

Analítica de los Negocios

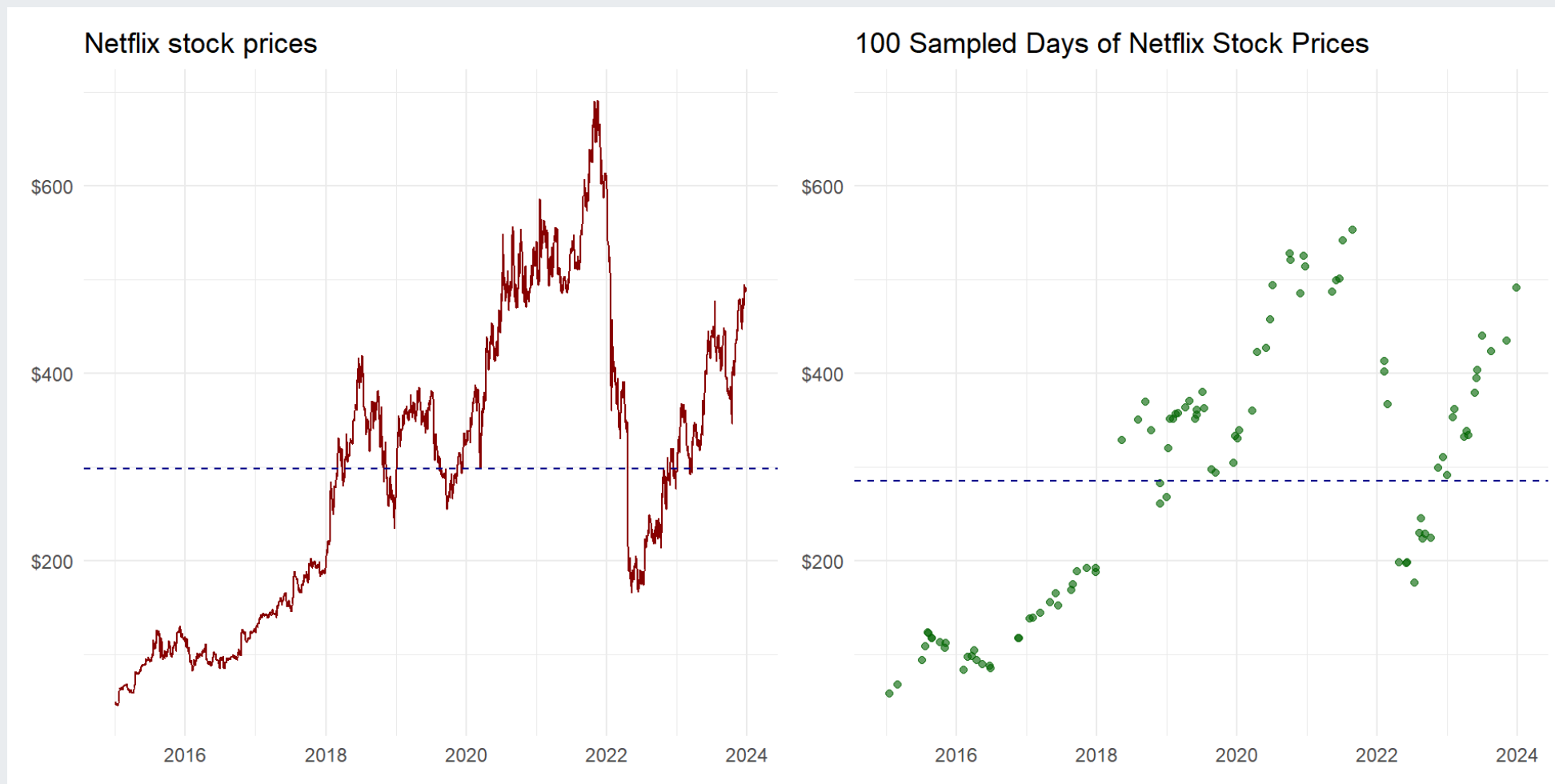
Pruebas de Hipótesis

Carlos Cardona Andrade

Plan para hoy

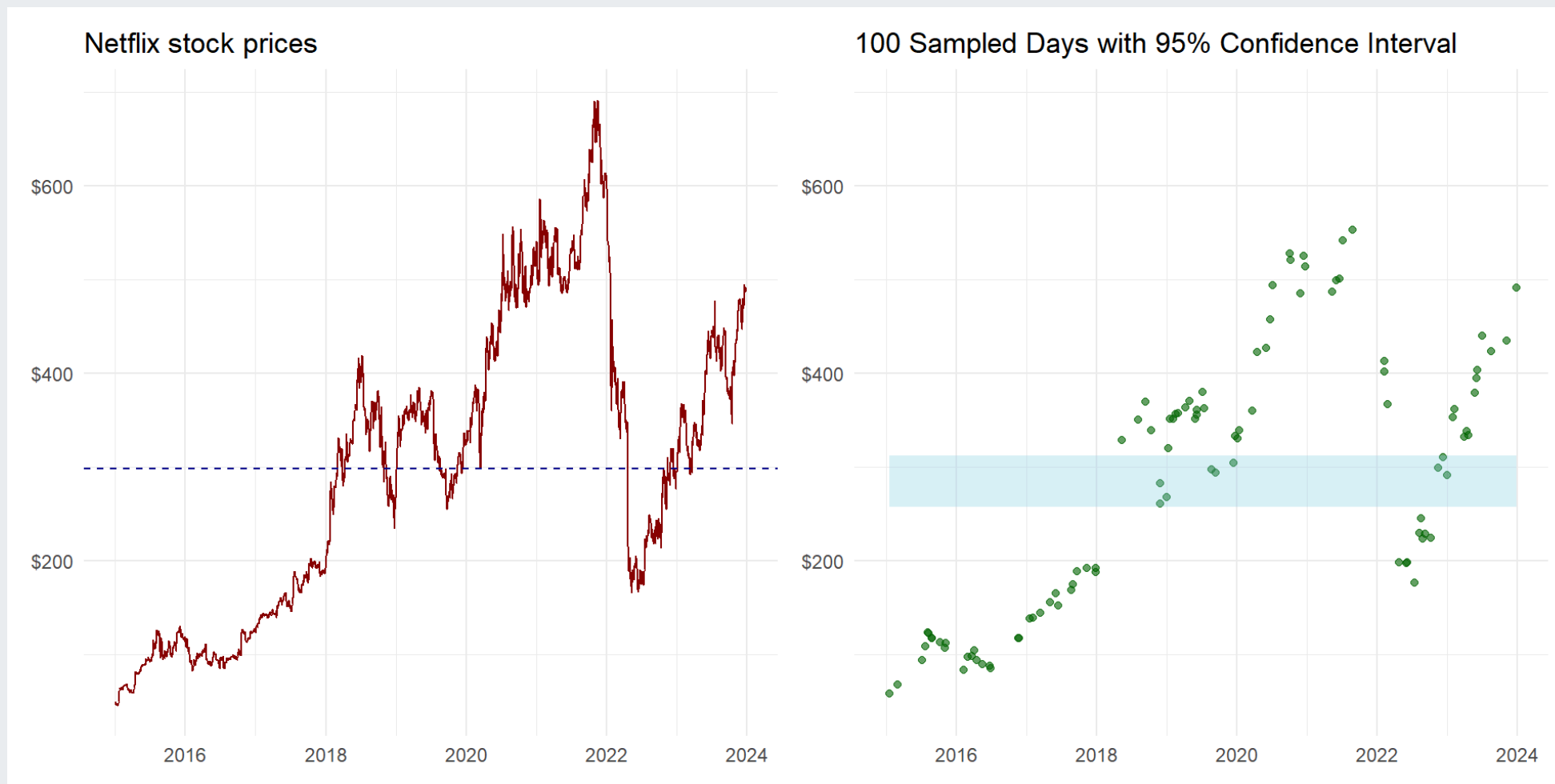
1. Intervalos de Confianza
2. Intro a Pruebas de Hipótesis
3. Pruebas de Hipótesis con dos muestras

