Analítica de datos

Intervalos de confianza y pruebas de hipótesis

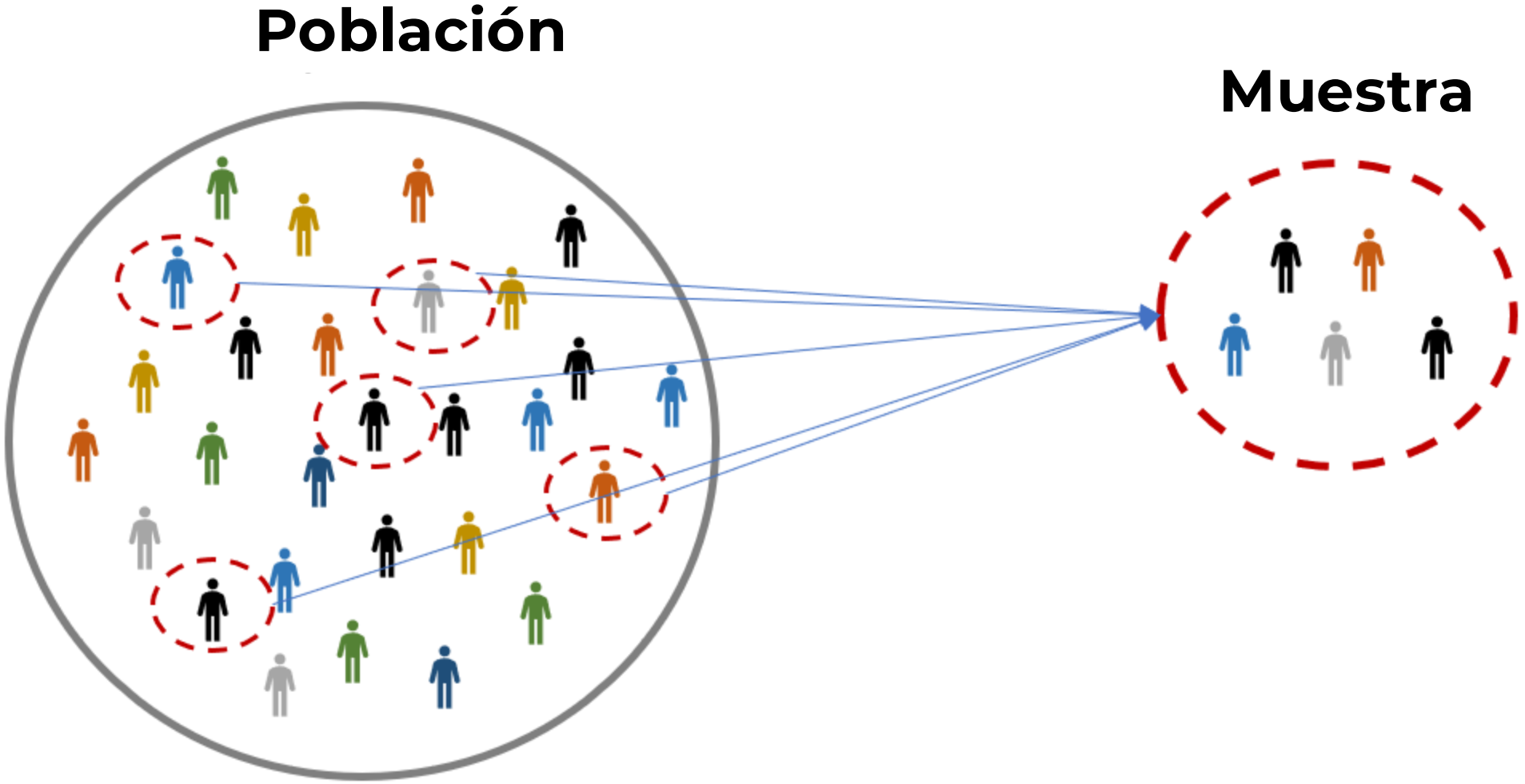
Pontificia Universidad
JAVERIANA
Bogotá

Profesor: Nicolás Velásquez

Intervalos de confianza para la media muestral

- La media muestral es la que uno calcula con su muestra de datos, como lo hicimos la clase pasada.

