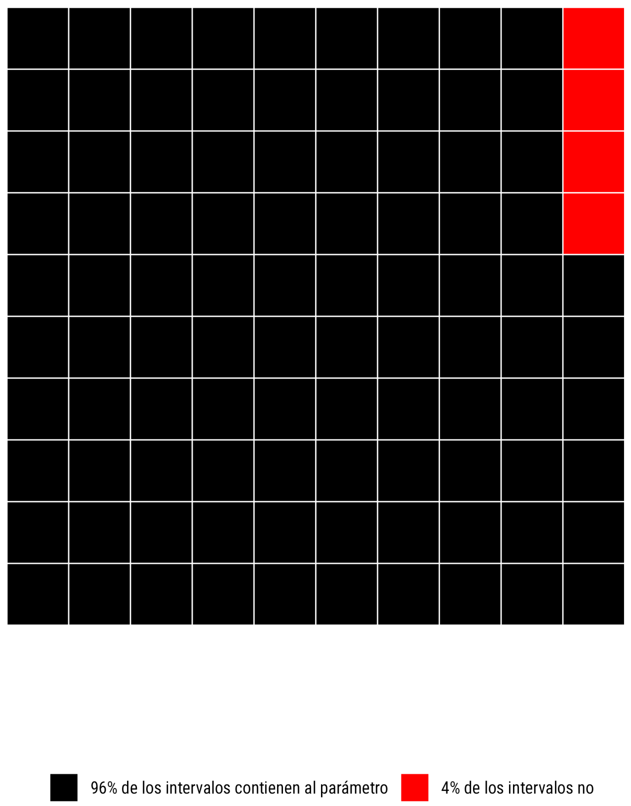


A medida que empleamos muestras de mayor tamaño, por ejemplo de 50 observaciones, en vez de 30 observaciones, los intervalos de confianza son más estrechos. En otras palabras, el rango intervalar es más pequeño.

En estas simulaciones, intervalos pueden capturar al parámetro poblacional unas 92% a 96%. No obstante, si aumentáramos la cantidad de muestras de la distribución muestral, por ejemplo a 10000 muestras, la expectativa es que el 95% de los intervalos generados capturaría al parámetro de la población.