Intervalos de Confianza - Muestra 1



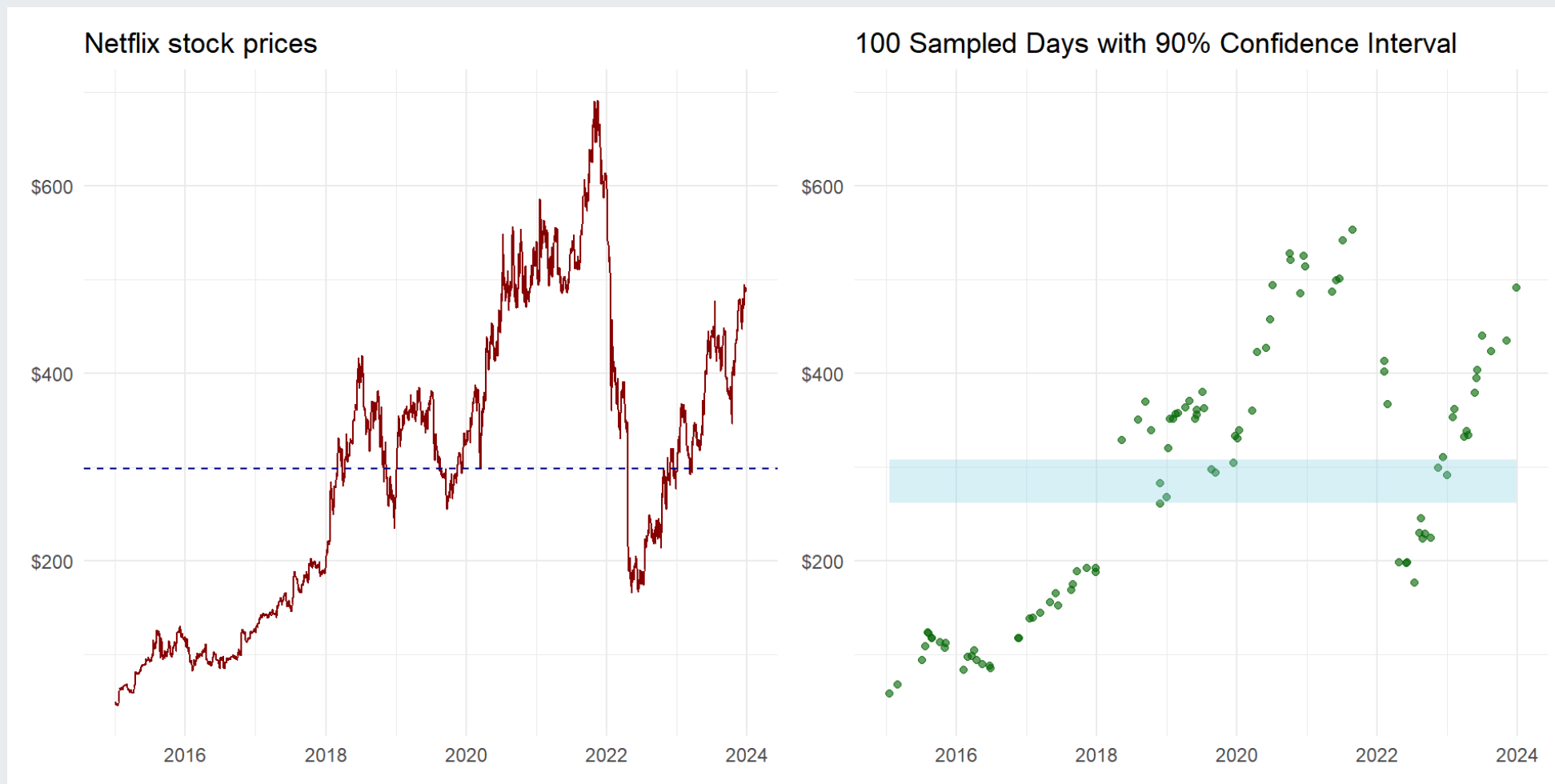
El precio promedio de las acciones de Netflix fue $\mu = 298.64$

Intervalos de Confianza - Muestra 1



El intervalo de confianza al 95% es [257.79, 312.49]

Intervalos de Confianza - Muestra 1



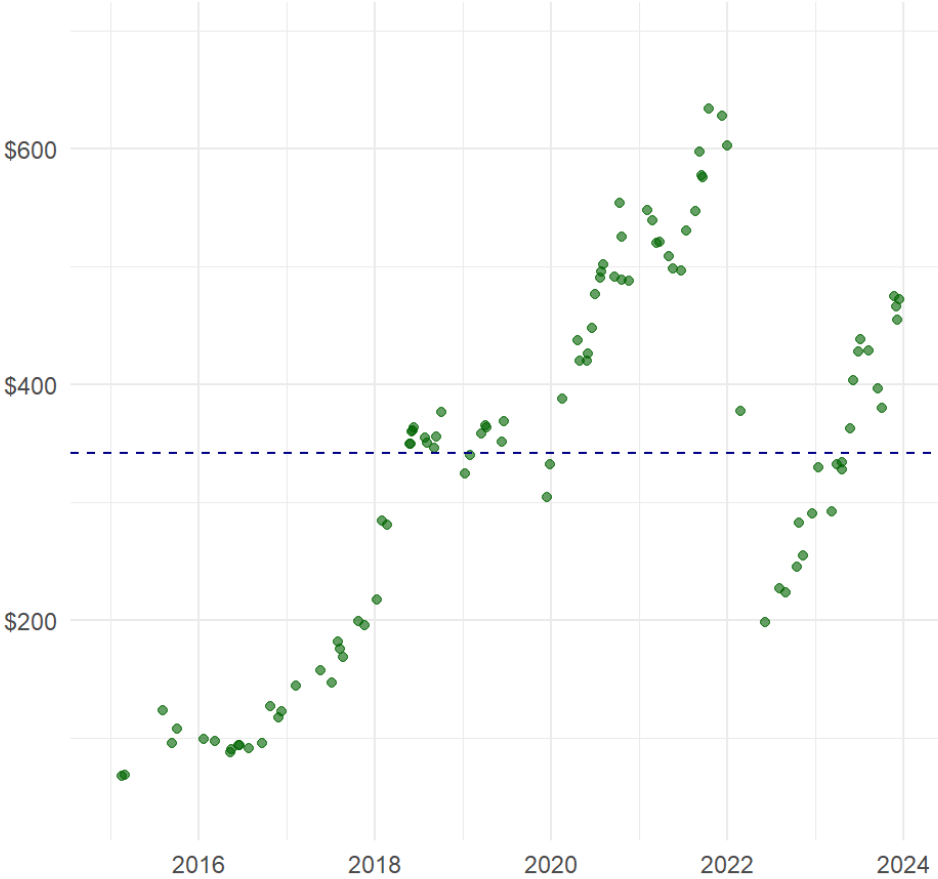
El intervalo de confianza al 90% es [262.25, 308.03]

