


Analítica de los Negocios

Distribución muestral e Intervalos de Confianza

Carlos Cardona Andrade

¿Qué hemos visto hasta ahora?

- Una (muy!) breve introducción a 
- Los paquetes `ggplot` y `dplyr` para visualizar y transformar datos
- Cómo visualizar y describir variables numéricas
 - Tendencia central, dispersión, histogramas, gráficos de dispersión, etc
- Cómo visualizar y describir variables categóricas
 - Tablas de frecuencia, gráficos de barras, etc
- La Distribución Normal

Plan para hoy

1. Distribución Muestral
2. Intervalos de Confianza

Población y Muestra

- Los estadísticos descriptivos discutidos anteriormente describen una **muestra**, pero no a la **población**
- Las medidas que describen a una población se llaman **parámetros**. Utilizamos letras griegas para referirnos a ellos

Medida	Parámetro poblacional	Estadístico muestral
Media	μ	\bar{x}
Varianza	σ^2	S^2
Desviación Estándar	σ	S

¿Por qué nos importa tanto la población vs la muestra?

- La inferencia estadística es el acto de generalizar a partir de una muestra para sacar conclusiones sobre una población
- Nos interesan los parámetros de la población, pero no podemos observarlos directamente. En su lugar, calculamos estadísticas a partir de una muestra para hacer inferencias sobre ellos

$$\bar{x} \xrightarrow{\text{🙏 ojalá 🙏}} \mu$$

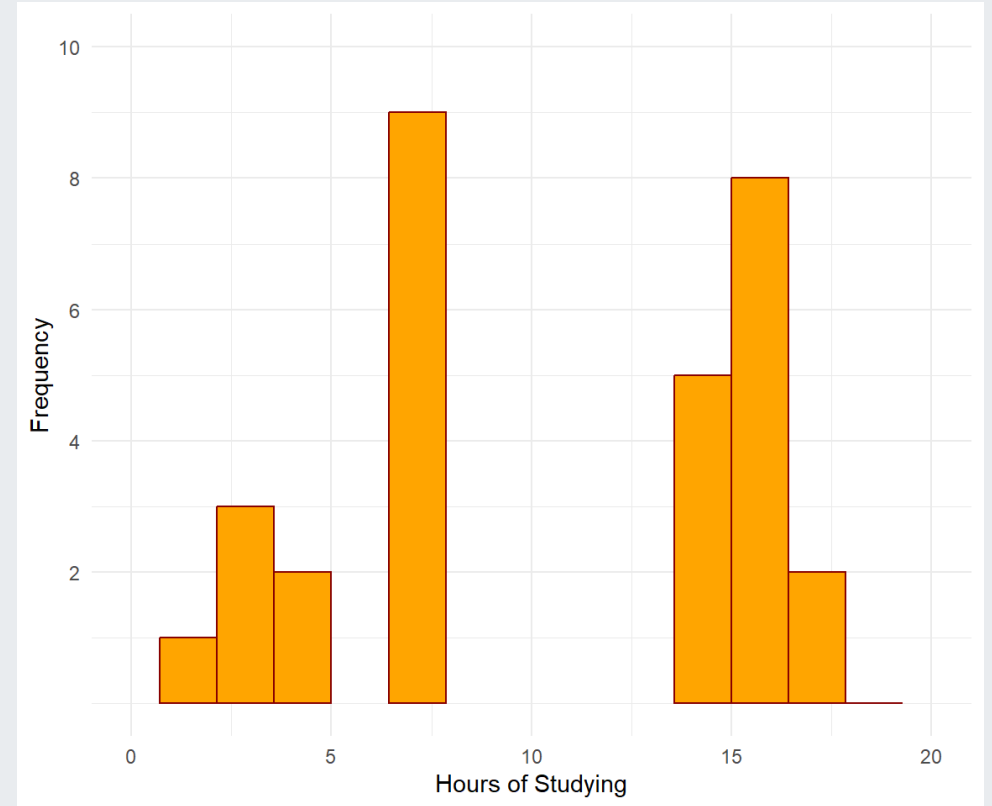
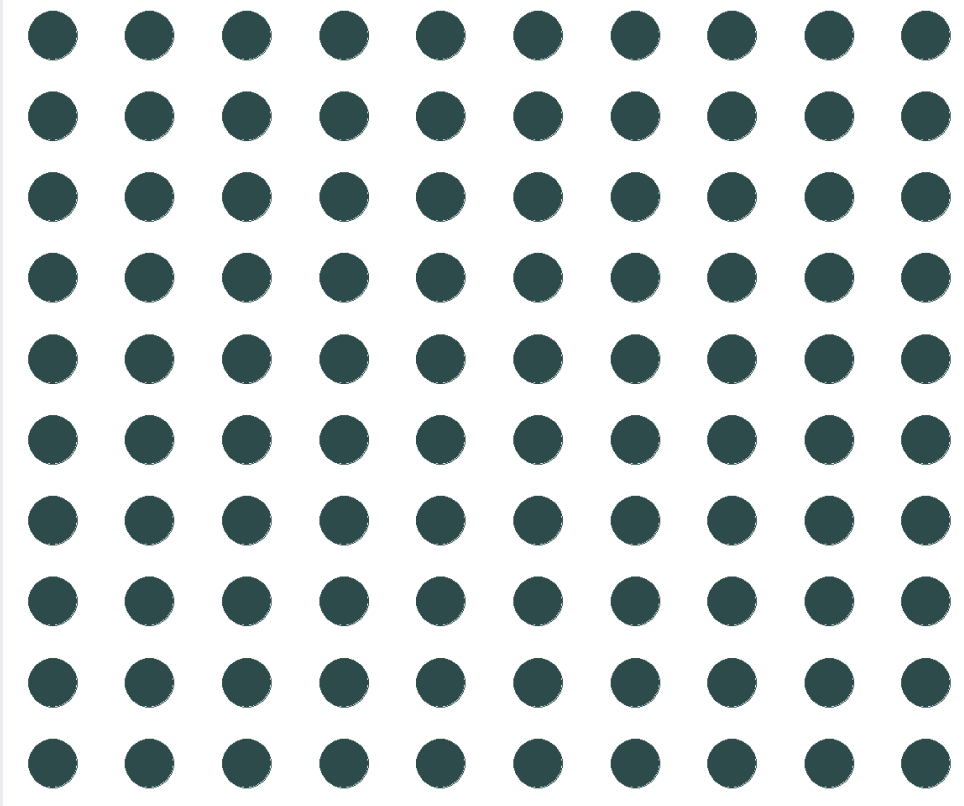
- Como parte de este proceso, es importante cuantificar la incertidumbre asociada a los estadísticos muestrales

¿Por qué nos importa tanto la población vs la muestra?

Para responder a esta pregunta, utilicemos el siguiente ejemplo:

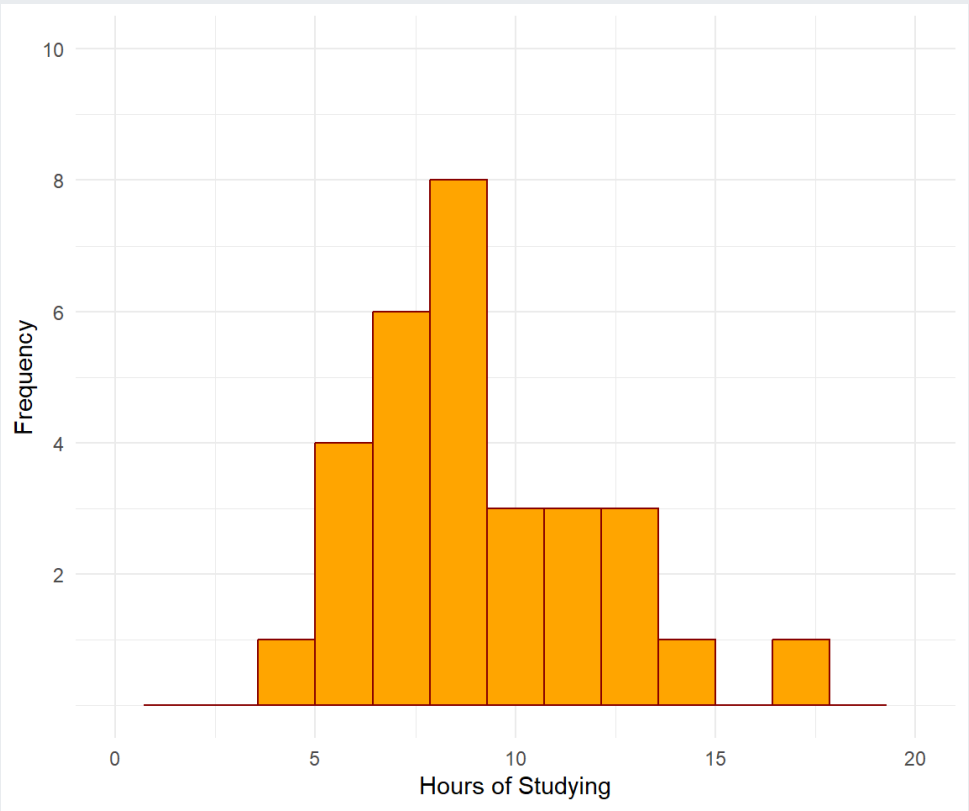
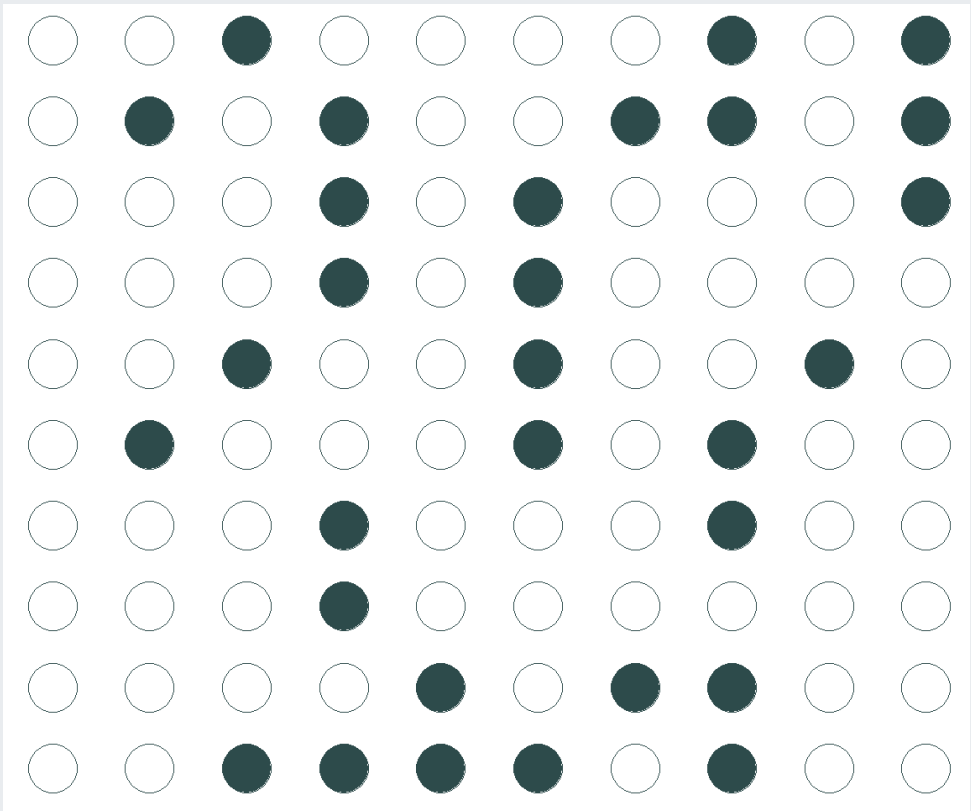
- 100 estudiantes están tomando Analítica de los Negocios y la universidad quiere saber el número de horas de estudio de todo ellos. Asumamos que tenemos los datos para todos los estudiantes
- Luego tomemos muestras de 30 estudiantes y miremos cómo el número de horas de estudio varían entre ellas

La distribución poblacional de horas de estudio



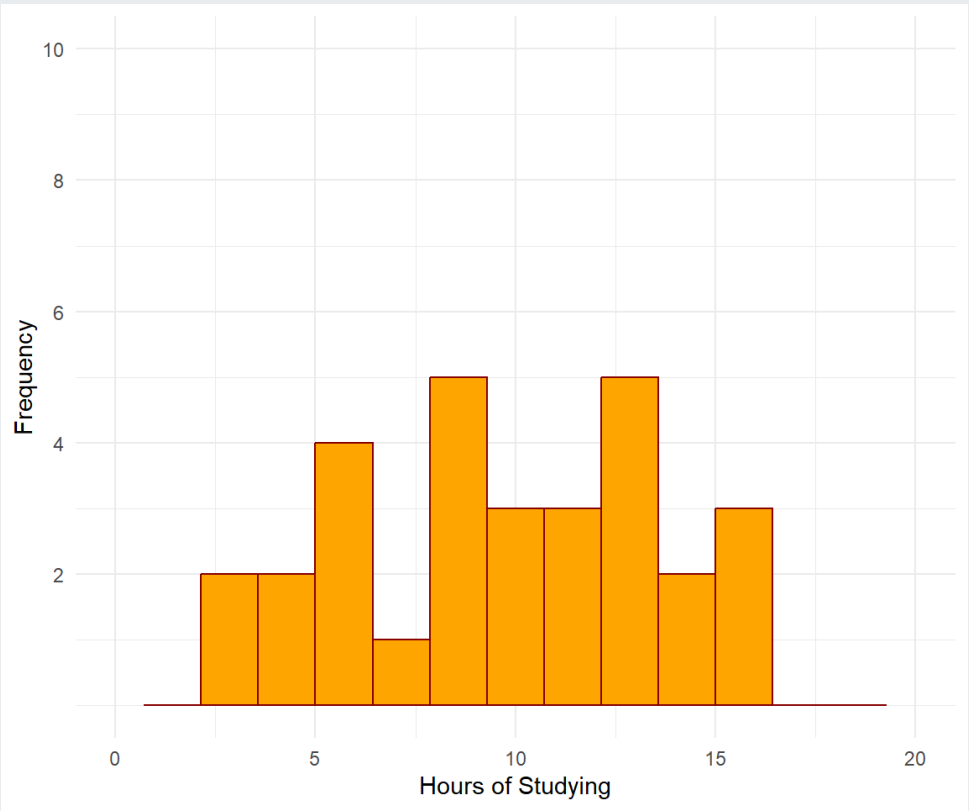
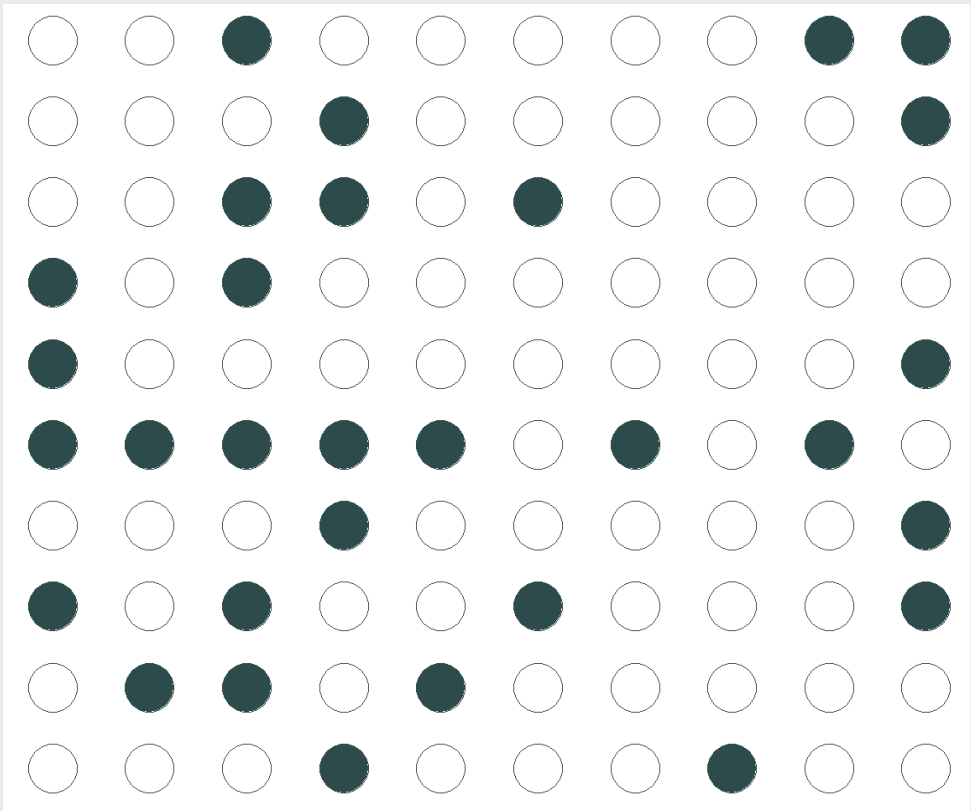
La media **poblacional** () es 9.9

