

# **Analítica de Datos**

**Probabilidad, Z-Score y Distribuciones  
Muestrales**

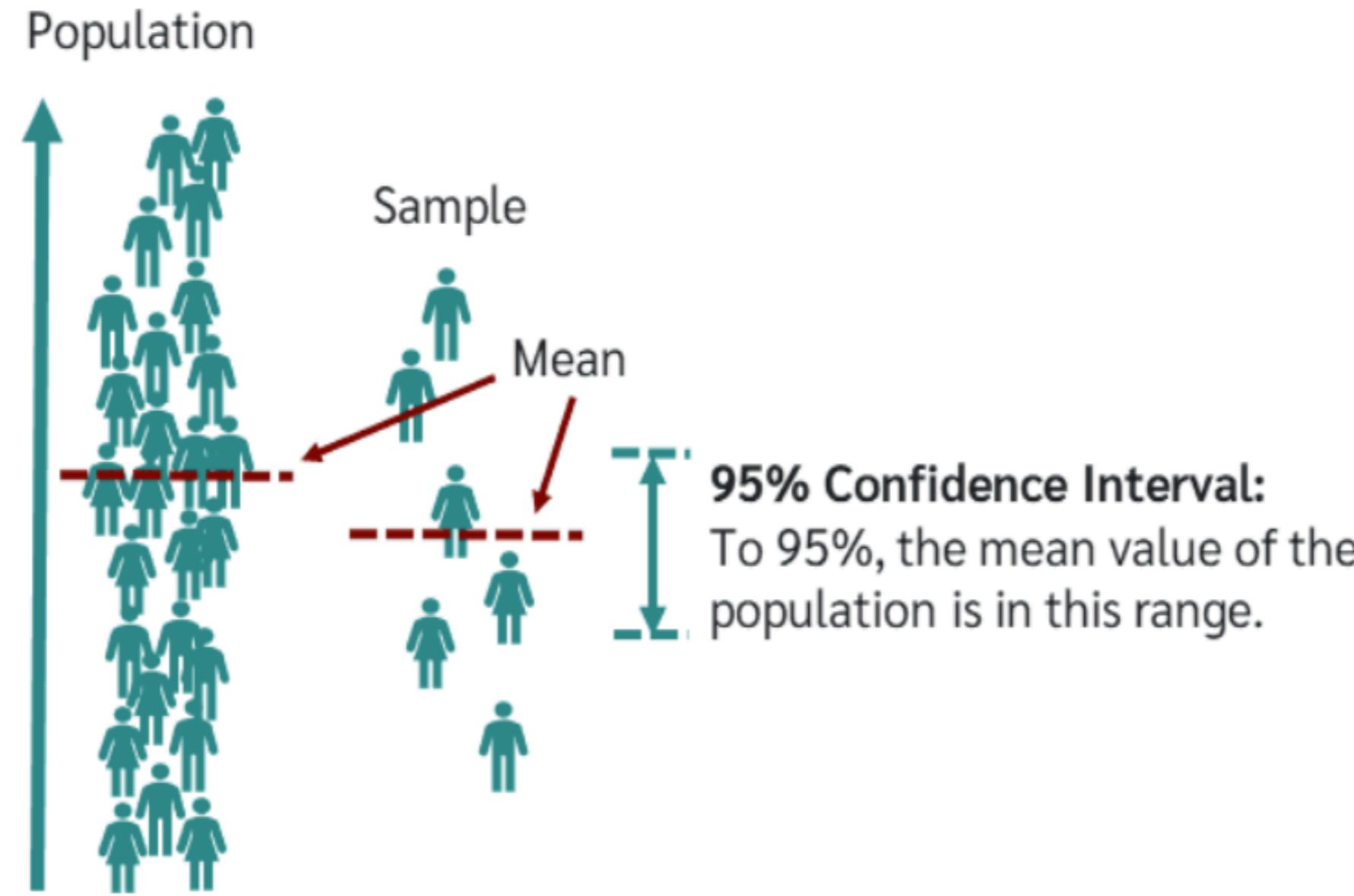
**Carlos Cardona Andrade**