Intervalos de Confianza - Muestra 2

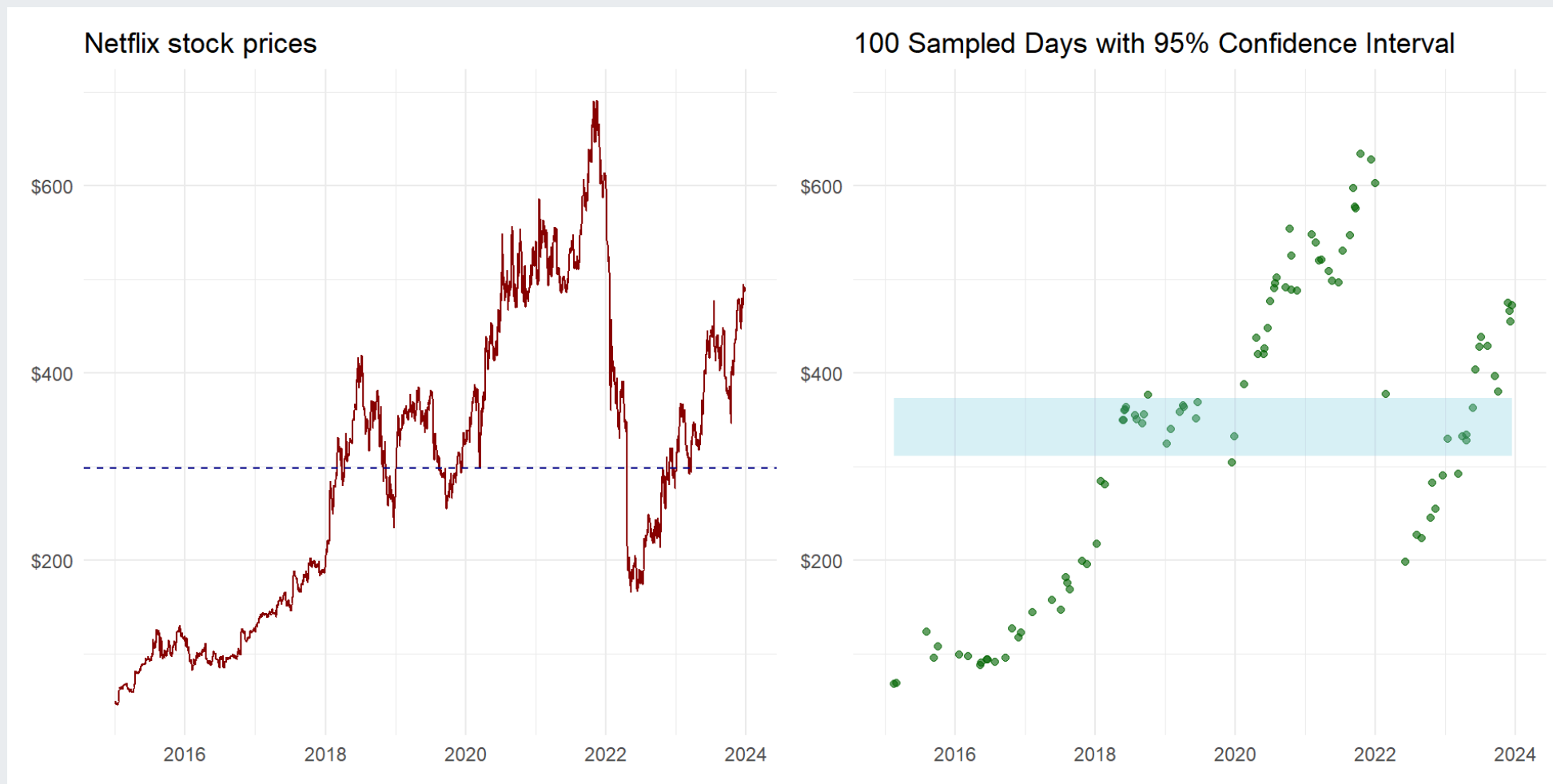
Netflix stock prices



100 Sampled Days of Netflix Stock Prices

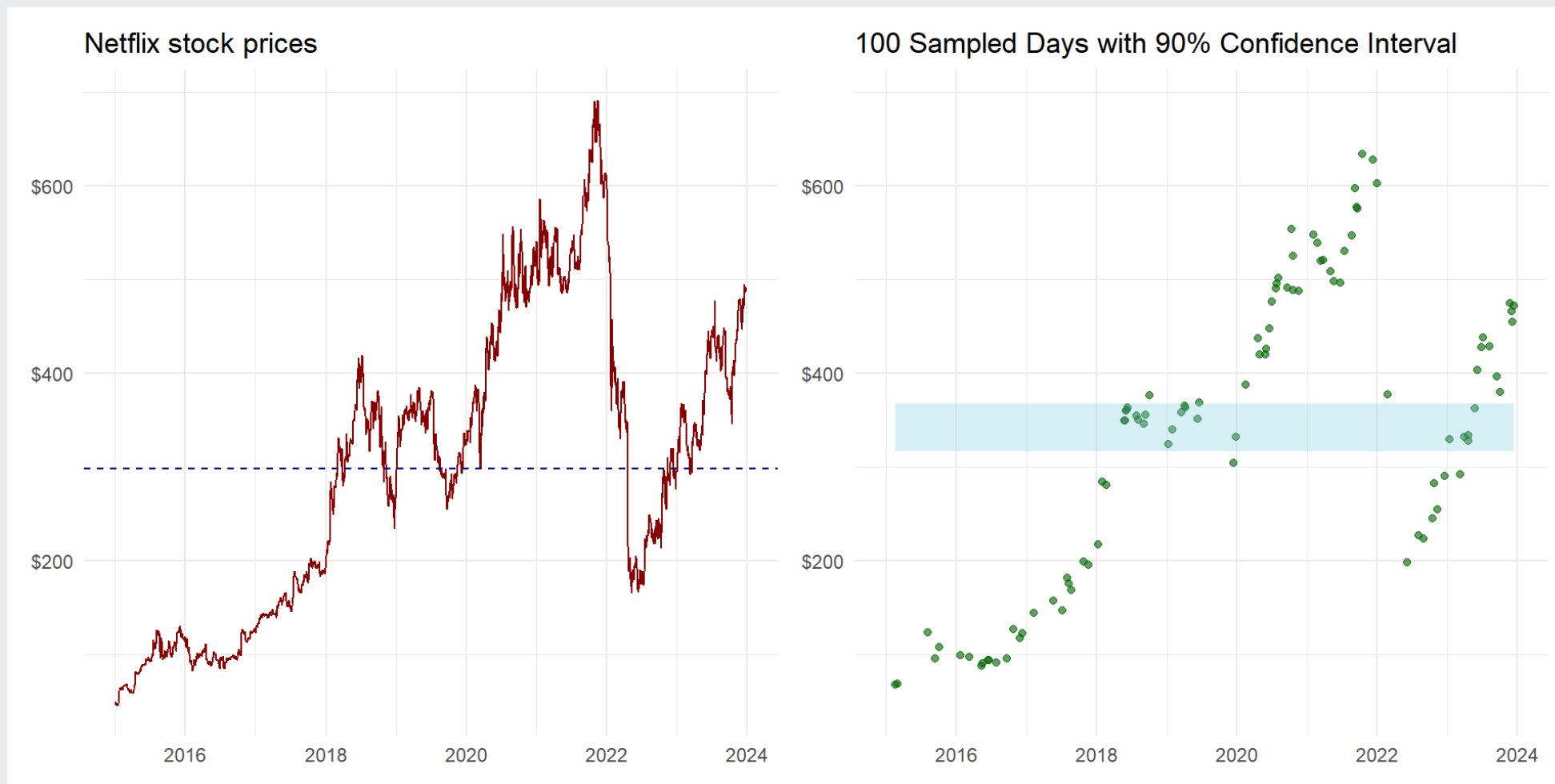


Intervalos de Confianza - Muestra 2



El intervalo de confianza al 95% es [311.46, 372.85]

Intervalos de Confianza - Muestra 2



El intervalo de confianza al 90% es [316.47, 367.84]

Intro a Pruebas de Hipótesis

Muriel Bristol y las tazas de té



En la década de 1920, la botánica inglesa Muriel Bristol afirmó que podía distinguir si la leche se había agregado antes o después del té en una taza. El estadístico Ronald Fisher, diseñó un experimento para poner a prueba su afirmación.