Horas de estudio para la Muestra 1



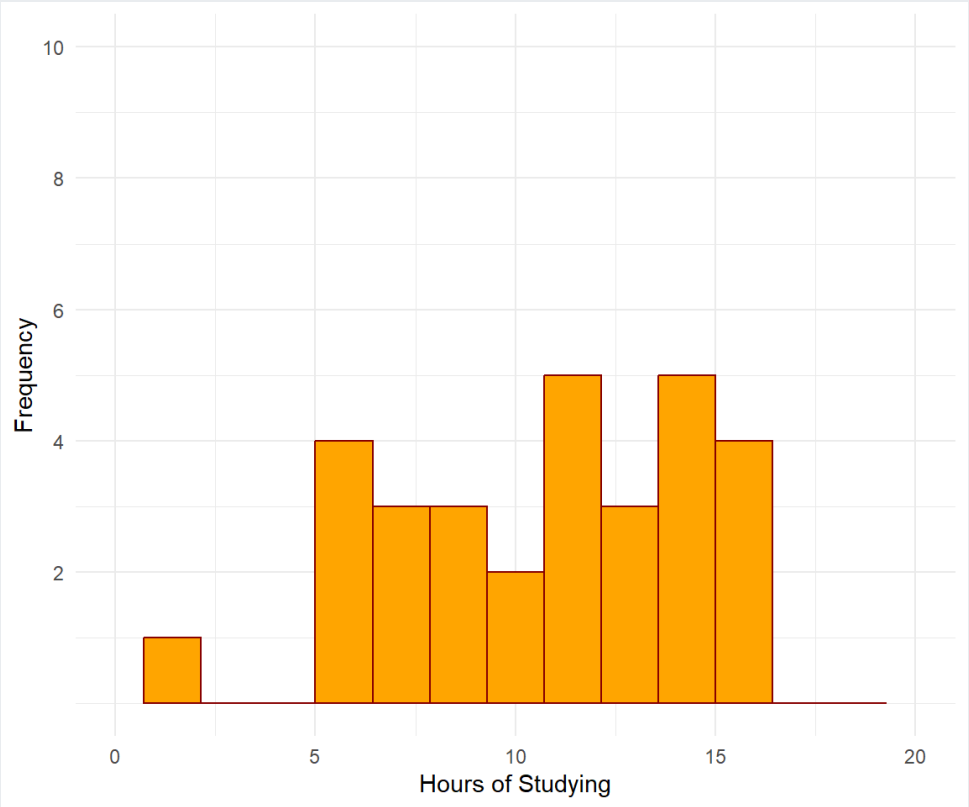
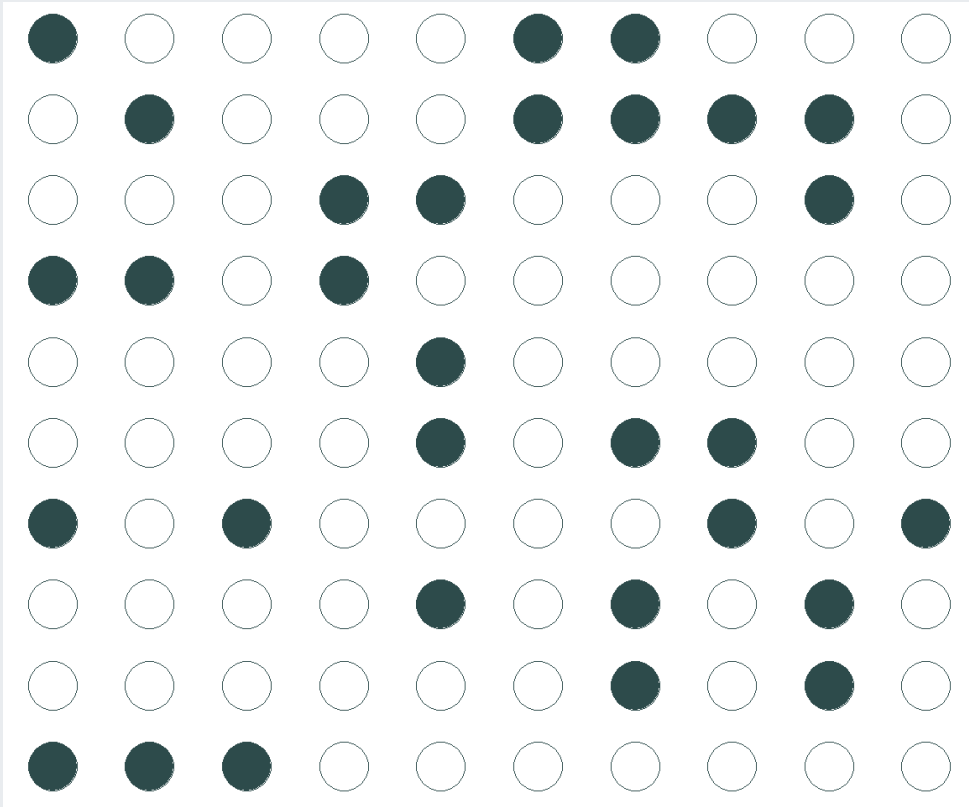
La media **muestral** es 9.1

Horas de estudio para la Muestra 2



La media **muestral** es 9.7

Horas de estudio para la Muestra 3



La media **muestral** es 10.8

¿Por qué nos importa tanto la población vs la muestra?

- Como vimos en los anteriores ejemplos, algunas muestras pueden estar lejos del verdadero valor del parámetro poblacional (el número de horas de estudio promedio para todos los estudiantes tomando Analítica).
- Diferencias entre muestras individuales y la población generan **incertidumbre** para quien analiza el problema

¿Por qué nos importa tanto la población vs la muestra?

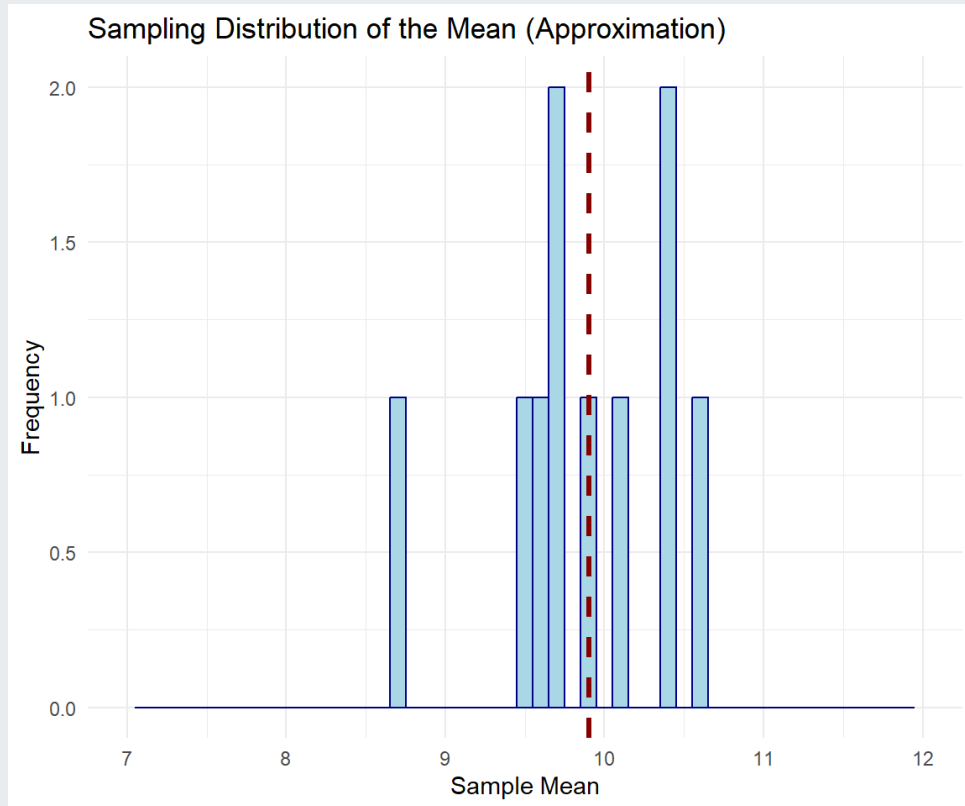
Respuesta: !Porque la incertidumbre importa!

Cuando tomamos una muestra, no sabemos si es una *buena* muestra (está cerca a) o una *mala* muestra (difiere mucho de)

Distribución Muestral de la Media

Para analizar la variabilidad de las medias muestrales con en nuestro ejemplo, tomemos 10 muestras, calculemos la media de horas de estudio en cada una y representemos los resultados con un histograma...