- Un intervalo de confianza toma en cuenta que la media muestral es un estimado de la media poblacional y nos da un rango de valores que podría tomar la media poblacional.


Población
Muestra

- **Ejemplo:**

  - Queremos **estimar el salario medio poblacional** de las personas que viven en Chapinero Alto (y que tienen empleo remunerado con salario).

  - Población: personas en Chapinero Alto con salario.

  - **Tomamos una muestra** de 200 personas.

  - **Contruimos la media muestral**, que es un aproximado de la poblacional (pero no es la poblacional).

- Obtenemos media muestral: $6.000.000

- **Con los datos, podemos también constuir un intervalo de confianza para la media poblacional**:

  “La media poblacional está entre $5,120,000 y $6.880.000 con 95% de confianza.”

- Esto quiere decir que es muy posible que la verdadera media poblacional esté en ese intervalo.

- Desde nuestra perspectiva, la probabilidad de que la media poblacional esté en ese intervalo es de 95%.

La razón por la que podemos construir dicho intervalo es porque la media muestral es una variable, cuyo valor depende (en parte) de la suerte, y que tiene una distribución de forma conocida...

Vamos a explorar un poco este asunto, antes de pasar a calcular intervalos de confianza para la media...

Distribución de la media muestral

- La distribución de la media muestral es una distribución de todos los posibles valores que puede tomar la media.

- ¡La media muestral es una variable cuyo valor específico depende de la suerte!

- Un ejemplo....

“Deduciendo” la distribución de la media muestral

- **Suponemos que hay una población.**
- Tamaño N=4.
- La variable de interés,

  es X, la edad de un individuo.
- X puede tomar los valores:

  18, 20, 22, 24 (años).

“Deduciendo” una distribución muestral

```r
# datos
edades <- c(18, 20, 22, 24)

# Calcular la media
media <- mean(edades)
print(paste("La media es:", media))

# Calcular la desviación estándar
desviacion <- sd(edades)
print(paste("La desviación estándar es:", desviacion))
```

Media poblacional = 21

Desviación estándar poblacional: 2.5819

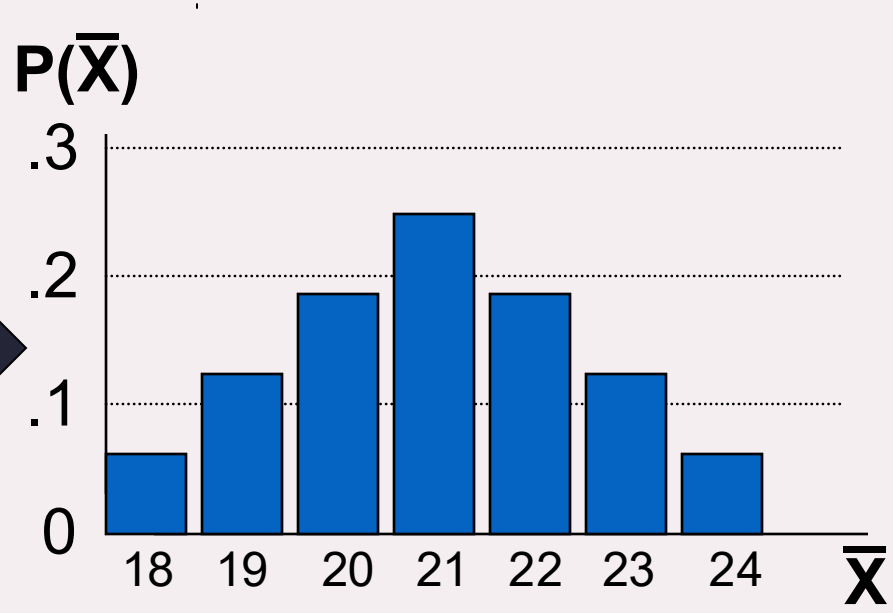
“Deduciendo” una distribución muestral

Distribución de la media muestral con n=2.