Distribución Muestral de 100 observaciones



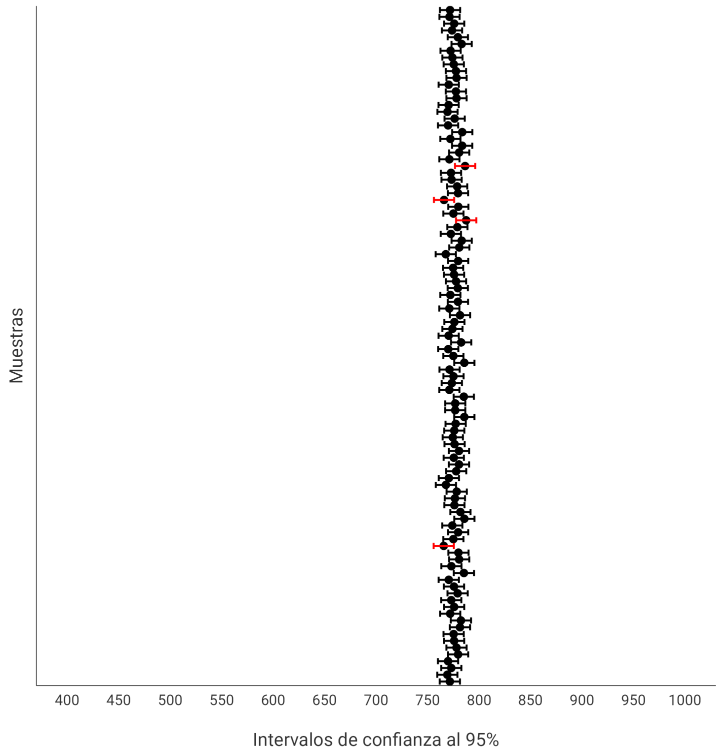


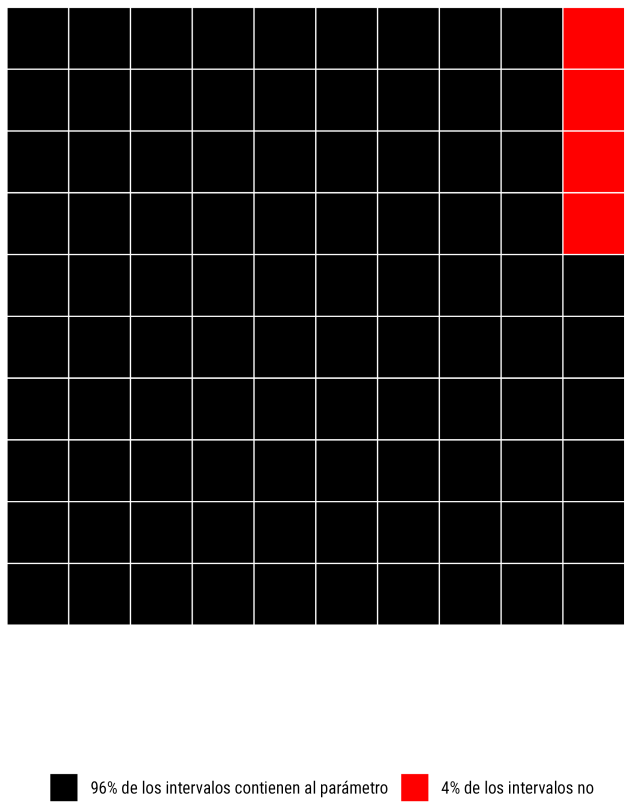
Nótese que, aunque la distribución muestral tuviera 30, 50, 100 o 500 casos, dado que la distribución muestral es normal, estos intervalos capturan al parámetro poblacional solo un 95% de las veces.

En otras palabras, el tamaño muestral no influye sobre la cantidad de veces en que un intervalo