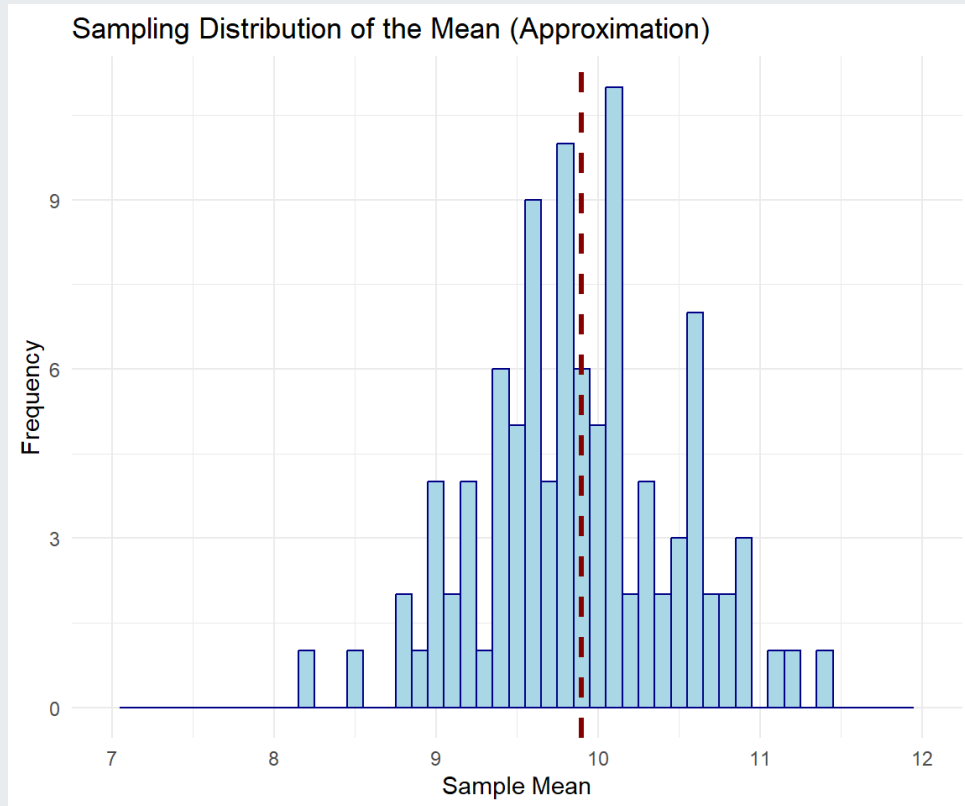
Distribución Muestral de la Media para 10 muestras



Así se ve la distribución de las medias muestrales para 10 muestras, con la media poblacional representada por la línea punteada (9.9 horas de estudio en promedio).

Sin embargo, no se pueden sacar muchas conclusiones con base en el gráfico.

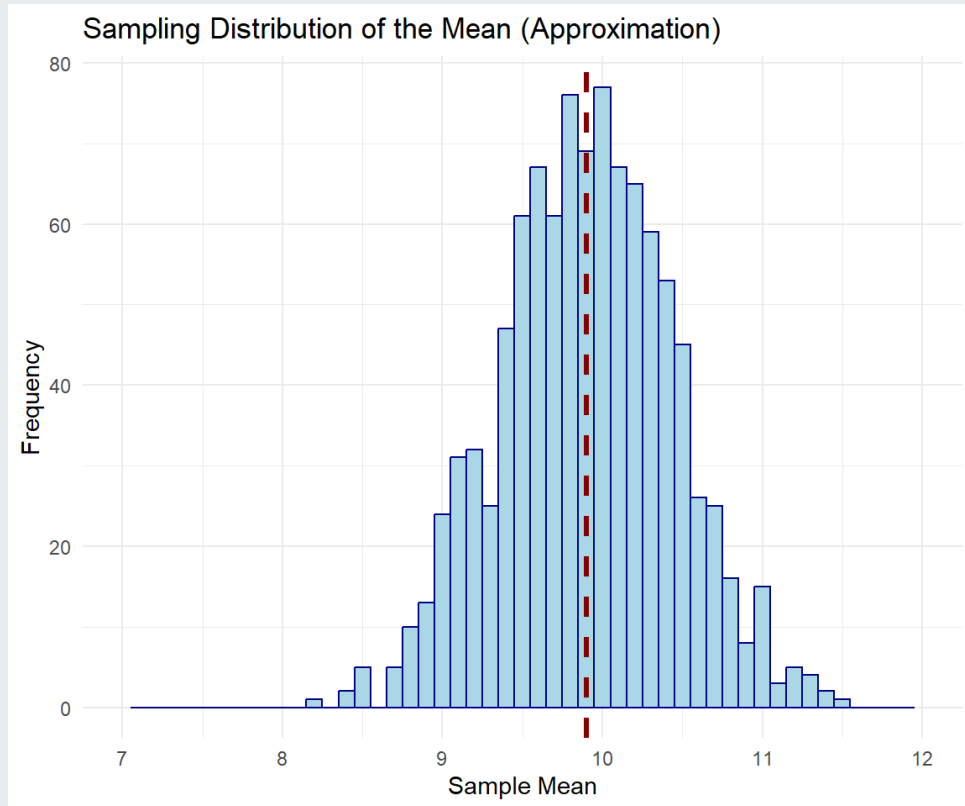
Distribución Muestral de la Media para 100 muestras



¿Qué pasa si tomamos 100 muestras?

🤔 Hmm cierta forma ya conocida empieza a emerger...

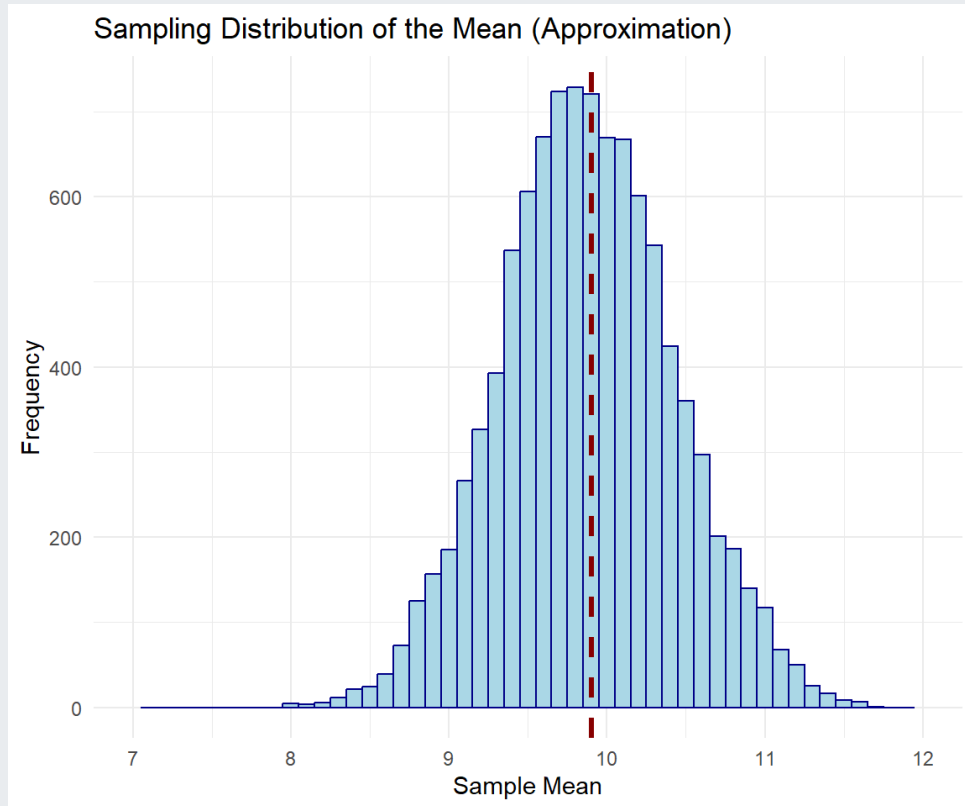
Distribución Muestral de la Media para 1000 muestras



¿Qué pasa si tomamos 1000 muestras?

Ya es claro que la distribución de las medias muestrales tiende a ser normal...

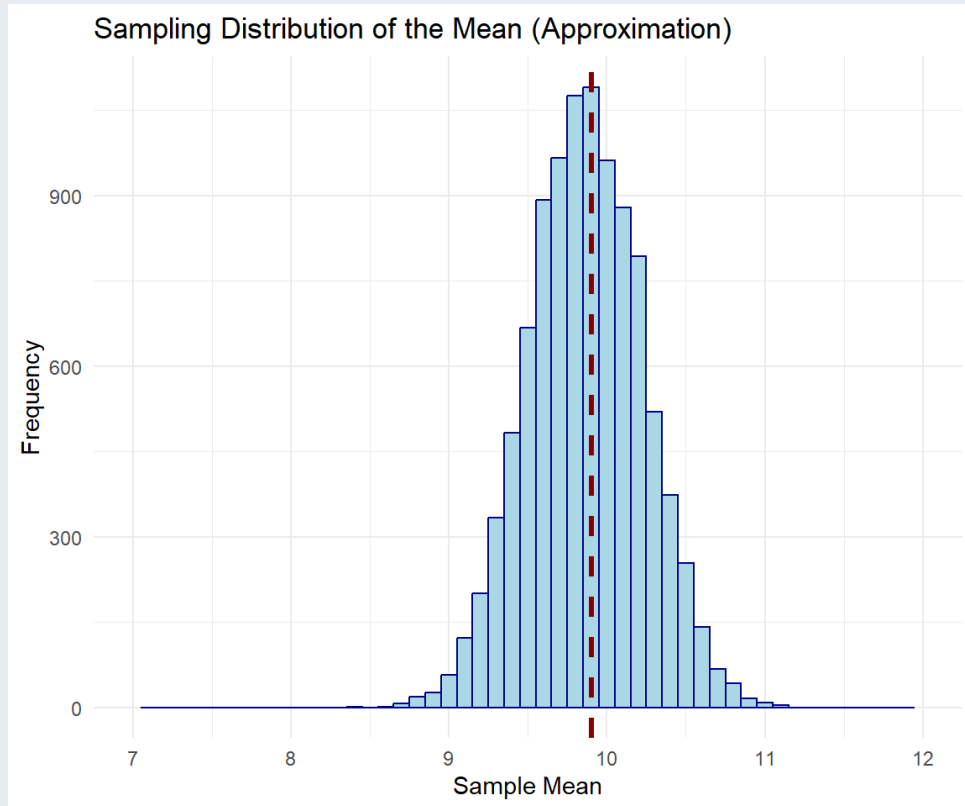
Distribución Muestral de la Media para 10000 muestras



¿Qué pasa si tomamos 10000 muestras?