Fisher le presentó ocho tazas de té, cuatro con la leche agregada antes y cuatro después, y le pidió que identificara cuáles pertenecían a cada grupo.

¿Por qué es importante?

Este experimento marca el nacimiento del concepto de prueba de hipótesis en estadística. Fisher planteó dos hipótesis:

- Hipótesis nula (H_0): La mujer elige al azar
- Hipótesis alternativa (H_a): La mujer realmente distingue la diferencia

Si ella identifica correctamente las tazas más veces de lo que se esperaría por azar, Fisher rechazaría la hipótesis nula en favor de la alternativa.

¿Cuáles son los elementos de una prueba de hipótesis?

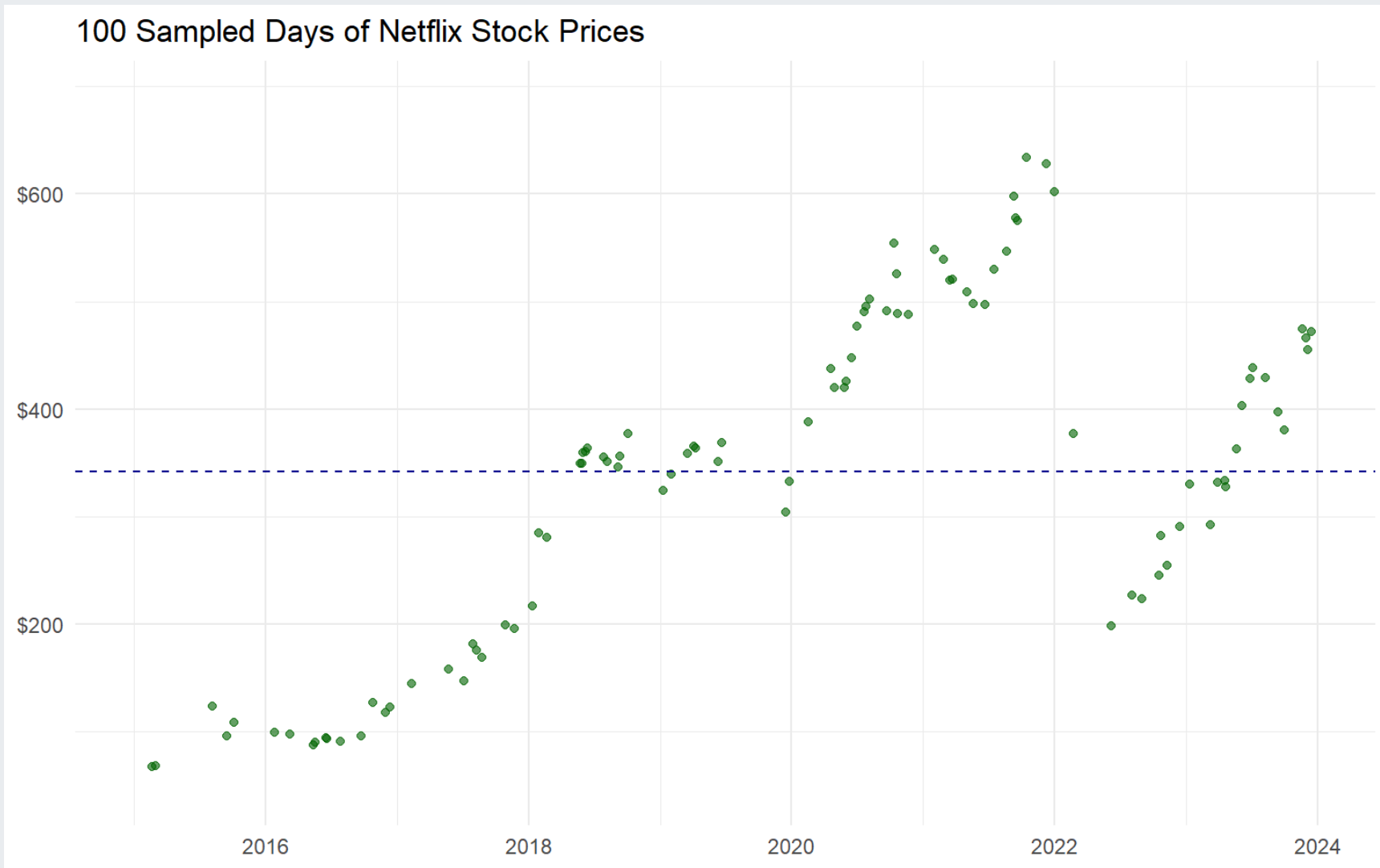
- Hipótesis nula (H_0): Es la suposición inicial que se pone a prueba.
- Hipótesis alternativa (H_a): Es la afirmación que se intenta demostrar.
- Estadístico de prueba: Un valor calculado a partir de los datos que permite evaluar la evidencia en contra de H_0 .
- Nivel de significancia (α): La probabilidad de rechazar cuando en realidad es verdadera.
- Regla de decisión: Un criterio basado en el estadístico de prueba para determinar si se rechaza o no H_0 .

Ejemplo: Precio de la acción de Netflix 2015-2023

- Retomemos el ejemplo de la muestra de los 100 días de precios de la acción de Netflix

```
1 library(tidyfinance)
2
3 # Descargamos los datos de 2015 a 2023
4 netflix_data <- download_data(
5   type = "stock_prices",
6   symbols = "NFLX",
7   start_date = "2015-01-01",
8   end_date = "2023-12-31"
9 )
10
11
12 # Establecemos la semilla
13 set.seed(10)
14
15 # Tomamos la muestra de 100 días
16 netflix_sample <- sample_n(netflix_data, 100)
```

¿Cómo es la distribución el precio de la acción de Netflix?



Ejemplo: Precio de la acción de Netflix 2015-2023

- Uno de sus colegas *cree* que el precio promedio de la acción de Netflix fue 310
- El precio promedio de la acción para la muestra es 342.16 y la desviación estándar es 154.71
- ¿Son estos datos suficientes para probar que el precio de la acción es mayor al precio que su considera fue el real durante ese periodo?
- Elaboremos una prueba de hipótesis con un nivel de significancia del 5% () para evaluar esta afirmación

Hipótesis

- **Población:** todos los días de 2015 a 2023
- El **parámetro de interés** es el precio promedio de la acción en *todos* los días de ese período
- Hay dos explicaciones de por qué la media muestral es mayor que el precio estimado por su colega:
 1. La media real de la población es diferente.
 2. La media real de la población es 310, y la diferencia entre la media real de la población y la media de la muestra se debe simplemente a la variabilidad natural del muestreo.
- (El precio promedio de la acción es 310)
- (El precio promedio de la acción es 310)

Maneras incorrectas de establecer las hipótesis

- y **SIEMPRE** se expresan en términos de parámetros de población, no de estadísticas de muestra.
- N_i :
- n_i :

el número de horas de estudio en la muestra es 310

el número de horas de estudio en la muestra es 310

son correctas.