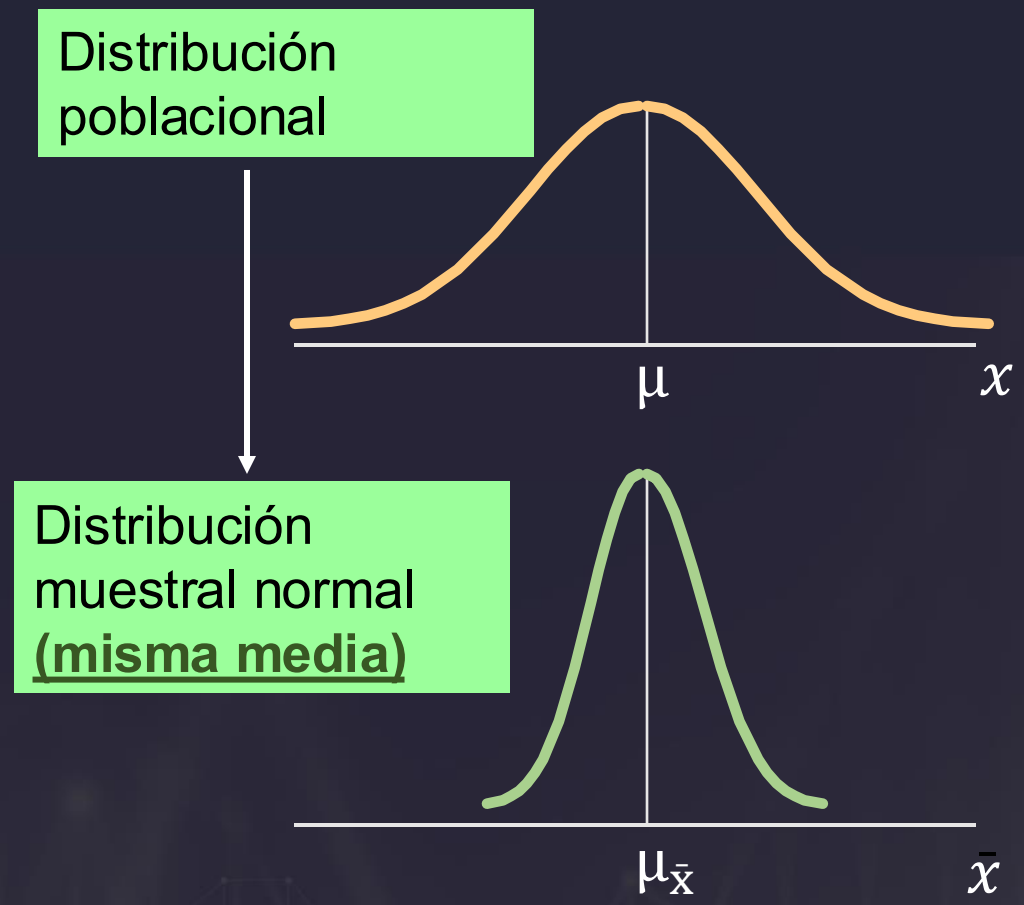
16 posibles medias muestrales

1er Obs	2a Obs			
	18	**20**	**22**	**24**
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

Distribución muestral



- Sabemos la distribución de la media muestral (básicamente una distribución “normal”)
- Dicha distribución tiene una media igual a la media poblacional que es lo que queremos estimar.
- Por lo tanto, podemos hacer inferencia de la media poblacional a partir de la distribución de la media muestral.



Supuesto importante:

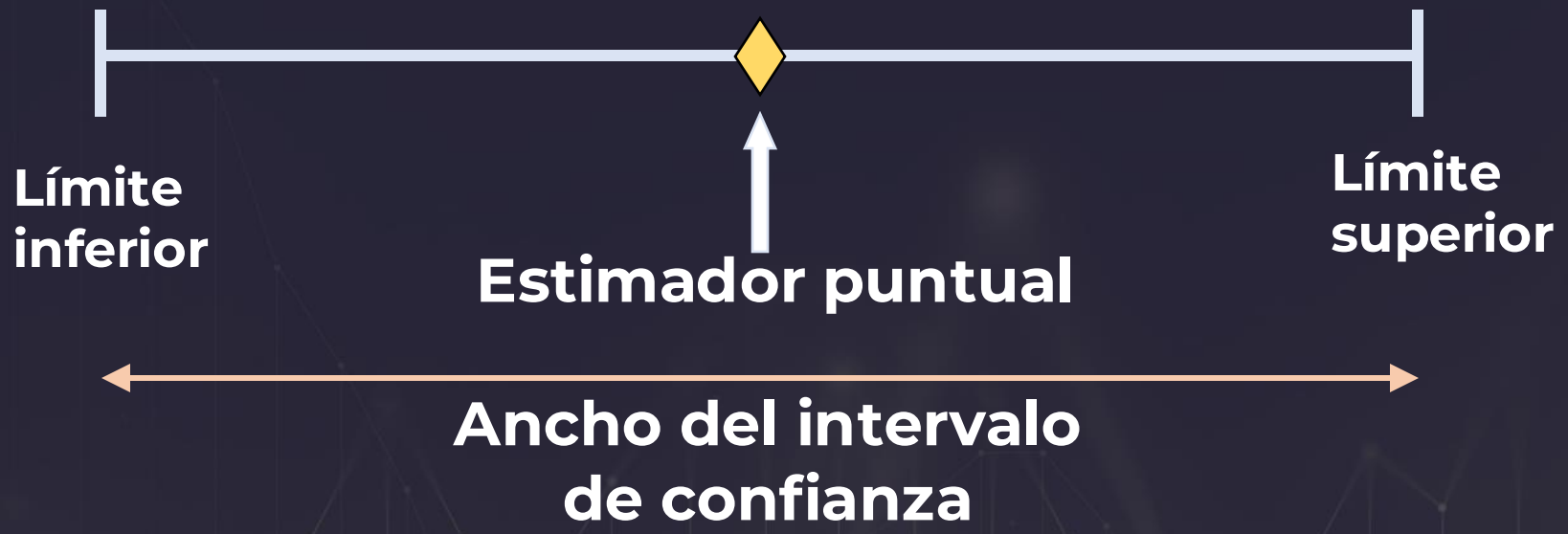
De aquí en adelante, una muestra consiste de más de **30** observaciones.

A mayor tamaño de la muestra más precisa es nuestra estimación

Ahora sí, veamos qué es un intervalo de confianza...

Estimación puntual e intervalos de confianza

- Un estimador puntual es UN número. En nuestro caso, la media muestral.
- Un intervalo de confianza da mayor información sobre la variabilidad del estimado (la media muestral):



Intervalos de confianza

- ¿Qué tanta incertidumbre está asociada con un estimador puntual de un parámetro poblacional?

- Un intervalo de confianza provee mayor información sobre la incertidumbre con la que estamos calculando la media muestral:

  - Entre más pequeño el intervalo, menos incertidumbre.