...y que la distribución está centrada en la media poblacional.

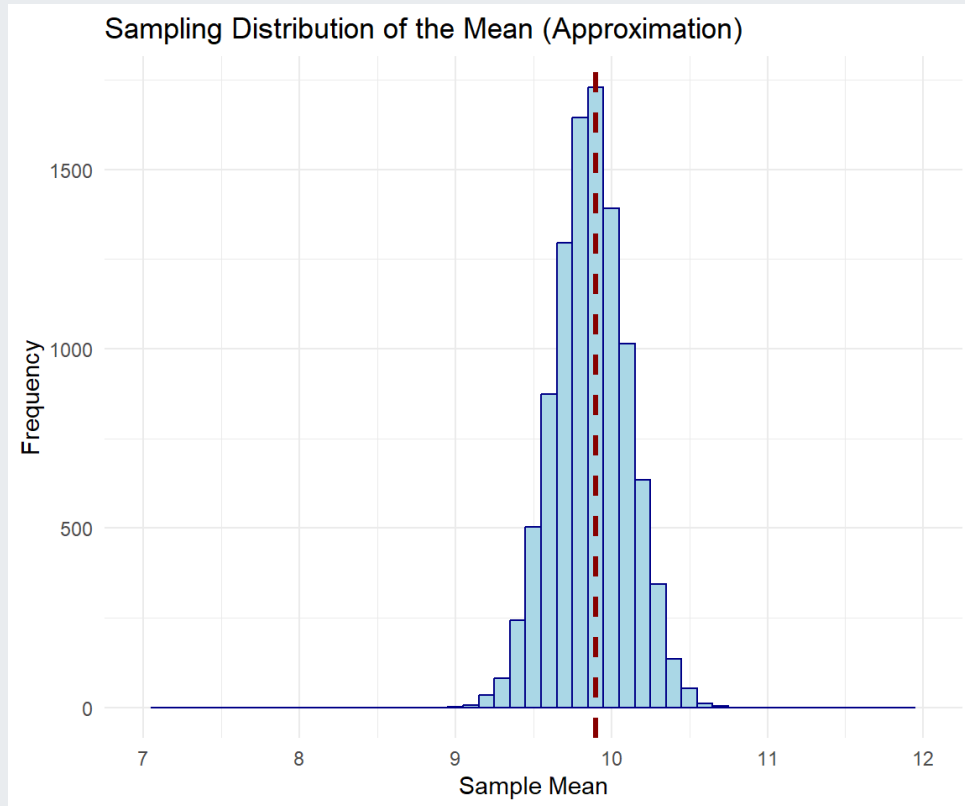
Distribución Muestral de la Media con un n mayor



¿Qué pasa si tomamos 10000 muestras pero de 50 estudiantes?

La variabilidad de las medias muestrales es menor.

Distribución Muestral de la Media con un n mayor



¿Qué pasa si tomamos 10000 muestras pero de 70 estudiantes?

A medida que el tamaño de la muestra aumenta, las medias muestrales tienden a concentrarse más alrededor del parámetro poblacional

¿Qué podemos concluir de los histogramas anteriores?

- La distribución muestral de la media muestral puede no ser normal cuando el tamaño de la muestra es pequeño, pero se vuelve más normal a medida que el tamaño de la muestra aumenta.
- La media muestral puede no ser igual a la media poblacional, pero su distribución se centra en la media poblacional.
- Con una muestra más grande, la variabilidad de la media muestral alrededor de la población disminuye.

Teorema del Límite Central

Para una población con una media bien definida y una desviación estándar se cumplen estas tres propiedades para la distribución de la media muestral , siempre que se cumplan ciertas condiciones:

1. La media de la distribución muestral de la media tiende a la media poblacional .
2. La desviación estándar de la distribución de las medias muestrales es

Esto se llama el *error estándar* (SE) de la media.

3. Para un suficientemente grande, la forma de la distribución muestral de las medias es aproximadamente normal.

Teorema del Límite Central

Básicamente el teorema se resume de la siguiente manera:

E identifica las tres características básicas de una distribución:

1. Forma Normal
2. Tendencia central
3. Dispersión

Horas de Estudio y el TLC

Volviendo al ejemplo de nuestros estudiantes, la media es 9.9 y la desviación estándar es 3.6. Entonces la distribución muestral de es aproximadamente:

Dado que el TLC dice que la distribución de las medias muestrales es normal, podemos calcular probabilidades bajo la curva. Por ejemplo:

Horas de Estudio y el TLC

```
1 pnorm(10, mean = 9.9, sd = 0.65, lower.tail=FALSE)
```

```
[1] 0.4388655
```

En nuestra simulación con las 1000 muestras simuladas de , ¿cuántas medias muestrales están por encima de 10?

```
1 sum(sample_means > 10)
```

```
[1] 425
```

Este número se acerca a la aproximación con el TLC, el cual establece que son aproximadamente 43% () de las medias muestrales.

Condiciones para el TLC

- ✓ **Independencia:** la muestra debe ser tomada aleatoriamente
- Si las muestras son independientes, por definición el valor de una muestra no debería “influnciar” los valores de otras muestras

Condiciones para el TLC

✓ Tamaño de la muestra /Distribución

- Si los datos son numéricos, usualmente es una muestra suficiente para que el TLC se aplique
- Si sabemos que los datos se distribuyen como una normal, la distribución de las medias muestrales también será normal, sin importar

Intervalos de Confianza

Intervalos de Confianza



- Usar un estadístico muestral para estimar un parámetro es como pescar con una lanza en un lago fangoso
- Si tiramos la lanza donde creemos ver un pez, lo más probable es que fallemos
- Si reportamos una estimación puntual, lo más probable es que no le peguemos al parámetro poblacional