de 95% de confianza va a capturar al parámetro poblacional.

A partir de los 30, la distribución muestral, toma una forma normal. De esta forma, a partir de este tamaño, podemos construir intervalos de esta forma.

Distribución Muestral de 400 observaciones



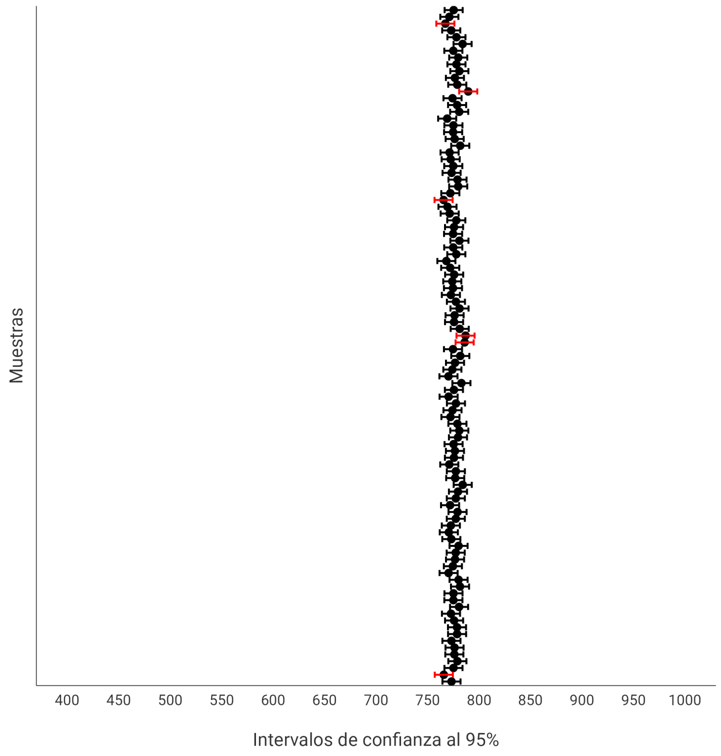


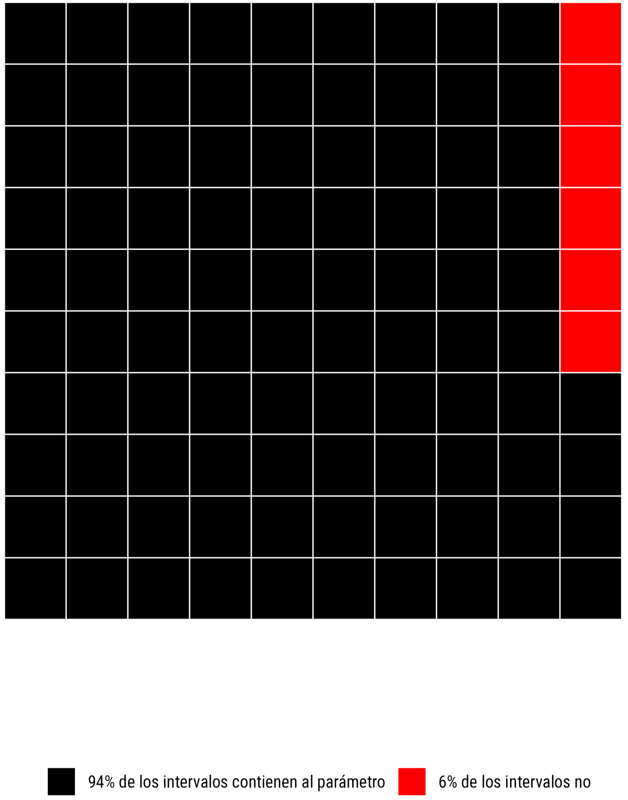
Para crear intervalos de confianza de 95% empleamos el puntaje z = 1.96.