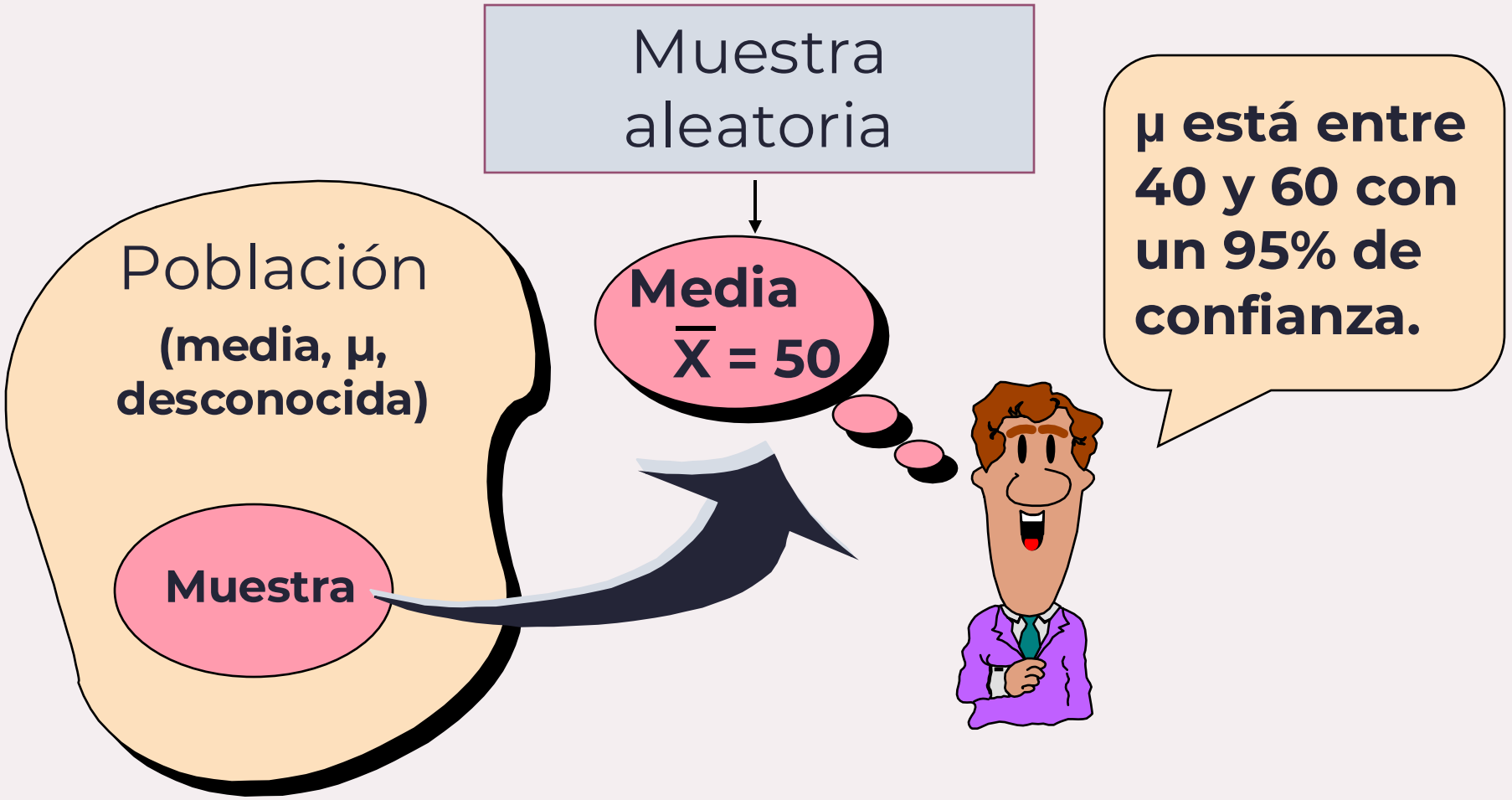
Ejemplo

- El nivel de confianza nos dice la probabilidad, desde nuestra perspectiva, que la verdadera media poblacional esté en el intervalo.

- **En la práctica** solamente se toma una muestra.

- Basándonos en la muestra, podemos decir que tenemos 90% de confianza de que el intervalo contiene **la verdadera media poblacional** (para un intervalo de confianza del 90%, como en el ejemplo).

Proceso de estimación



Fórmula general

- La fórmula general para todos los intervalos de confianza es:

Estimado (media muestral) ± Margen de error

```
# Primero creemos unas variables
p <- 0.45
N <- 1000
```

Fórmula general

Estimado (media muestral) ± Margen de error

```
# Primero incluyamos unas variables
p <- 0.45
N <- 1000

# Ahora calculemos el intervalo
# la función sample() que se utiliza para generar una muestra aleatoria.

x <- sample(c(0, 1), size = N, replace = TRUE, prob = c(1-p, p))
x_hat <- mean(x)
se_hat <- sqrt(x_hat * (1 - x_hat)/ N)
c(x_hat - 1.96 * se_hat, x_hat + 1.96 * se_hat)
```

- Si repiten este código varias veces y creando intervalos y verán la variación aleatoria.