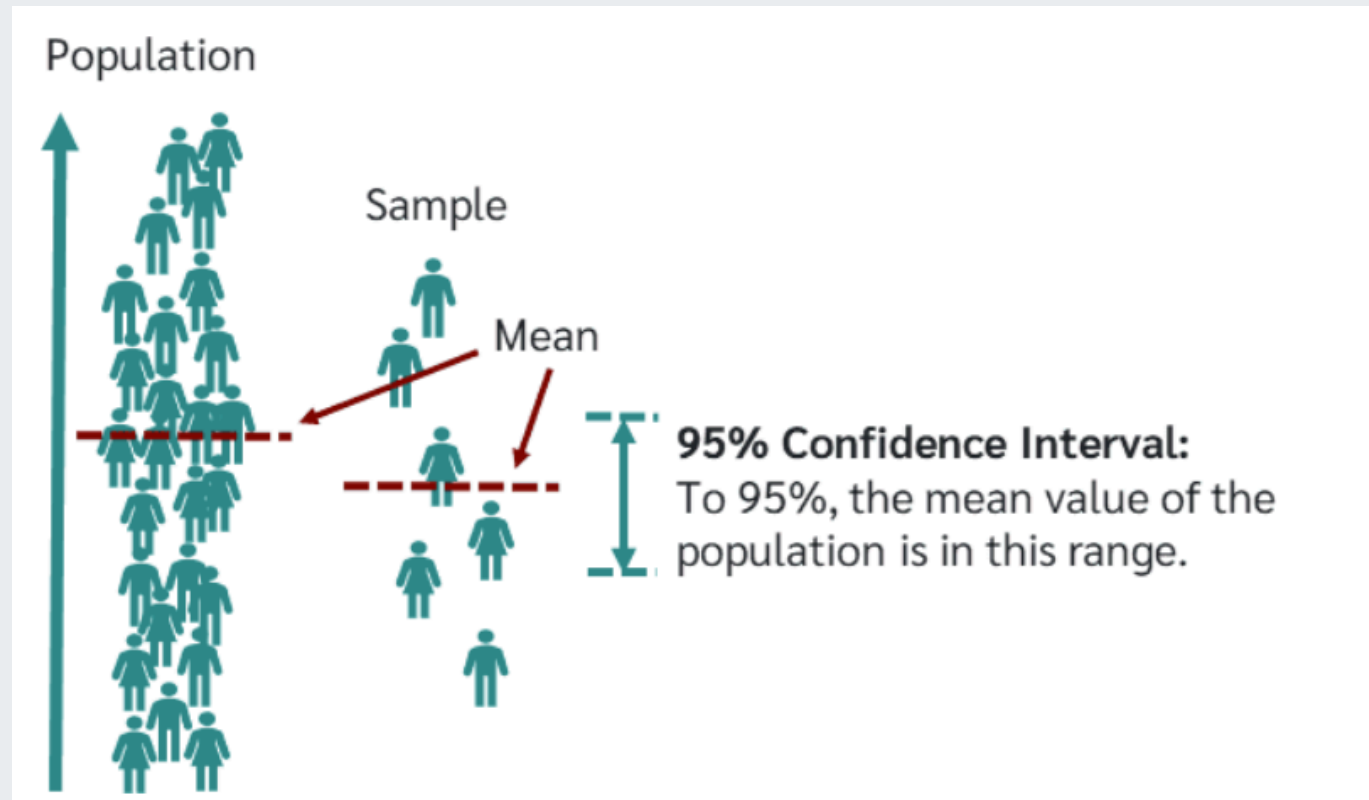
Intervalos de Confianza



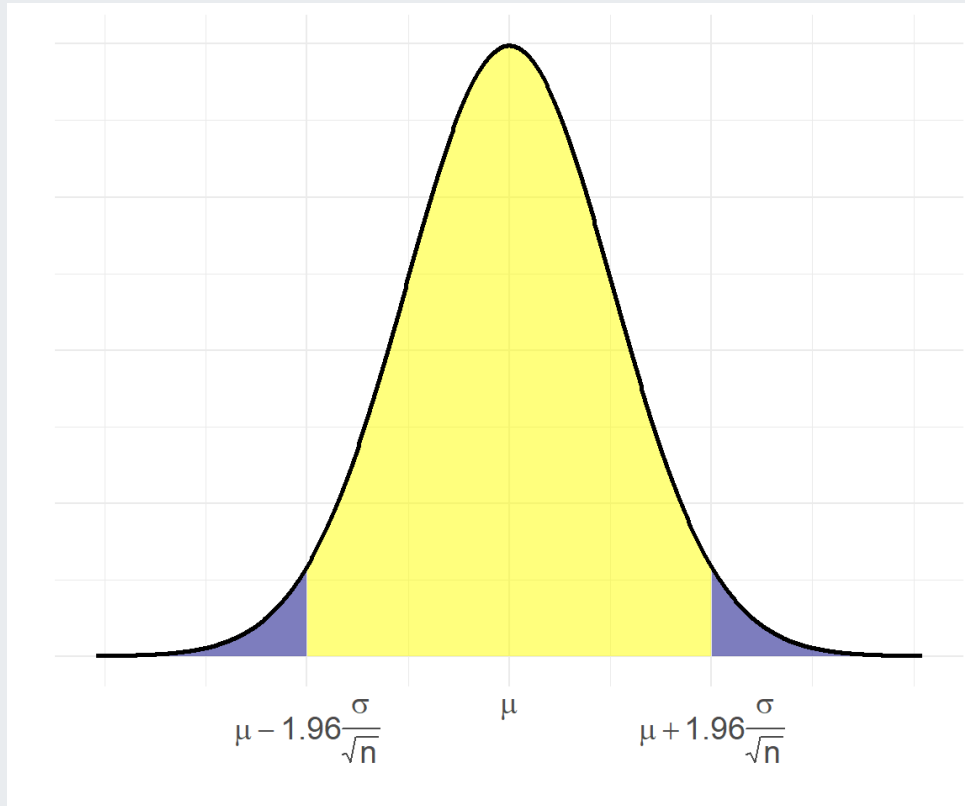
- Por otro lado, usar intervalos de confianza es como pescar con una red
- Si tiramos la red donde creemos ver un pez, tenemos una buena oportunidad de atraparlo
- Si reportamos un rango de valores plausibles, tenemos también una gran oportunidad de capturar al parámetro poblacional

¿Qué es un intervalo de confianza?

Con un intervalo de confianza tomamos una estimación puntual de la muestra y, con el conocimiento que tenemos sobre las distribuciones muestrales, tratamos de acercarnos al parámetro poblacional



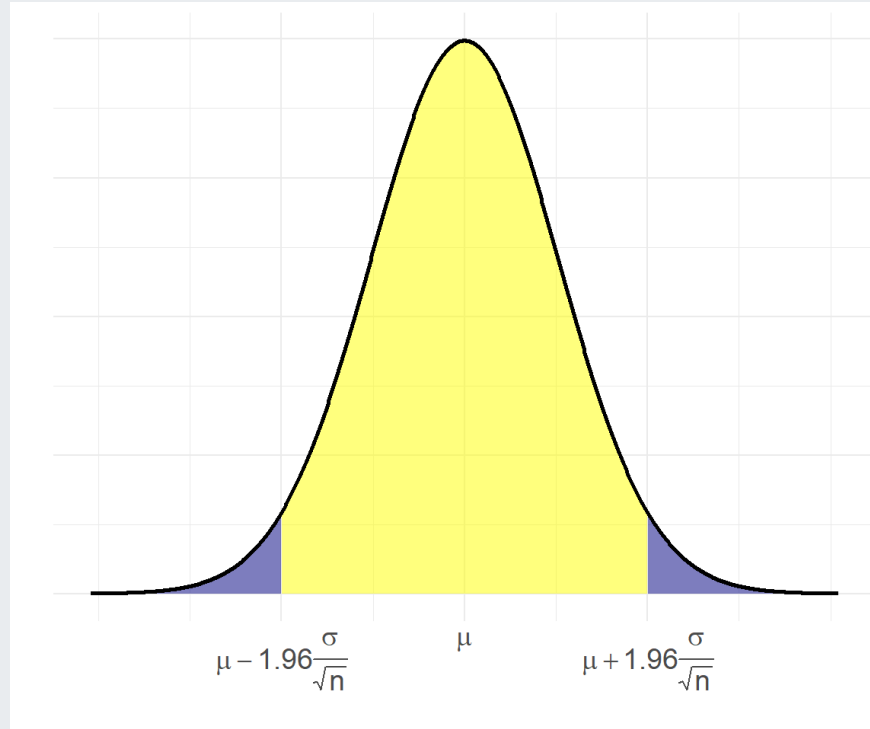
¿Qué es un intervalo de confianza?



- EL TLC dice que
- Para una curva normal, 95% de su área está dentro de 1.96 SD con respecto al centro.
- Eso significa que, para un 95% de las veces, va a estar a de

¿Qué es un intervalo de confianza?

Por lo tanto, un intervalo de confianza de al 95% será:



Procedimientos para encontrar un IC al 95%

1. Tomen una muestra aleatoria de tamaño n y calculen la media muestral \bar{x}
2. Si n es grande, el intervalo de confianza a un 95% de será:

El problema radica en que usualmente es desconocida para nosotros...

Los parámetros poblacionales son desconocidos...

Cuando la desviación estándar poblacional es desconocida, la reemplazamos con la mejor aproximación que tenemos - la desviación estándar muestral .