Bajo -1.96 se encuentran un 2.5% de los valores, de esta distribución probabilística. Sobre 1.96 se encuentra otro 2.5% de todos los valores posibles.

Multiplicamos este valor, llamémosle valor z crítico, por la desviación estándar de la distribución muestral. Dicho de otra forma. Multiplicamos el valor crítico, por la el error estándar de la media. Con esta operación, obtenemos el margen de error.

Distribución Muestral de 500 observaciones





El margen de error es luego, sumado y restado sobre la media de cada muestra aleatoria generada.

Con lo anterior, generamos dos valores: el limite inferior del intervalo, y el límite superior del intervalo.

En términos generales, estamos empleando la siguiente formula:

$$CI95\%$$

$$[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} ; \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}]$$

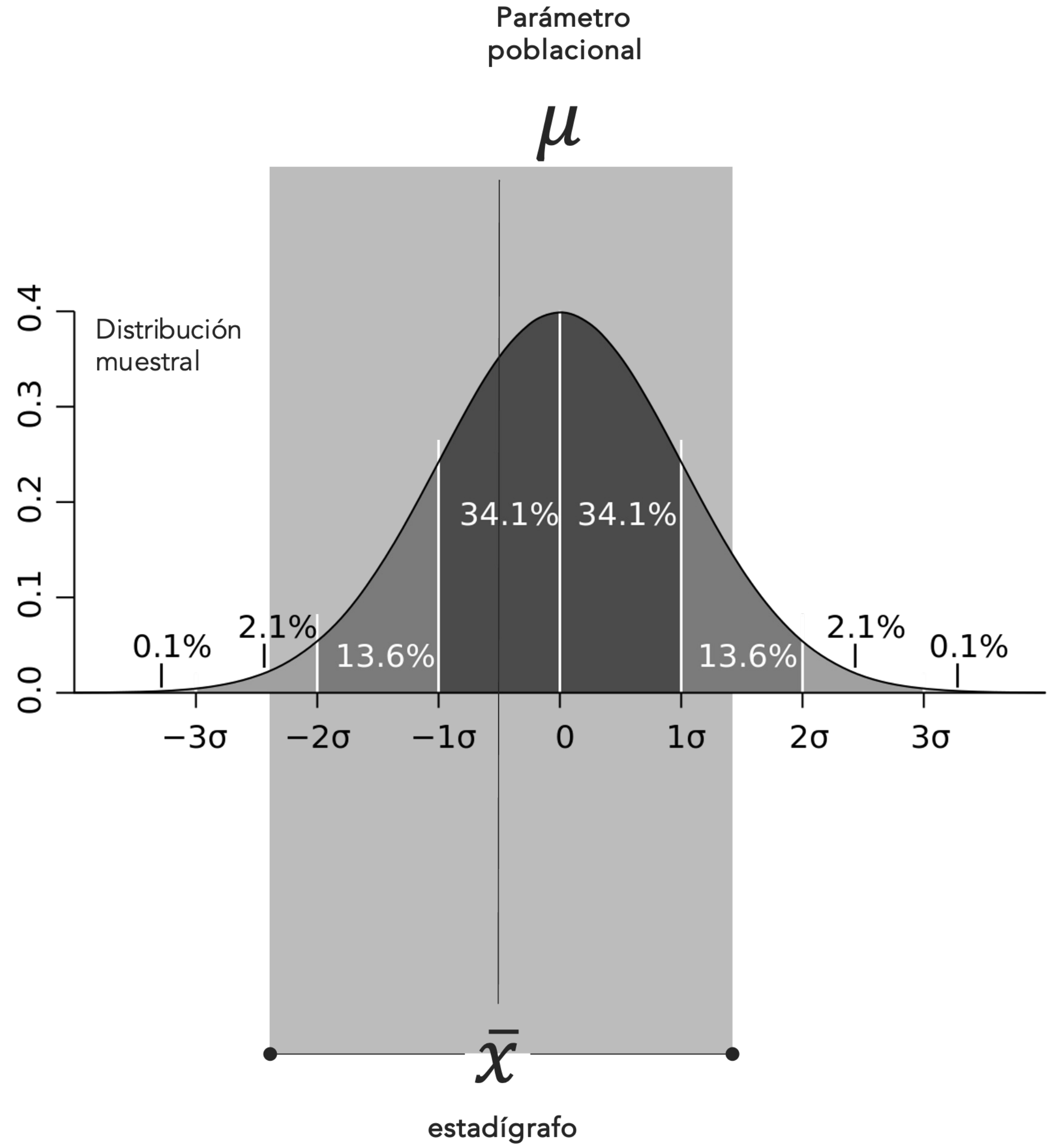
Intervalo de confianza

Un intervalo es un rango de valores que se extienden entre dos límites, uno superior y otro inferior.

Hablamos de **intervalo de confianza**, al intervalo que se construye alrededor de los estadígrafos, para garantizar que **el parámetro poblacional fuera contenido en una proporción de las veces**, es decir con una cierta probabilidad, si generásemos una distribución muestral.

El intervalo de confianza más tradicional es el de 95%. También es posible construir intervalos de 99%. Intervalos de 99% son a veces empleados para muestras grandes (e.g., 1000 observaciones). Mientras que los intervalos de 95% se emplean con muestras de menos de 1000 casos.

En esta clase hemos visto como crear intervalos de confianza para medias. También es posible crear estos intervalos para otros estadígrafos.



CI para	Estadísgrafo	Margen de error	Cuándo se usa tradicionalmente
Media poblacional (μ)	$\bar{x}$	$\pm z \frac{\sigma}{\sqrt{n}}$	x es normal n ≥ 30 σ es conocido
Media poblacional (μ)	$\bar{x}$	$\pm t_{n-1} \frac{s}{\sqrt{n}}$	n < 30 σ no es conocido
Proporcional poblacional (p)	$\bar{p}$	$\pm z \sqrt{\frac{\bar{p}(1-\bar{p})}{n}}$	Cada proporción p, o 1-p, es mayor o igual a 10 observaciones.
Diferencia de dos medias poblacionales ($\mu_1 - \mu_2$)	$\bar{x}_1 - \bar{x}_2$	$\pm z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	Ambas distribuciones son normales; de tamaño 30 o mayor; y/o con varianzas conocidas.
Diferencia de dos medias poblacionales ($\mu_1 - \mu_2$)	$\bar{x}_1 - \bar{x}_2$	$\pm t_{n_1+n_2-2} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$	Muestras de menos de 30 casos; y escenarios de varianzas no conocidas
Diferencia de dos proporciones poblacionales ($p_1 - p_2$)	$\bar{p}_1 - \bar{p}_2$	$\pm z \sqrt{\frac{\bar{p}_1(1-\bar{p}_1)}{n_1} + \frac{\bar{p}_2(1-\bar{p}_2)}{n_2}}$	p y 1-p, son proporciones generadas por muestras mayores a 10 observaciones, en ambos grupos.

Ref: https://www.dummies.com/article/academics-the-arts/math/statistics/surveying-statistical-con¦dence-intervals-188336/