Forma sencilla en R

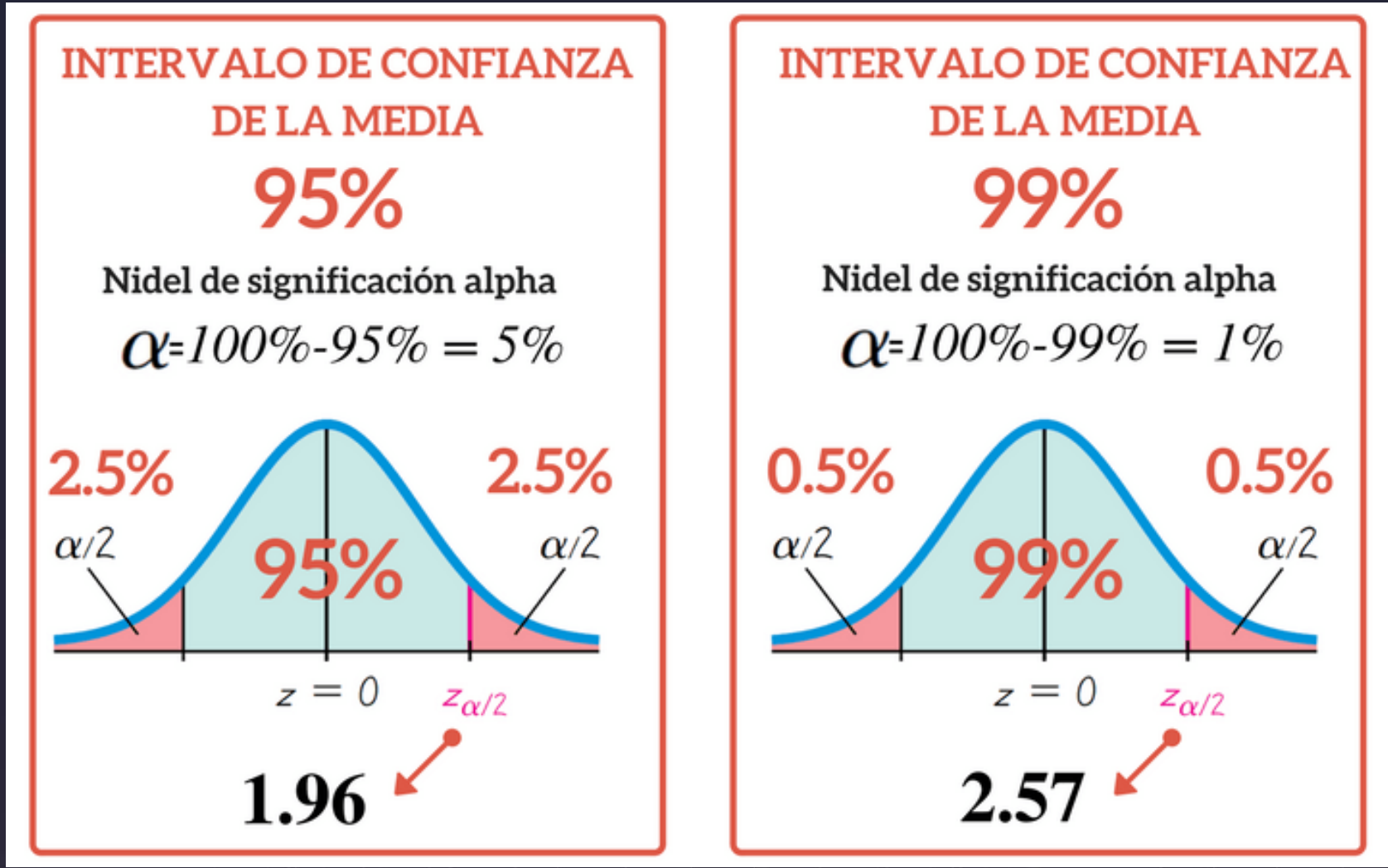
```
# Se puede realizar la prueba t para lograr esto
resultado <- t.test(x)

# Mostrar el intervalo de confianza
print(resultado$conf.int)
```

Nivel de confianza

- El nivel de confianza es:
  - Expresado en términos % (menor a 100%).
  - Un nivel que fijamos antes de construir el intervalo.
  - Es el % de intervalos, que construidos de la misma forma, contendrían el verdadero parámetro.
  - Ejemplo: Nivel de confianza = 95%.