Así, un aproximado al intervalo de confianza a un 95% para es:

- Sin embargo, este reemplazo es peligroso:
 - es una estimación pobre de si es pequeño
 - es bastante sensible a datos atípicos

Ejemplo: Expectativa de Vida en 2007

Primero, tomemos la población de países en 2007:

```
1 library(gapminder)
2 library(tidyverse)
3 gapminder_2007 <- gapminder |>
4   filter(year==2007)
```

¿Cuál es la expectativa de vida promedio ese año? (Media poblacional)

```
1 pop_mean <- mean(gapminder_2007$lifeExp, na.rm = TRUE)
2 pop_mean
```

```
[1] 67.00742
```

Ejemplo: Expectativa de Vida en 2007

Seleccionemos una muestra de 30 países de manera aleatoria y calculemos los estadísticos muestrales:

```
1 # Establecer una semilla garantiza que la muestra va a ser la misma
2 set.seed(1)
3 # Tomamos una muestra de 30 países
4 sample_gapminder <- gapminder_2007 |>
5   sample_n(30)
6 # Calculamos la media muestral
7 sample_mean <- mean(sample_gapminder$lifeExp, na.rm = TRUE)
8 sample_mean
```

```
[1] 70.161
```

```
1 # Calculamos el error estándar
2 std_error <- sd(sample_gapminder$lifeExp, na.rm = TRUE) / sqrt(nrow(sample_
3 std_error
```

```
[1] 1.811705
```

Ejemplo: Expectativa de Vida en 2007

```
1 # Calculamos los intervalos de confianza
2 ic_inf <- sample_mean - 1.96 * std_error
3 ic_sup <- sample_mean + 1.96 * std_error
4
5 # Definámoslo como intervalo
6 ic <- c(ic_inf, ic_sup)
7 ic
```

```
[1] 66.61006 73.71194
```

¿Qué pasa si calculamos el IC con la desviación estándar poblacional?

```
1 # Calculamos el error estándar con la desviación estándar poblacional
2 std_error_2 <- sd(gapminder_2007$lifeExp, na.rm = TRUE) / sqrt(nrow(sample_
3
4 # Calculamos los intervalos de confianza
5 ic_inf_2 <- sample_mean - 1.96 * std_error_2
6 ic_sup_2 <- sample_mean + 1.96 * std_error_2
7
8 # Definámoslo como intervalo
9 ic_pop <- c(ic_inf_2, ic_sup_2)
10 ic_pop
```

[1] 65.84073 74.48127

Ejemplo: Expectativa de Vida en 2007

Otra manera de calcular el intervalo de confianza es con la función `t.test()`:

```
1 result <- t.test(sample_gapminder$lifeExp, conf.level = 0.95)
2 result
```

One Sample t-test

```
data: sample_gapminder$lifeExp
t = 38.727, df = 29, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 66.45565 73.86635
sample estimates:
mean of x
 70.161
```

```
1 # Extraemos los intervalos
2 ic_2 <- result$conf.int
3 ic_2
```

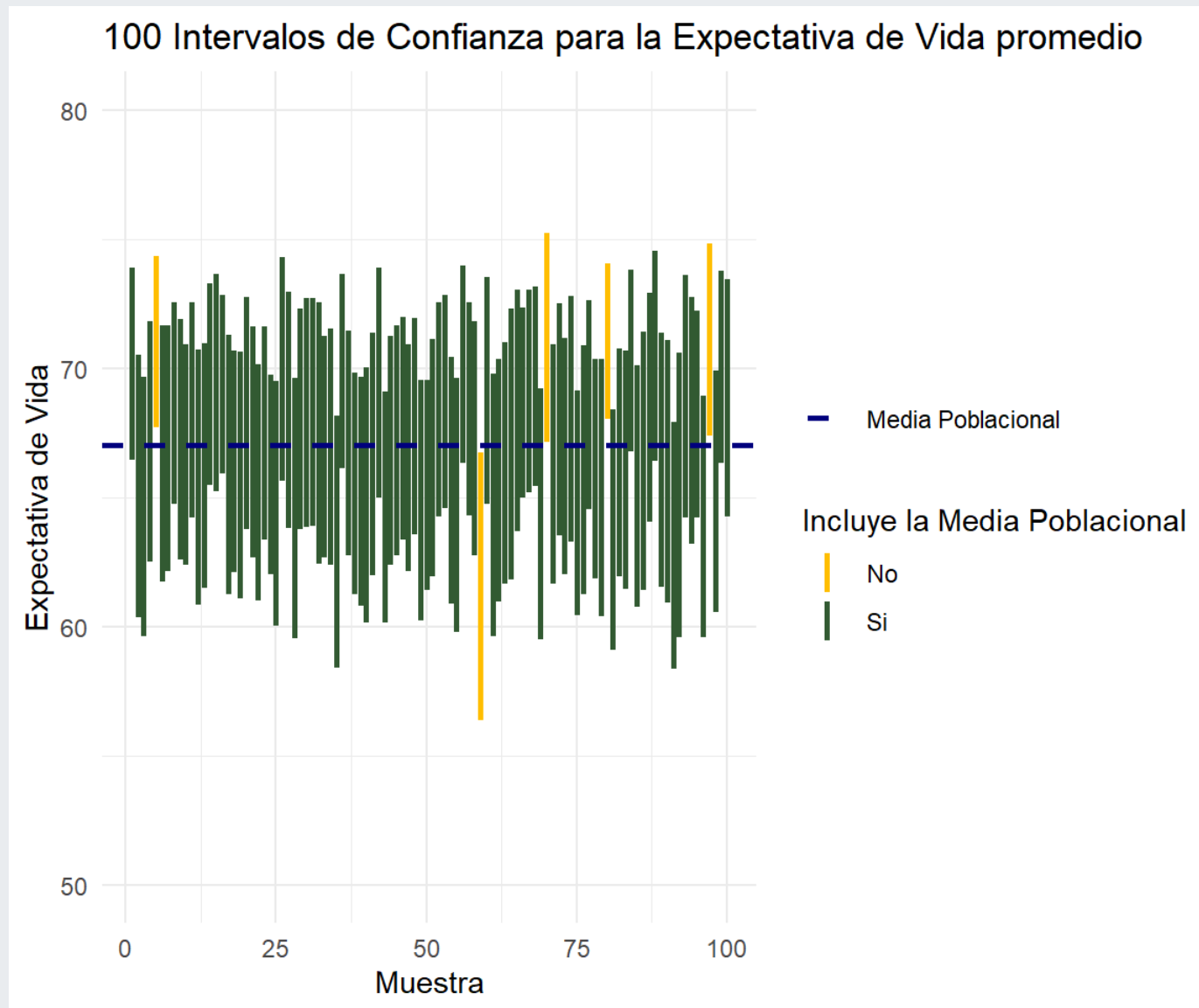
```
[1] 66.45565 73.86635
attr(,"conf.level")
[1] 0.95
```

¿Qué significa “nivel de confianza al 95%”?

¿Qué es lo que tiene un 95% de probabilidad de ocurrir?

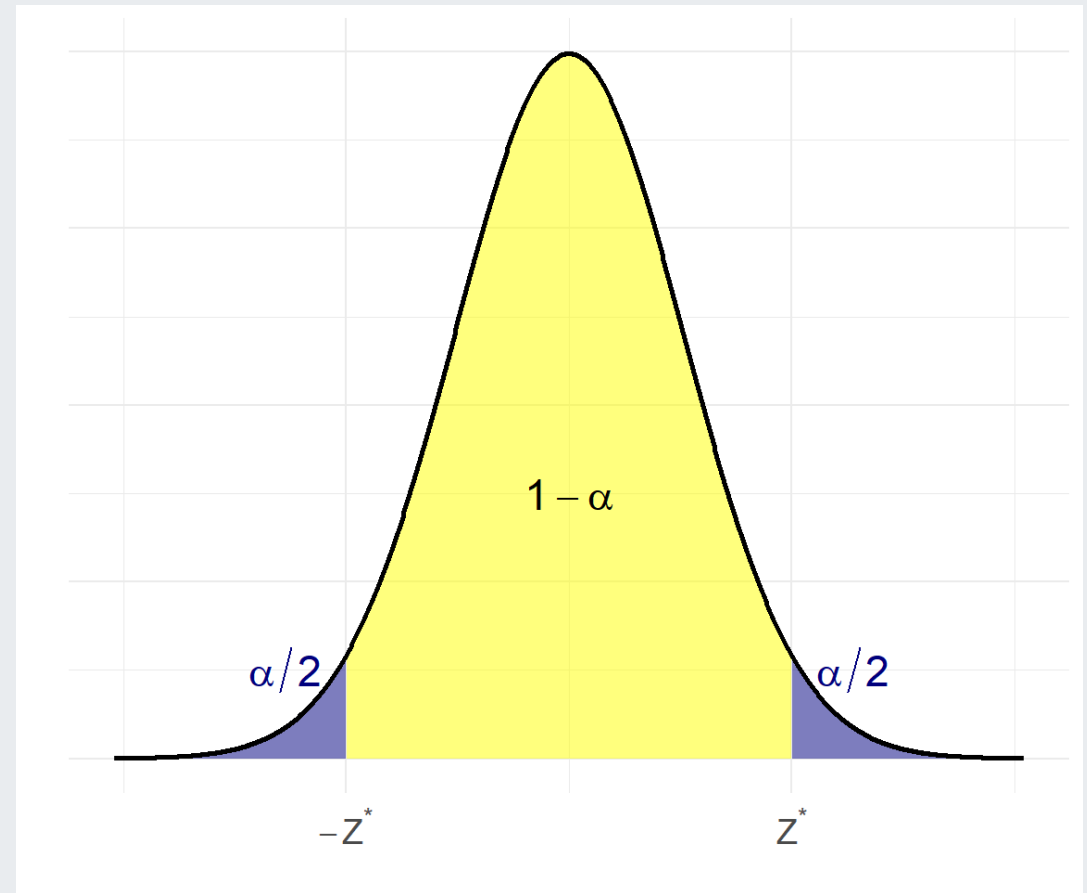
- Es el procedimiento para construir el intervalo del 95%
- Aproximadamente el 95% de los intervalos construidos siguiendo el procedimiento explicado anteriormente cubrirán el verdadero valor de la media poblacional
- Después de tomar la muestra y construir un intervalo, el intervalo construido cubre o no lo cubre. No lo sabemos. Solo Dios lo sabe.
- Es como la lotería: antes de elegir los números y comprar un boleto, tenemos alguna probabilidad de ganar el premio. Después de obtener el boleto, o ganamos o perdemos