La manera correcta de establecer las hipótesis

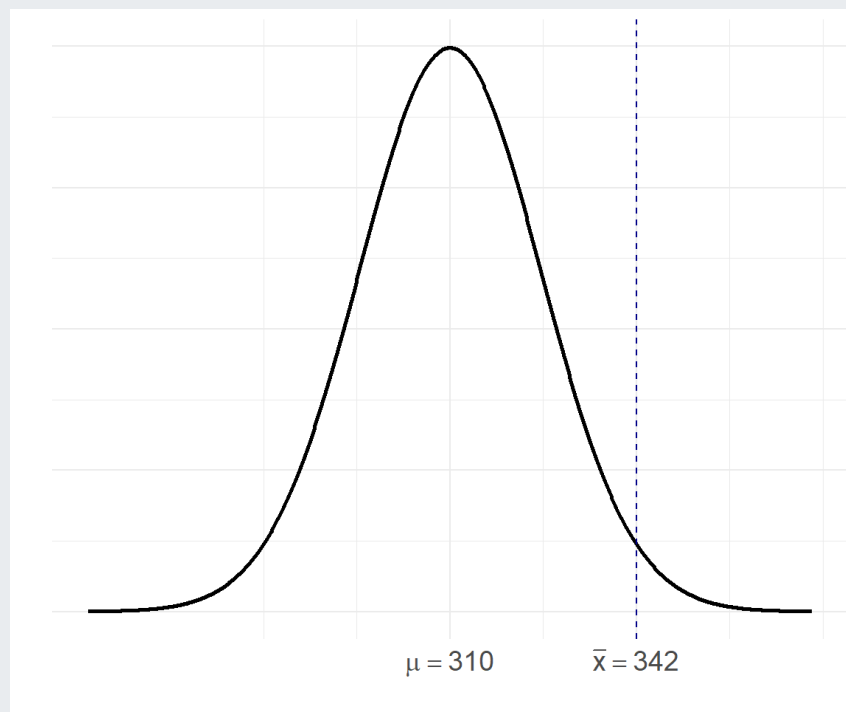
Las hipótesis son:

También siempre **especifiquen claramente** qué es

e.g., es el precio promedio de la acción de Netflix entre 2015 y 2023

El estadístico de la prueba

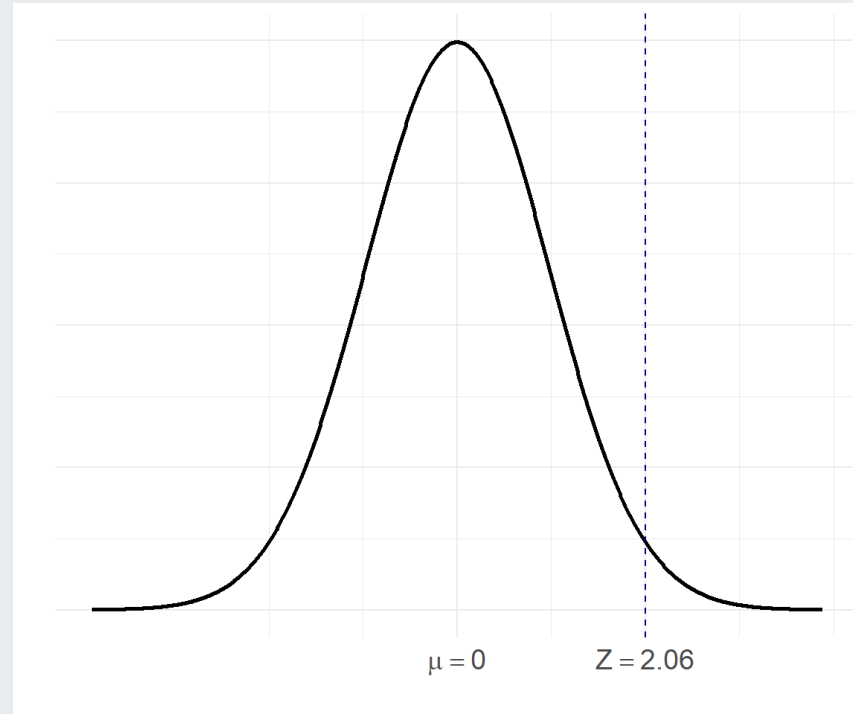
Por el TLC y asumiendo que es cierta, la distribución de la media muestral es:



Por ahora usamos el z-score

Para medir qué tan *inusual* es la media muestral observada en relación con su distribución muestral, el estadístico de prueba que usamos es el z-score.

El estadístico de la prueba



El estadístico de la prueba



- Las medias muestrales que son probables de obtener si es cierta son las medias muestrales cercanas a la hipótesis nula.
- Las medias muestrales que son poco probables de obtener si es cierta son aquellas lejanas a la hipótesis nula

¿ Qué significa “alta” y “baja” probabilidad?

- Esto se establece a partir de una probabilidad específica, la cual se conoce como nivel de significancia (se denota con α), para la prueba de hipótesis.
- El valor es una probabilidad pequeña que se utiliza para identificar muestras de poca probabilidad o inusuales.
- Por convención, los valores más comunes son:
- Por ejemplo, si usamos $\alpha = 0,05$, separamos el 5% de las medias más improbables (valores extremos) del 95% de las medias muestrales más probables (valores centrales).

Región y Z-scores críticos

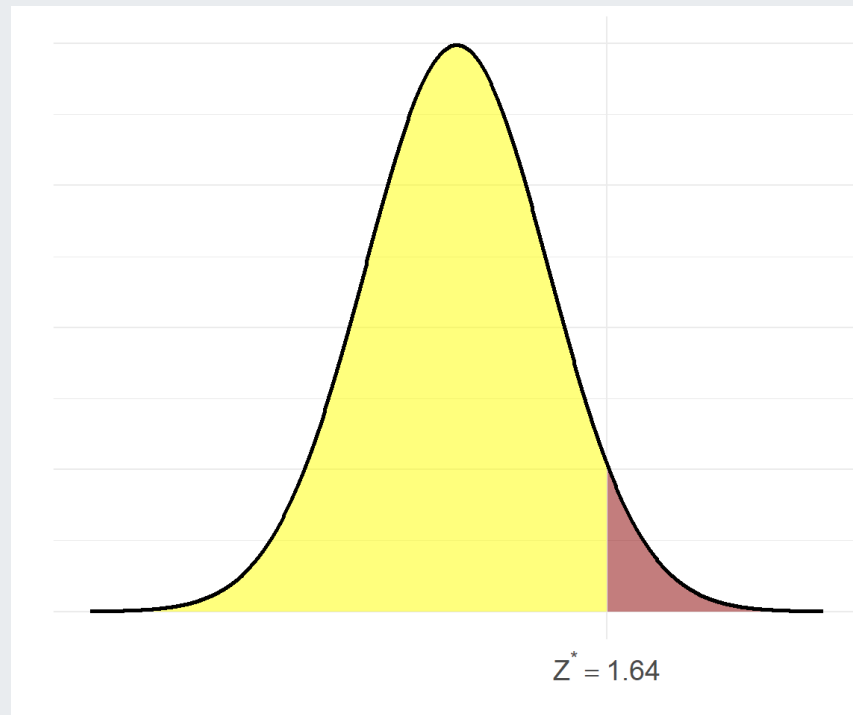
- Los valores extremos que son poco probables, definidos por el nivel de significancia, constituyen lo que se conoce como **región crítica**
- Estos valores son **inconsistentes** con la hipótesis nula
- Al igual que con los intervalos de confianza, para determinar la ubicación exacta de los límites se utilizan el z y la distribución normal para encontrar el z -score crítico.

¿Cuál es nuestro crítico?