    Con 95% de confianza, el intervalo considerado contendrá el verdadero valor de la media poblacional.

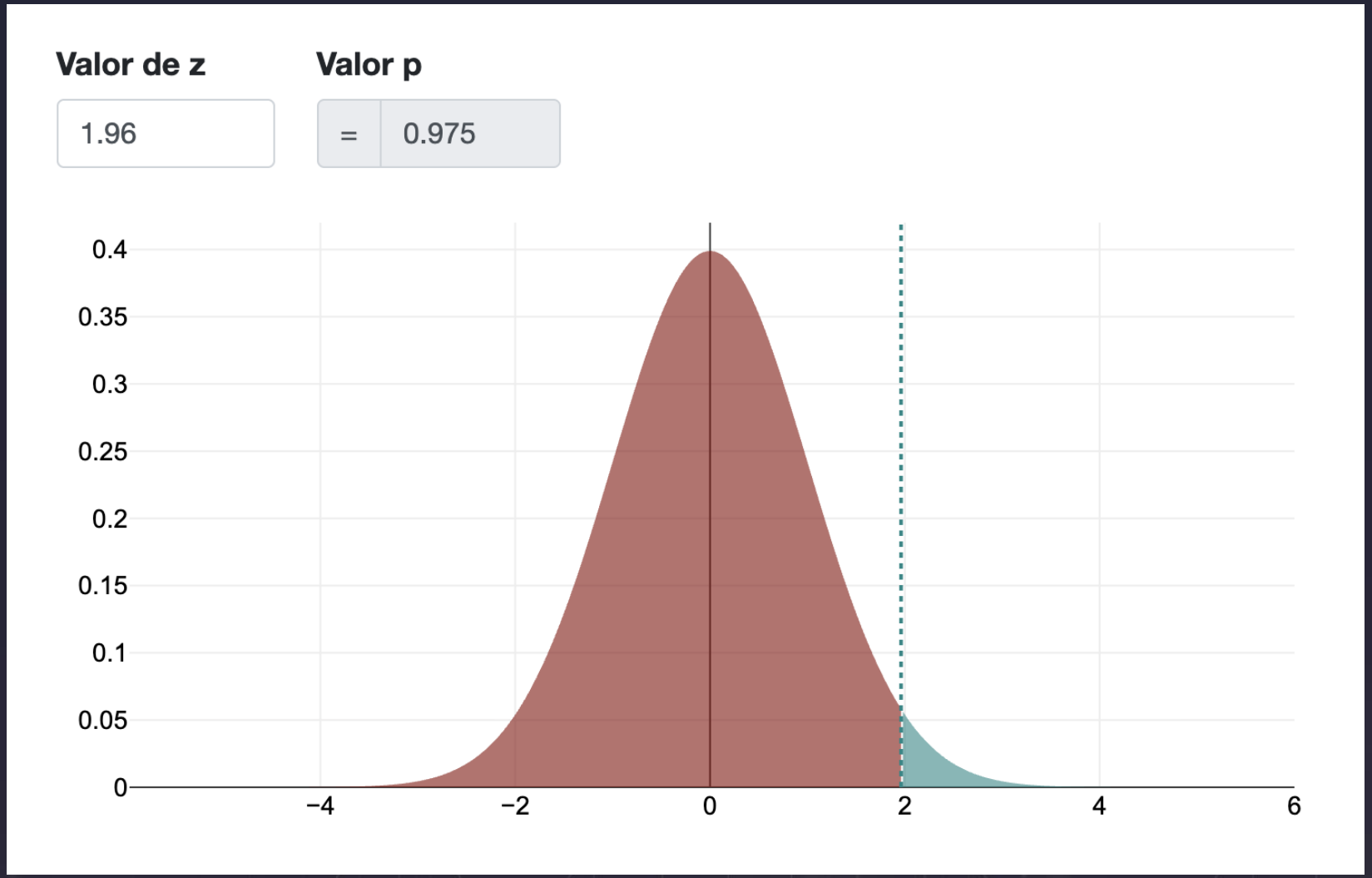
Nivel de confianza



Nivel de confianza = 95%

Desv. normal x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233