Simulación de Intervalos de Confianza del 95% para 100 muestras



IC a otros niveles de confianza

Para un nivel de confianza , queremos encontrar el tal que:

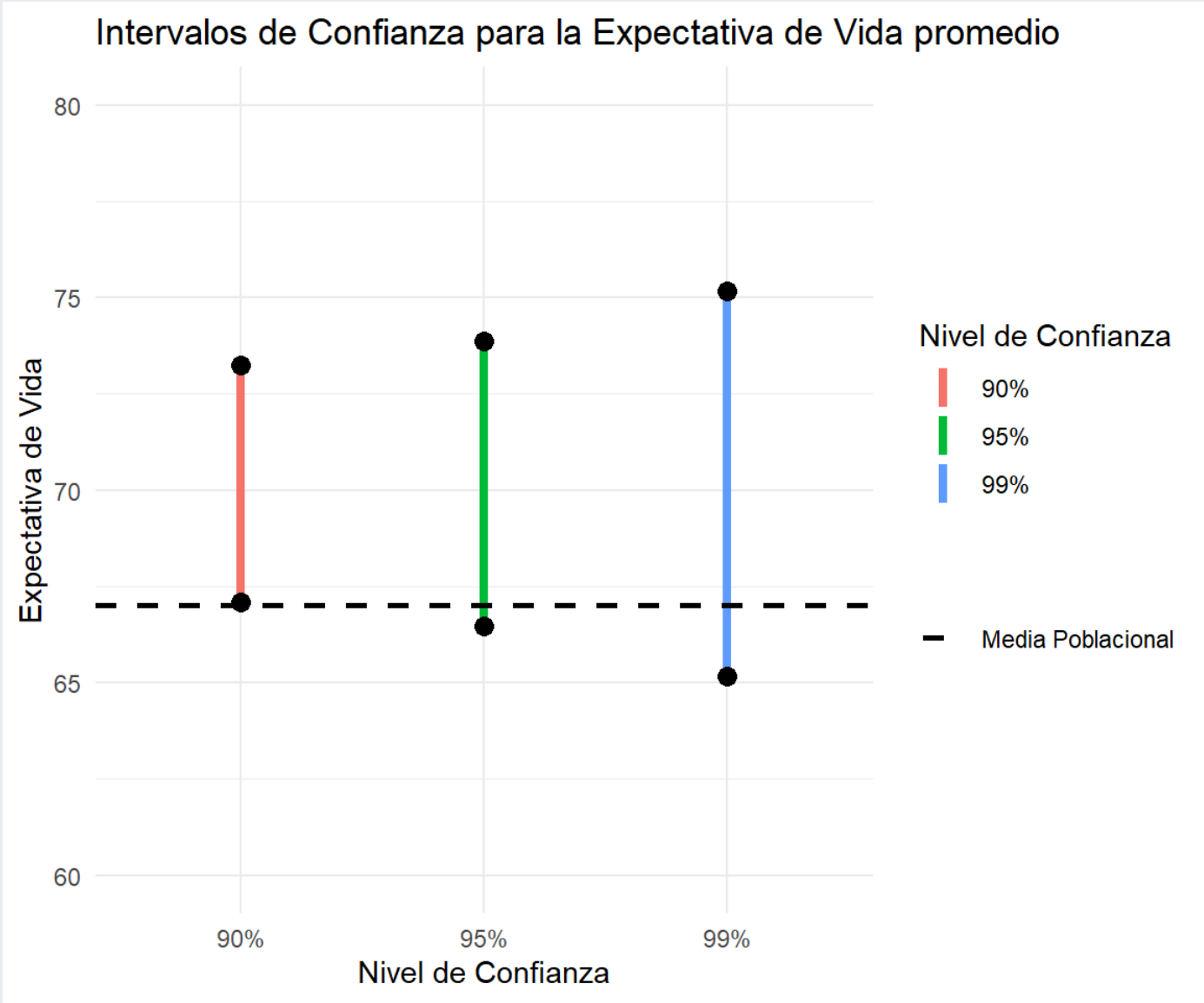


IC a otros niveles de confianza

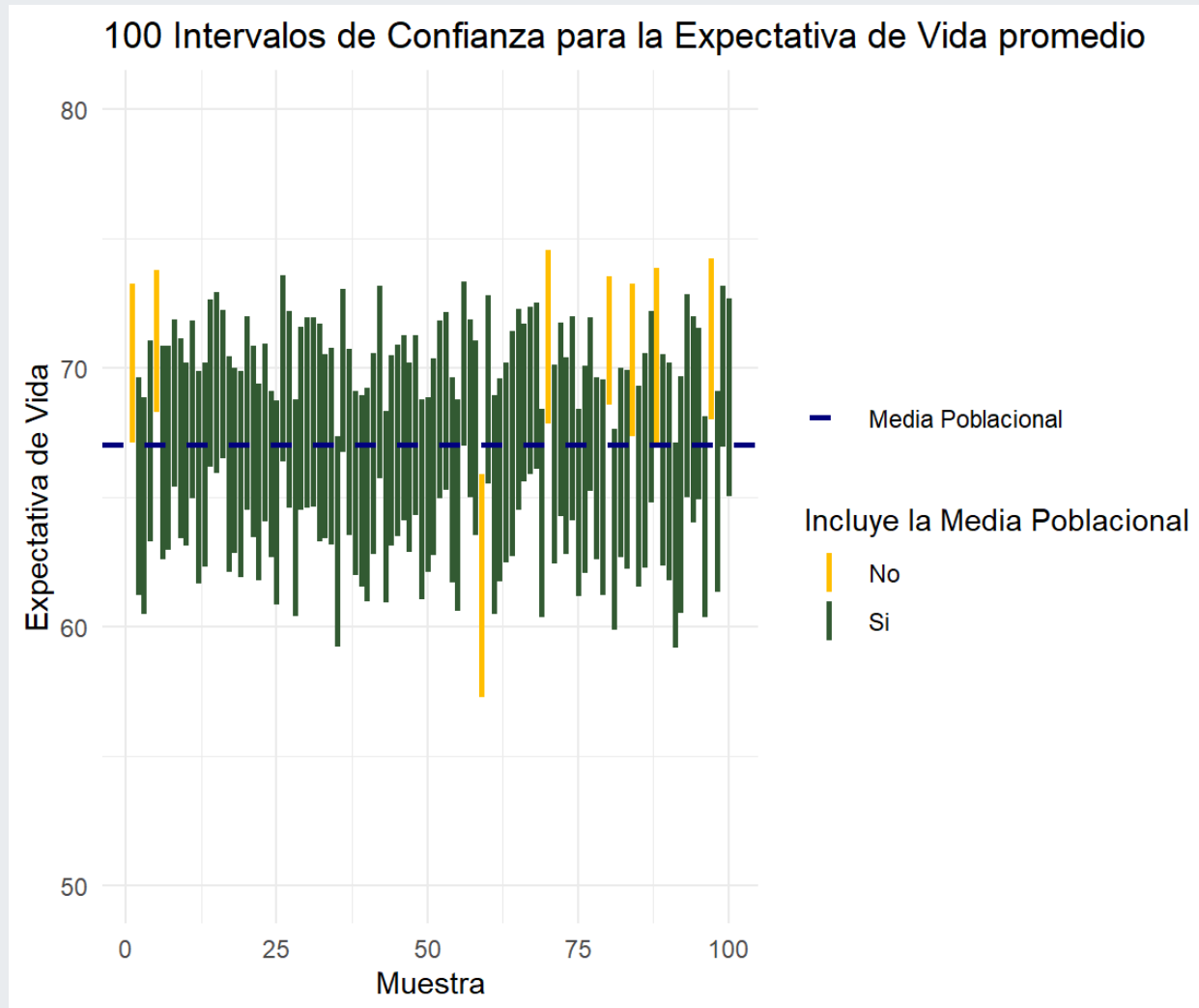
En general, un intervalo de confianza al nivel es

- es conocido como el margen de error
- Los niveles de confianza más usados son:
 - 90% IC: ,
 - 95% IC: ,
 - 99% IC: ,

IC con distintos niveles de confianza



Simulación de Intervalos de Confianza del 90% para 100 muestras



Menor Confianza, Mayor Riesgo de Error

¿Qué tan informativo es un IC?

Pero... un intervalo demasiado amplio no dice mucho