Metodología Cuantitativa

Interpretación de los CI95%

Cómo interpretar intervalos de confianza

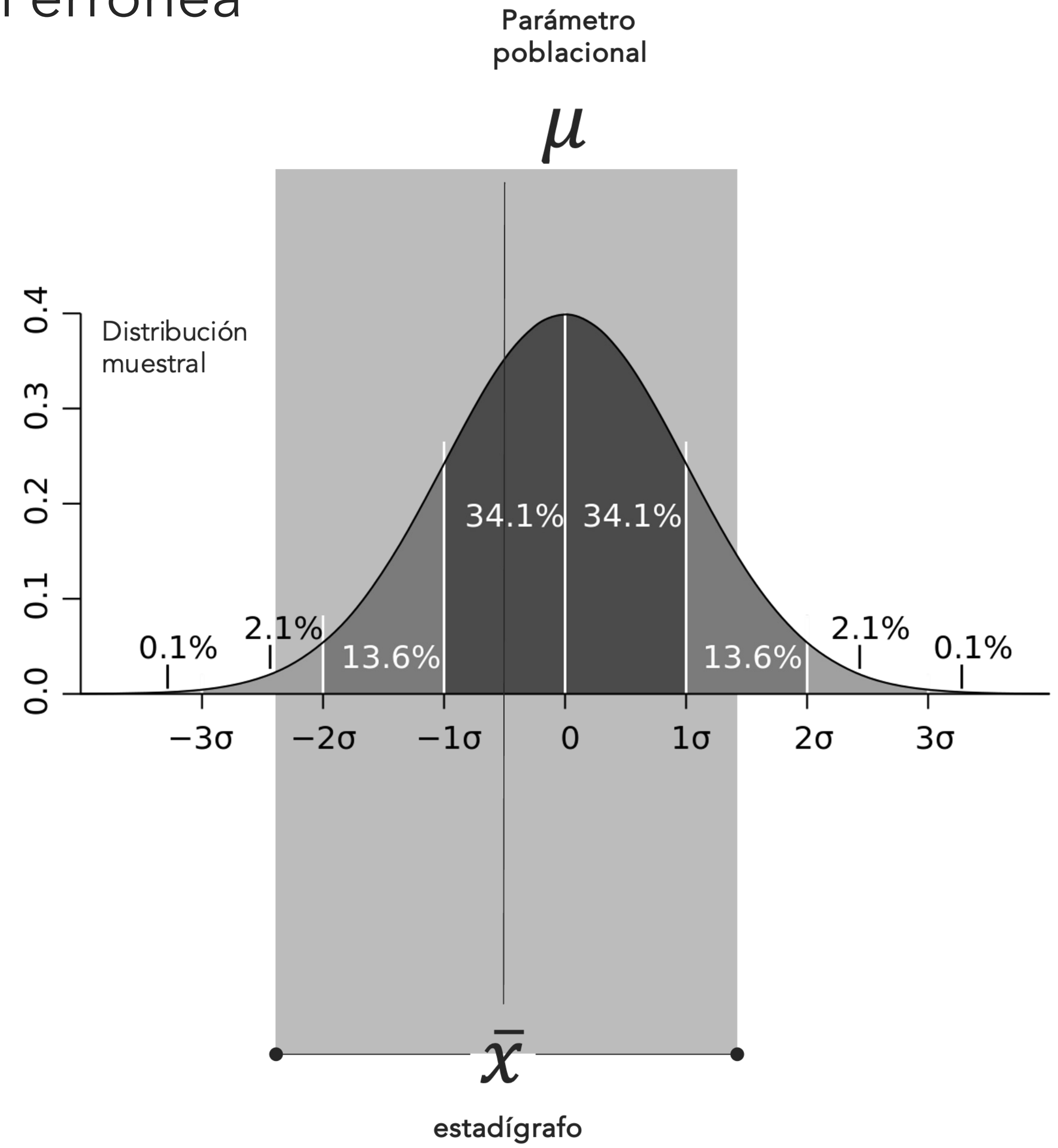
Intervalo de confianza y su interpretación errónea

Si un intervalo de confianza de 95% ha sido creado para estimar el valor de un parámetro de la población, la probabilidad que el parámetro de interés ***se encuentre entre los límites del intervalo de confianza, es igual a 95 de 100****. […] Esto es una mala comprensión de lo que implica un intervalo de confianza.*

– Huck, 2009, p147

Como vimos en las simulaciones, y en la selección de diferentes muestras aleatorias sobre datos reales (ver anexos), **estos intervalos "pueden" contener o no**, al parámetro de la población.

El "95% de las veces" es una expectativa respecto a los intervalos de confianza sobre las medias en la distribución muestral.