# Tabla de contenido

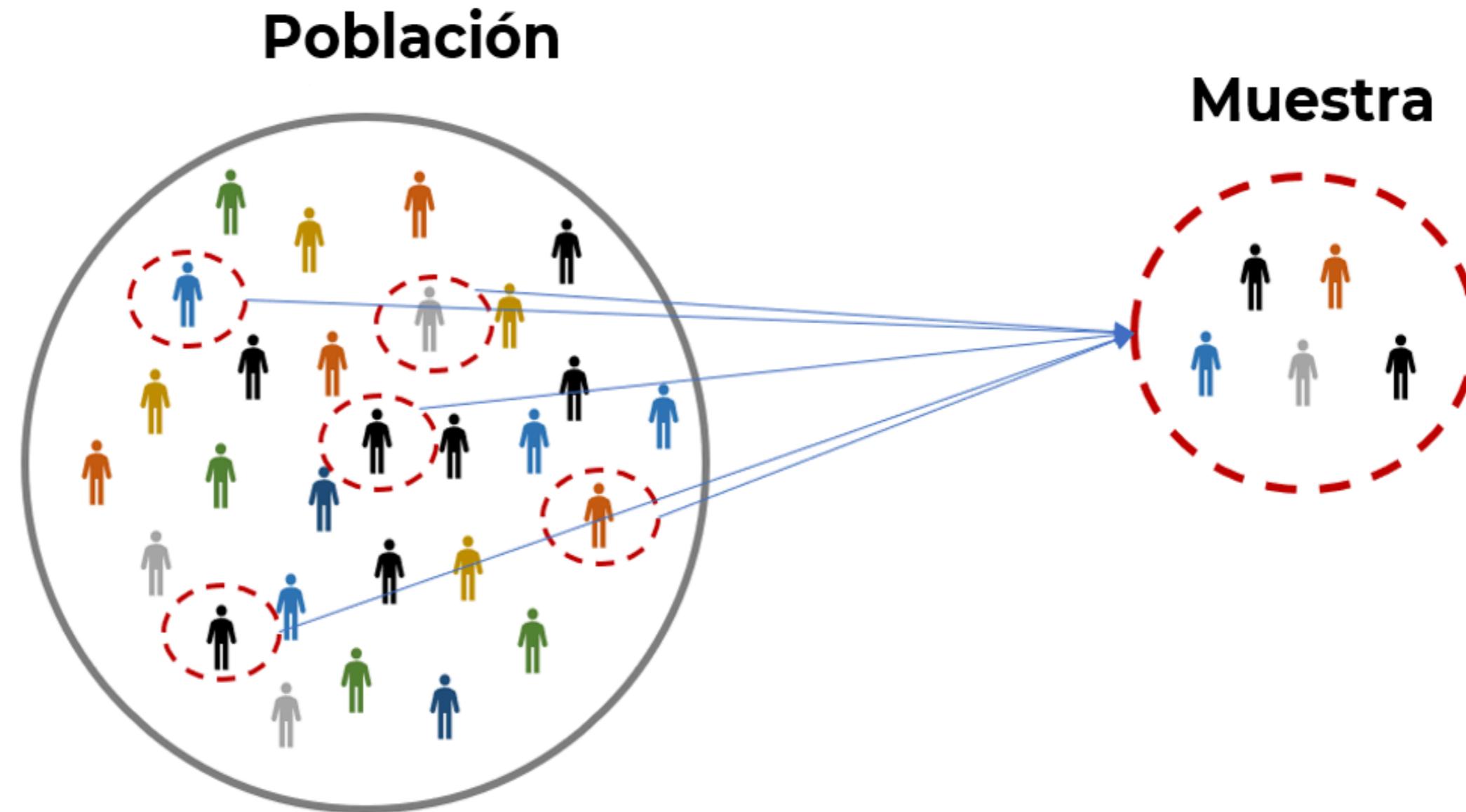
1. Probabilidad
2. Z-Score
3. Distribuciones Muestrales
4. Intervalos de Confianza