Si nuestro $\alpha = 0.05$, el z-crítico será 1.64 de acuerdo a la tabla de la distribución normal

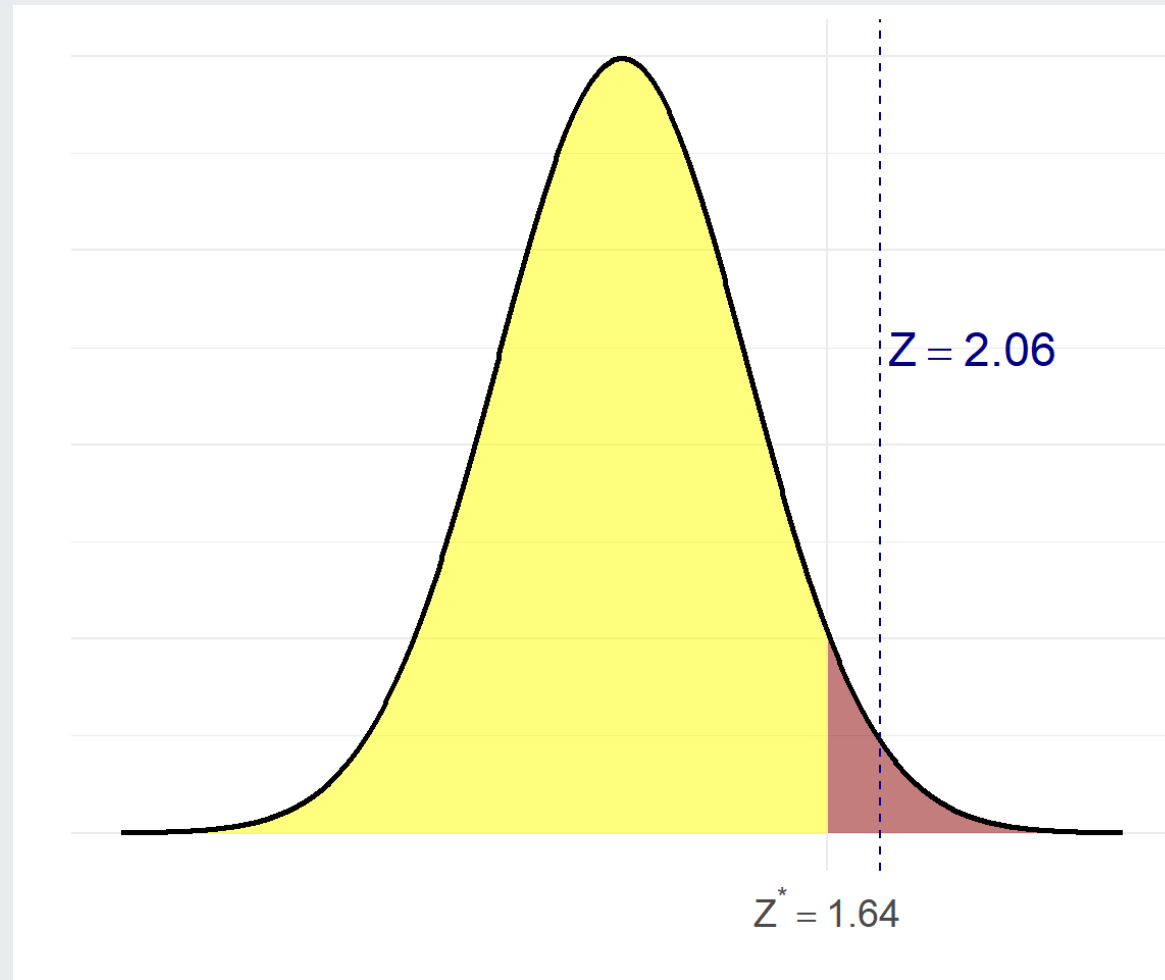
```
1 qnorm(0.95, mean = 0, sd = 1)
```

```
[1] 1.644854
```



¿Cuál es el resultado de la prueba?

La media muestral se ubica en la región crítica!



¿Cuál es el resultado de la prueba?

- Rechazamos
- Los datos proveen evidencia convincente de que el precio promedio de las acciones de Netflix entre 2015 y 2023 es mayor a 310
- En otras palabras, la media muestral es **estadísticamente diferente** de 310

Prueba de hipótesis a dos colas

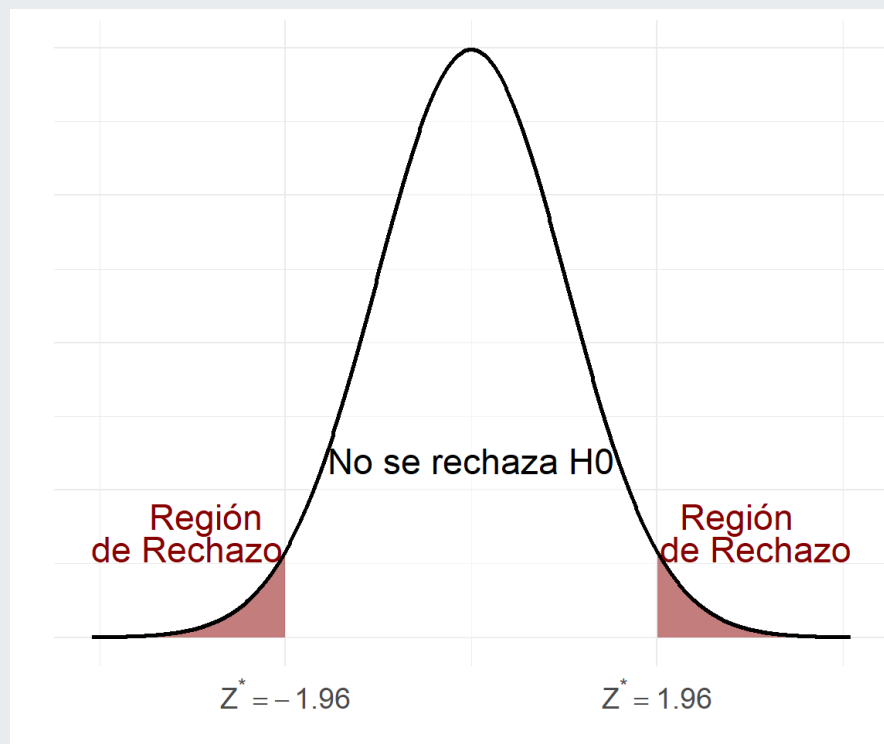
Si quisiéramos saber si los datos proveen evidencia consistente que el precio promedio de las acciones de Netflix es **diferente** que el 310 descrito por el colega, la hipótesis alternativa cambiaría:

Prueba de hipótesis a dos colas: z-crítico

Si nuestro $\alpha = 0.05$, el z-crítico será 1.96 de acuerdo a la tabla de la distribución normal

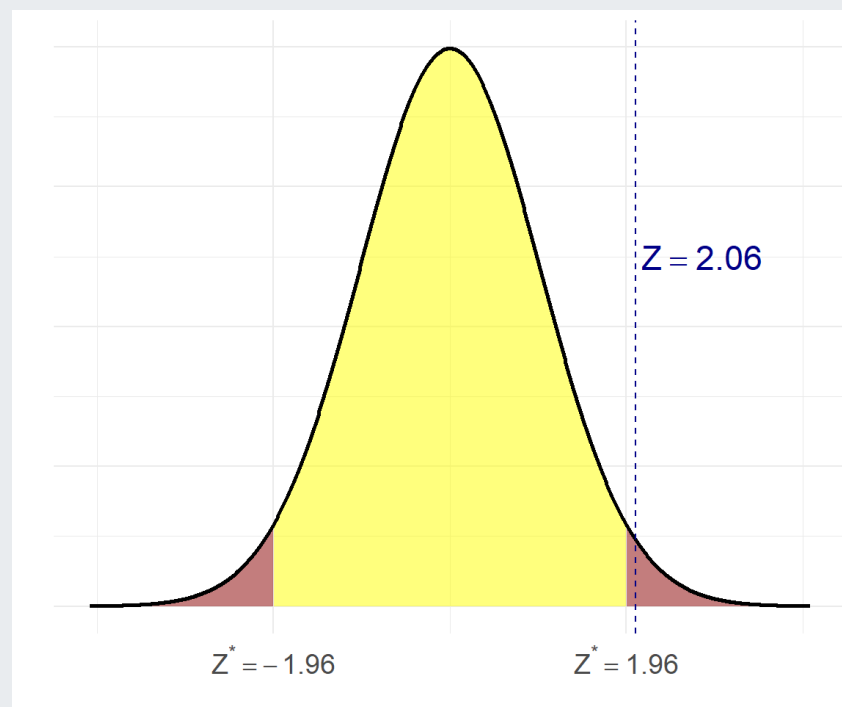
```
1 qnorm(0.975, mean = 0, sd = 1)
```

```
[1] 1.959964
```