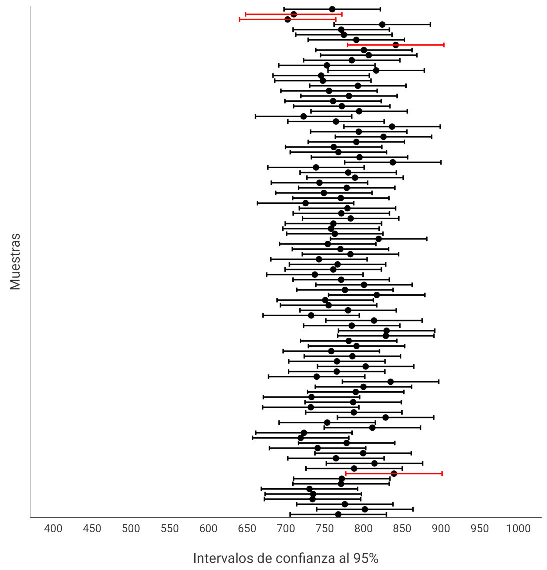


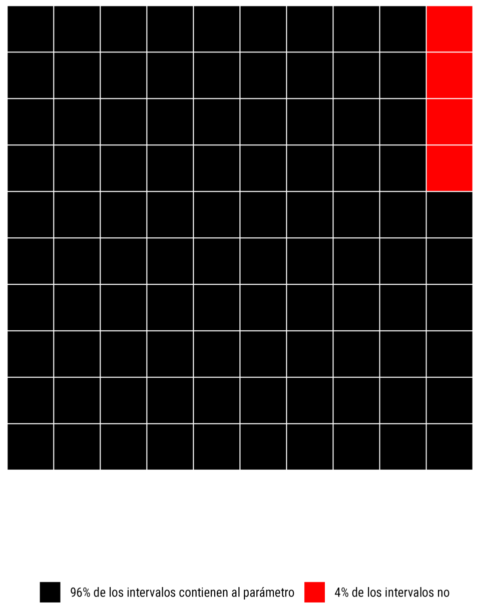
Intervalo de confianza y su interpretación

Dado el teorema del limite central, si muestras aleatorias fueran recogida infinitas veces se posee la expectativa de que **la distribución muestral resultante,** puede contener al parámetro de la población 95% de las veces, una vez construido los intervalos de confianza sobre las medias de cada muestra.

La probabilidad de contener 95% de las veces al parámetro, es una afirmación respecto a los intervalos de la distribución muestral, no respecto a una media de una sola muestra.

Uno de los aspectos críticos a distinguir en la interpretación de los CI95%, es el objeto de la afirmación anterior. Sería un error categorial (Ryle, 1949), asumir que el 95% de probabilidad es respecto a la media de una sola muestra. La interpretación correcta de un intervalo, es que tenemos un solo intervalo de 95% confianza, de una distribución muestral.





Cuando se produce una distribución muestral de n≥30 tamaño, podemos construir intervalos de 95% de confianza alrededor de todas las medias de cada muestra. De todos intervalos, 95% de estos intervalos van a capturar al parámetro poblacional. **Cuando estimamos resultados con una sola muestra aleatoria, tenemos solo uno de estos intervalos posibles.**

Intervalo de confianza y su forma de presentación

EXCERPTS 6.4–6.5 • *Confidence Intervals Around a Mean and a Percentage*

ABC scores were obtained from 46 children, the mean score was 51 and 95% CI was 46–56 (SD 18).

Source: Fernell, E., Hedvall, A., Norrelgen, F., Eriksson, M., Höglund-Carlsson, L., Barnevik-Olsson, M., et al. (2010). Developmental profiles in preschool children with autism spectrum disorders referred for intervention. *Research in Developmental Disabilities*, *31*(3), 790–799.

Of the 18–30 year olds, 49.8% (95% confidence interval: 48.5%–51.2%) were men and 50.2% (48.8%–51.5%) were women.

Source: Cavazos-Rehg, P. A., Spitznagel, E. L., Krauss, M. J., Schootman, M., Bucholz, K. K., Cottler, L. B., et al. (2010). Understanding adolescent parenthood from a multisystemic perspective. *Journal of Adolescent Health*, *46*(6), 525–531.

Los puntajes ABC fueron obtenidos de una muestra de 46 niños, con un promedio de 51 y un intervalo de confianza al 95% 46-56 (DS = 18).

De los jóvenes de 18 a 30 años, 49.8% (intervalo de confianza al 95%: 48.5%-51.2%) son hombres, y 50.2% (48.8%-51.5%) son mujeres.

Intervalo de confianza y su forma de presentación

In Table 3, we present the school level estimates. The predictor with the highest size is again authoritarianism, with a standardised effect in the constrained model of 0.84 ($SE = 0.03$, CI 95% [0.78, 0.89], $p < .001$). For Guatemala, this effect is smaller than the rest of the countries, reaching a standardised effect of 0.59 ($SE = 0.07$, CI 95% [0.46, 0.72], $p < .001$). Thus, schools with higher levels of authoritarianism present higher levels of tolerance of corruption in comparison to other schools. In addition, schools with higher civic knowledge present lower levels of tolerance of corruption, reaching a standardised effect of -0.19 ($SE = 0.04$, CI 95% [-0.26, -0.11], $p < .001$) in the constrained model. Inspecting the country specific results, we found that Guatemala presents a larger coefficient than the rest of the compared countries, where its standardised effect is of -0.41 ($SE = 0.09$, CI 95% [-0.59, -0.23], $p < .001$), whereas Paraguay and the Dominican Republic present non-significant effects on this estimate. However, all country estimates overlap with the confidence interval of the