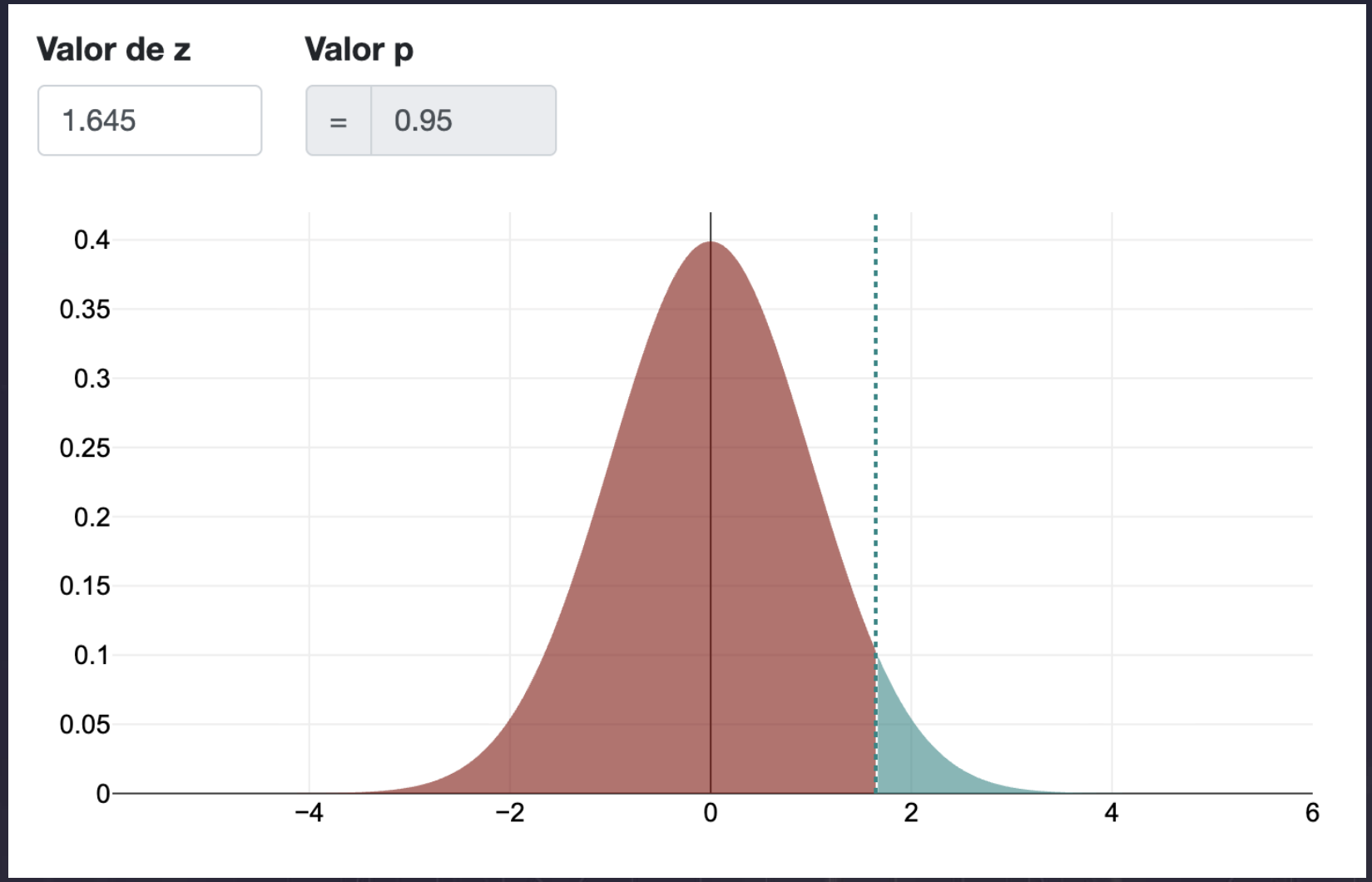
Nivel de confianza = 95%



Nivel de confianza = 90%

Desv. normal x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233

Nivel de confianza = 90%



Nivel de confianza

Continuara...