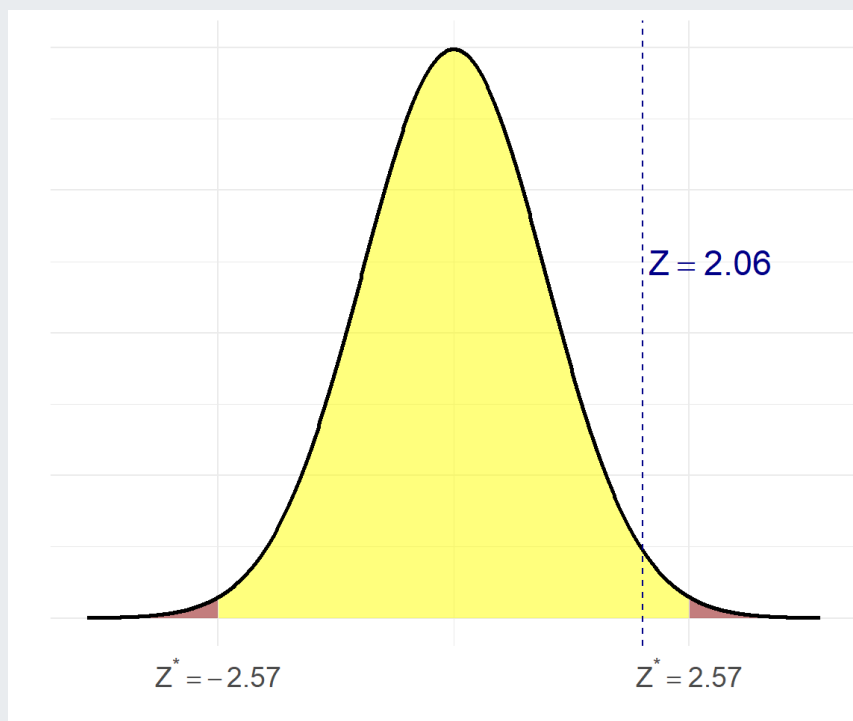
Prueba de hipótesis a dos colas: resultado al 5%

- En este caso, una media muestra mucho menor a 310 también sería evidencia en favor de
- Como , seguimos rechazando



Prueba de hipótesis a dos colas: resultado al 1%

- ¿Qué pasa con el resultado de la prueba si el nivel de significancia cambia a ?
- Como , ya no rechazamos



IC y prueba de hipótesis a dos colas

En una prueba de hipótesis a dos colas:

No rechazar la hipótesis nula es equivalente a que esté dentro del intervalo de confianza del % para μ . Es decir:

IC y prueba de hipótesis a dos colas

Retomemos:

Rechazamos
cuando:



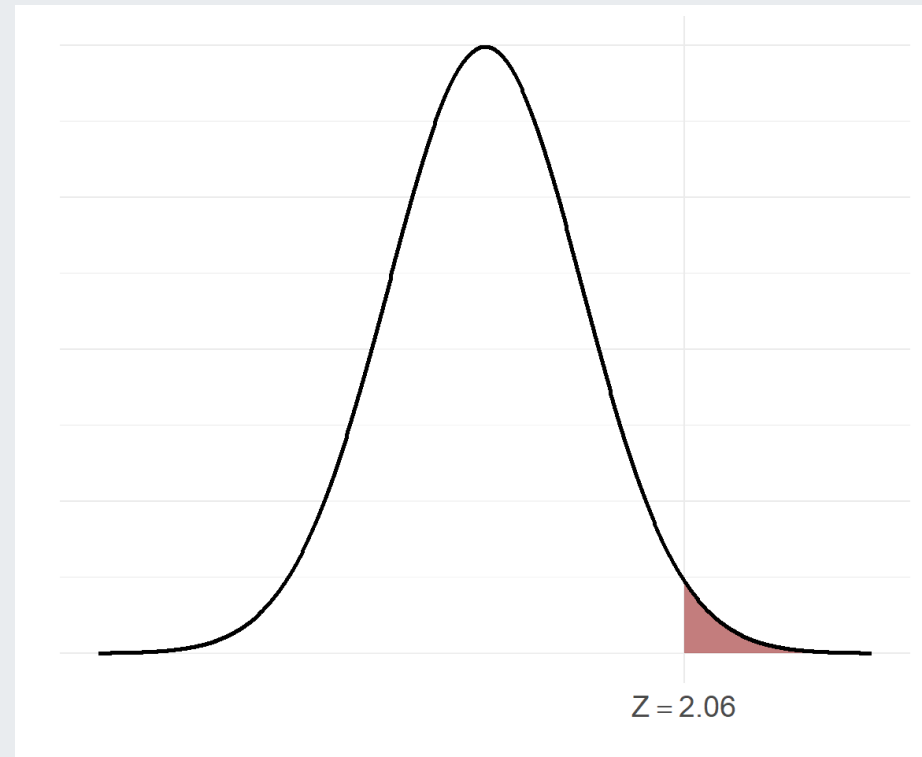


Ejercicio 1

1. Utilicen el código de la diapositiva 14 para generar una muestra aleatoria de 100 días pero con `set.seed(1)`
2. Usando el comando `t.test()`, evalúen nuevamente la hipótesis de que el precio promedio de la acción entre 2015 y 2023 fue 310 dólares:
 - A un nivel de significancia del 5%
 - A un nivel de significancia del 1%

p-value

- El z-score tiene una probabilidad asociada dada la forma de la distribución normal.
- Por ende, la decisión de una prueba de hipótesis puede basarse tanto en el z-score como en su probabilidad asociada (p-value).



```
1 pnorm(2.06, mean = 0, sd = 1, lower.tail = FALSE)
[1] 0.01969927
```

p-value

- Una prueba de hipótesis se rechaza si:
- Para nuestro , sabemos que
- Si $\alpha=0.05$, se rechaza la hipótesis nula dado que
- Esto sucede para un nivel de significancia del 5%, ¿qué sucede si este cambia a 1%?

El p-value es la probabilidad de observar nuestros datos ($\chi^2=342$) si la hipótesis nula fuera cierta ()

Resumen - Prueba de hipótesis

1. Establezca las hipótesis

-
- $< 0 > 0$

2. Revise los supuestos y condiciones

- Independencia
- Normalidad:

3. Calcular el z-score y el p-value (dibujen la gráfica!)

4. Tomen una decisión:

- Si : Se rechaza
- Si : No se rechaza

Pruebas de Hipótesis con dos muestras

Prueba de hipótesis con dos muestras

- ¿Las empresas que cotizan en NYSE tienen mayores rendimientos promedio de acciones que las que cotizan en NASDAQ?
- ¿Las tasas de interés de las hipotecas ofrecidas por el banco A son más bajas que las ofrecidas por el banco B?
- ¿Los salarios promedio de los empleados en empresas tecnológicas son diferentes de los de las empresas manufactureras?
- El objetivo ahora es comparar las medias (o alguna cantidad) y de dos poblaciones

Muestras independientes y relacionadas

- Para dos poblaciones de interés y , las hipótesis en este caso son:
 -
 -
- Las muestras pueden ser:
 1. Independientes: medición de unidades en distintos grupos.
 2. Relacionadas: medición de la misma unidad antes y después de alguna intervención/suceso.