Carrasco, D., Banerjee, R., Treviño, E., & Villalobos, C. (2020). Civic knowledge and open classroom discussion: explaining tolerance of corruption among 8th-grade students in Latin America. *Educational Psychology*, *40*(2), 186–206. https://doi.org/10.1080/01443410.2019.1699907

También es posible construir intervalos de confianza no solo sobre medias, o sobre porcentajes. Los podemos crear alrededor de coeficientes de modelos. En este caso, se reportan los intervalos de confianza alrededor de coeficientes estandarizados de un modelo mixto, o de regresión multinivel. En este caso, se reporta el estimado primero, al interior del paréntesis, se reporta al error estándar, y al intervalo de confianza del estimado.

Metodología Cuantitativa

Pausa

Recreo de 15 min

Metodología Cuantitativa

Construcción óptima de CI95%

Qué factores influyen sobre el rango intervalar

Intervalo de confianza y tamaño muestral óptimo

Por convención los intervalos de confianza se construyen con una confianza de 95%. En casos que se cuente con muestras muy grandes (e.g., 1000 observaciones), se puede emplear intervalos de 99% de confianza.

Los intervalos de confianza, serán más amplios o más estrechos según el tamaño muestral, a un grado de confianza fijo.

En términos generales, los intervalos pueden ser más grandes o más pequeños según el error estándar del estimado. Dicho de otro modo, según la desviación estándar del estadígrafo en la distribución muestral, los intervalos pueden ser más grandes o más pequeños.

En el caso de los porcentajes, una muestra aleatoria optima es una muestra que de unos 385 casos. Este tamaño muestral asegura que una proporción estimada, tenga un margen de error de 5%.

Proporción	Margen de Error (MOE)	Z crítico	var(p)	Resultado	n óptimo
0.00	0.025	1.96	0.00	0.00	0
0.10	0.025	1.96	0.09	49.79	50
0.20	0.025	1.96	0.16	157.35	158
0.30	0.025	1.96	0.21	271.06	272
0.40	0.025	1.96	0.24	354.04	355
0.50	0.025	1.96	0.25	384.16	385
0.60	0.025	1.96	0.24	354.04	355
0.70	0.025	1.96	0.21	271.06	272
0.80	0.025	1.96	0.16	157.35	158
0.90	0.025	1.96	0.09	49.79	50
1.00	0.025	1.96	0.00	0.00	0

$$n\left[\left(\frac{z*\sigma}{MOE}\right)^2\right] \quad n\left[\left(\frac{1.96*(\hat{p}*(1-\hat{p})}{.025}\right)^2\right]$$

Para calcular muestras aleatorias óptimas, se emplear al z crítico, a la varianza, y al margen de error (vinculado al intervalo deseado).

Intervalo de confianza y tamaño muestral óptimo

Por convención, la mayoría de los estudios observacionales, se optimizan a muestras aleatorias simples de 400 casos (Ross, 2005). Por ejemplo, los estudios de gran escala (e.g., PISA, ICCS, ERCE) son comúnmente estimados al grado de precisión que lograría una muestra aleatoria simple de 400 casos.

Esta cantidad, asegura que la estimación de una proporción de 50%, si se le construyera un intervalo de confianza, el limite inferior sería de .475, y su intervalo superior fuera de .525; de esta forma el rango intervalar de este intervalo de confianza es de 5%.

Para optimizar intervalos de confianza de otros estimados (e.g., medias, diferencias de medias, interacciones, varianzas), se puede proceder de forma similar, pero se tiene que considerar el error estándar. Es decir, la desviación estándar del estadígrafo en la distribución muestral.

$$n\left[\left(\frac{1.96 * (\hat{p} * (1-\hat{p})}{.025}\right)^2\right]$$

$$n = 385 = \left[\left(\frac{1.96 * (0.5 * (1-0.5)}{.025}\right)^2\right]$$

Una muestra aleatoria de 385 casos puede generar un intervalo de confianza de rango intervalar de 5%, sobre una proporción de 50%. Esta cifra tiende a ser redondeada a 400 casos, y se emplea como un benchmark de calidad de precisión para crear planes muestrales en estudios observacionales.