Ejemplo: Apple and Amazon stock prices

- Volvamos a los datos de precios de las acciones en el período 2015-2023.
- Seleccionemos una muestra de 100 días de los precios de Apple y Amazon.
- Basádos en la muestra, ¿fueron iguales los precios promedio de estas dos compañías durante este período?

Ejemplo: Apple and Amazon stock prices

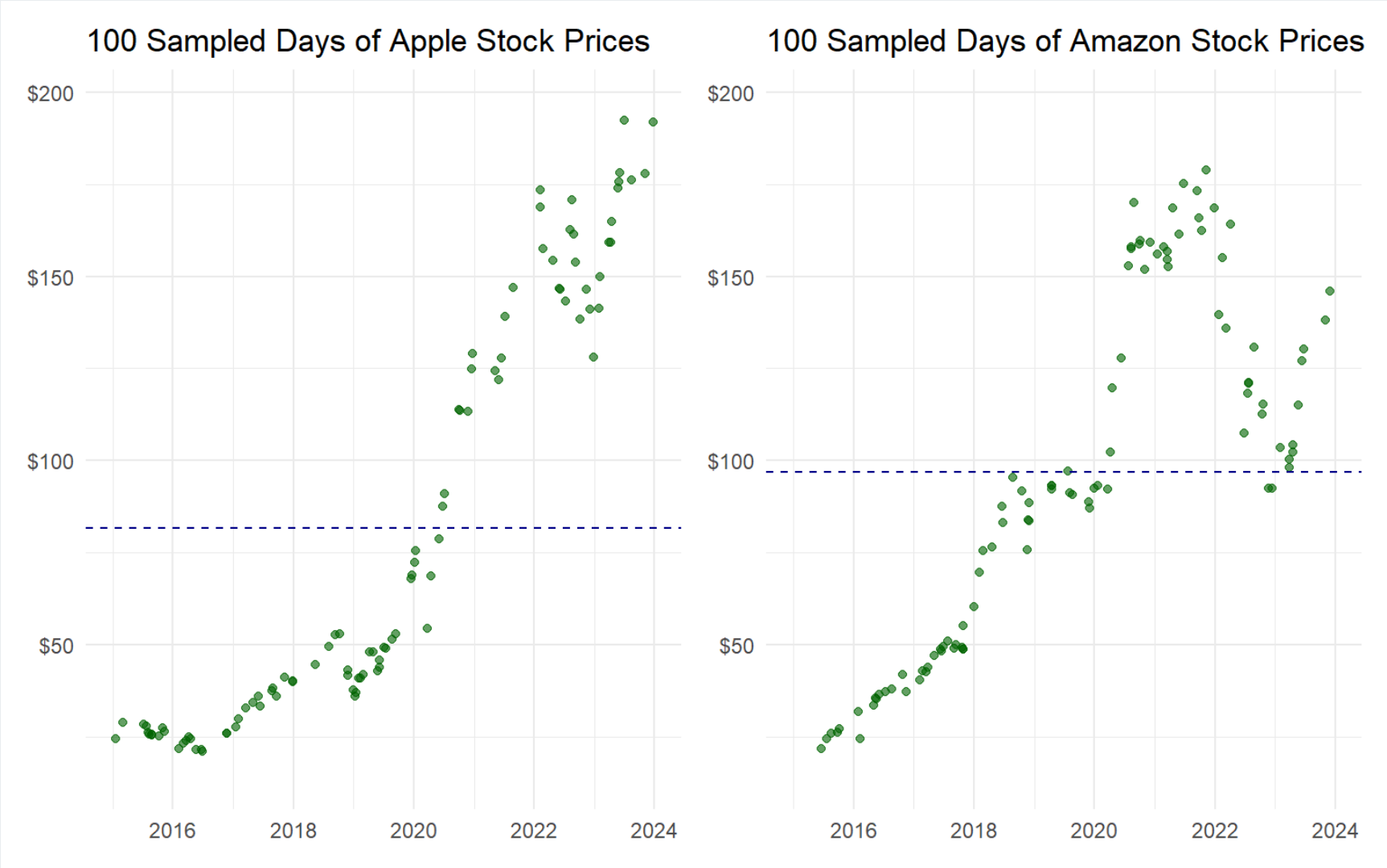
```
1 prices_data <- download_data(  
2   type = "stock_prices",  
3   symbols = c("AAPL", "AMZN"),  
4   start_date = "2015-01-01",  
5   end_date = "2023-12-31"  
6 )  
7  
8 # Establecemos la semilla para reproducibilidad  
9 set.seed(123)  
10  
11 # Tomamos una muestra de 100 días para cada empresa  
12 prices_sample <- prices_data |>  
13   group_by(symbol) |>  
14   sample_n(100)  
15  
16 # Verificamos la muestra  
17 nrow(prices_sample)
```

```
[1] 200
```

Ejemplo: Apple and Amazon stock prices



Ejemplo: Apple and Amazon stock prices



Prueba de hipótesis en R

Nuestra hipótesis nula es que no existe diferencia en el precio promedio de las acciones de ambas empresas durante el periodo.

Prueba de hipótesis en R

Nuevamente usamos el comando `t.test()`:

```
1 # Calculemos el test
2 test_resultado <- t.test(adjusted_close ~ symbol, data = prices_sample, con
3 test_resultado
```

Welch Two Sample t-test

```
data: adjusted_close by symbol
t = -2.0925, df = 191.3, p-value = 0.03771
alternative hypothesis: true difference in means between group AAPL and group
AMZN is not equal to 0
95 percent confidence interval:
 -29.9290902 -0.8839648
sample estimates:
mean in group AAPL mean in group AMZN
      81.52729      96.93382
```

Observen que el intervalo de confianza no incluye al cero. Por lo tanto
Rechazamos

Prueba de hipótesis en R

¿Cuánto es el p-value? ¿Es ?

```
1 test_resultado$p.value
```

```
[1] 0.03771285
```

```
1 test_resultado$p.value < 0.05
```

```
[1] TRUE
```

Ejercicio 2

1. Utilicen el código de la diapositiva 42 para generar una muestra aleatoria de 100 días para dos empresas diferentes. Pueden encontrar las siglas de las dos empresas acá: [Yahoo Finance](#).
2. Usando el comando `t.test()`, evalúen nuevamente la hipótesis de que el precio promedio de ambas acciones entre 2015 y 2023 fue diferente:
 - A un nivel de significancia del 5% y utilizando el p-value para la decisión de rechazar o no
 - A un nivel de significancia del 1% y utilizando el intervalo de confianza para la decisión de rechazar o no

Errores Tipo I y II

	Decisión: Rechaza	Decisión: No Rechazar
es verdadera	Error Tipo I (Falso Positivo) 😊	
es falsa	😊	Error Tipo II (Falso Negativo)

- (Casi) Nunca sabremos si o son verdaderas, pero se necesita considerar todas las posibilidades

Errores Tipo I y II





Ejercicio 3

1. Retomemos la situación en la que su colega piensa que el precio promedio de la acción de Netflix entre 2015 y 2023 fue de 310 dólares y nuestra muestra indica que el precio es mayor. En este contexto, ¿sería preferible cometer un error tipo I o tipo II? ¿Por qué?