Intervalo de confianza de otros estadígrafos

En el caso de medias y medianas, lo crítico es la desviación estándar del estadígrafo en la distribución muestral. En el caso de las medias, el error estándar es:

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

Mientras que en el caso de las medianas, el error estándar es

$$\frac{\frac{\pi}{2}\sigma}{\sqrt{n}}$$

Debido a que la mediana posee un error estándar de mayor tamaño a la media (es 1.57 veces más grande), entonces se requieren tamaños muestrales de mayor tamaño para que el rango intervalar logrado tuviera la misma distancia que en el caso de la media. En términos relativo, se dice que la media es más eficiente que la mediana, debido a que su error estándar es de menor tamaño.

La variabilidad de la distribución muestral de la media es menor a la variabilidad de la mediana.

En términos relativos se plantea que la variabilidad de la mediana 1.57 veces mayor a la variabilidad de las medias ($1.57 = \frac{\pi}{2}$). También se puede decir que la variabilidad de la media es 63.7% mas eficiente que la mediana.

$$ER = \frac{\frac{\sigma}{\sqrt{n}}}{\frac{\frac{\pi}{2}\sigma}{\sqrt{n}}} = \frac{\sigma}{\frac{\pi}{2}\sigma} = \frac{1}{1.57} = .637$$

Dónde ER = eficiencia relativa.

Nota: La eficiencia, es una propiedad de los estimadores.

Intervalo de confianza de otros estadígrafos

A pesar de la convención de optimizar a muestras aleatorias simples de 400 casos, esto no es suficiente para todos los estimados de interés.

Debido a que las medias, diferencias de medias, y los coeficientes de regresión dependen de el error estándar del estimado, optimizar muestras para lograr una determinada precisión, puede requerir producir distribuciones muestrales durante el diseño del plan muestral.

Por ejemplo, si quisiéramos evaluar una interacción casi nula (b = 0.05), en una regresión donde hubieran términos de prequeños (b = .30), el tamaño la muestra requerida puede llegar a 2500 observaciones. Muy por encima del benchmark de los estudios de encuestas.

Existe todo un marco de trabajo para estos propósitos, los cuales son llamados **estudios de poder estadístco.**

```
 library(InteractionPoweR)
> test_power <- power_interaction_r2(
+   alpha = 0.05,             # alpha, for the power analysis
+   N = 2500,                 # sample size
+   r.x1x2.y = .05,           # interaction effect to test (correlation between x1*x2 and y)
+   r.x1.y =  -.3,            # correlation between x1 and y
+   r.x2.y =   .3,            # correlation between x2 and y
+   r.x1.x2 =  .3             # correlation between x1 and x2
+ )
>
> test_power
        pwr
0.8094897
```

Los estudios de poder estadístico buscan responder a la pregunta de que tamaño muestral se requiere para estimar un estadígrafo con una determinada precisión.

Consisten en realizar simulaciones, con las cuales se construyen distribuciones muestrales. Empleando estas distribuciones del estadígrafo de interés, se toman decisiones de diseño.

Muchas gracias!

Carrasco, D., PhD,
Centro de Medición MIDE UC
Pontificia Universidad Católica de Chile
https://dacarras.github.io/

Referencias

Carrasco, D., Banerjee, R., Treviño, E., & Villalobos, C. (2020). Civic knowledge and open classroom discussion: explaining tolerance of corruption among 8th-grade students in Latin America. Educational Psychology, 40(2), 186–206. https://doi.org/10.1080/01443410.2019.1699907

Devlin, D., Guo, X., Kunin, D., & Xiang, D. (2018). Seeing Theory. Brown University. https://seeing-theory.brown.edu/doc/seeing-theory.pdf

Huck, S. W. (2009). Statistical Misconceptions. Routledge.

Rabe-Hesketh, S., & Skrondal, A. (2012). Multilevel and Longitudinal Modeling Using Stata, Volumes I and II, Third Edition (3rd ed.). Stata Press.

Ross, K. N. (2005). Sample design for educational survey research. In Quantitative research methods in education planning. International Institute for Educational Planning (IIEP). UNESCO. http://www.unesco.org/iiep/PDF/TR_Mods/Qu_Mod3.pdf

Rumsey, D. (2019). Statistics Essentials For Dummies. John Willey & Sons.

Ryle, G. (1949). The Concept of Mind (60th Anniv). Routledge.

Sterba, S. K. (2009). Alternative Model-Based and Design-Based Frameworks for Inference From Samples to Populations: From Polarization to Integration. In Multivariate Behavioral Research (Vol. 44, Issue 6, pp. 711–740). https://doi.org/10.1080/00273170903333574

Stigler, S. M. (2016). The seven pillars of statistical wisdom. Harvard University Press.