# La clase de hoy



# Breve Intro a: probabilidad

# Población y Muestra



# Población y Muestra

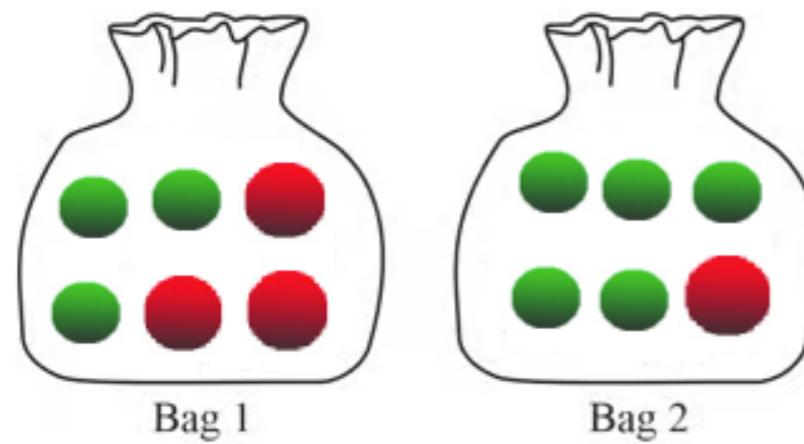
- Los estadísticos descriptivos discutidos anteriormente describen una **muestra**, pero no a la **población**.
- Las medidas que describen a una población se llaman **parámetros**. Utilizamos letras griegas para referirnos a ellos.

Medida	Parámetro poblacional	Estadístico muestral
Media	$\mu$	$\bar{X}$
Varianza	$\sigma^2$	$S^2$
Desviación Estándar	$\sigma$	$S$

# Probabilidad

- Una investigación inicia con una pregunta general sobre una población entera, pero se realiza usando una muestra.
- En esta situación, el rol de la estadística inferencial es utilizar la muestra como base para generalizar los resultados a la población.
- Para lograr este objetivo, los procedimiento inferenciales están construidos sobre el concepto de probabilidad.
- Específicamente, la relación entre población y muestra usualmente se define en términos de probabilidad.

# Probabilidad



- ¿Cuál es la probabilidad de sacar una bola verde?
- La probabilidad es una conexión entre población y muestra, la cual es base para la estadística inferencial que veremos más adelante.

# Probabilidad

- Una probabilidad se define como la siguiente proporción:

$$P = \frac{\# \text{ resultados deseados}}{\# \text{ resultados posibles}}$$

- Por ejemplo, al tirar un dado la probabilidad de obtener un 2 luego de lanzar un dado es:

$$P(2) = \frac{1}{6} = 0.166 = 16.66\%$$

# Probabilidad

1. Las probabilidades siempre están entre 0 y 1.
  - Una probabilidad igual a 0 indica que el evento nunca va a ocurrir.
  - Por otro lado, si es igual 1 indica que con toda seguridad el evento tendrá lugar.
2.  $\sum P = 1$
3. La probabilidad que un evento **no ocurra** es igual a 1 menos la probabilidad que el evento ocurra.
  - Al tirar un dado:

$$P(\sim 2) = 1 - P(2) = 1 - \frac{1}{6} = \frac{5}{6}$$

# Probabilidad

4. Si A y B son eventos alternativos (no se superponen), entonces  $P(A \text{ o } B) = P(A) + P(B)$
- Siguiendo con el ejemplo del dado:

$$P(2 \text{ o } 3) = P(2) + P(3) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6} = \frac{1}{3}$$

# Probabilidad

5. Si A y B son eventos que se superponen (ocurrencia conjunta), entonces  $P(A \text{ o } B) = P(A) + P(B) - P(A \text{ y } B)$
- ¿Cuál sería la probabilidad de sacar un número par o un 6?

$$P(\text{Par o } 6) = P(\text{Par}) + P(6) = \frac{3}{6} + \frac{1}{6} = \frac{4}{6} = \frac{2}{3} \quad \text{Incorrecto}$$

$$P(\text{Par o } 6) = P(\text{Par}) + P(6) - P(\text{Par y } 6) = \frac{3}{6} + \frac{1}{6} - \frac{1}{6} = \frac{3}{6} = \frac{1}{2} \quad \text{Correcto}$$

# Probabilidad

6. Si A y B son **independientes**, entonces  $P(A \text{ y } B) = P(A) * P(B)$

- ¿Cuál es la probabilidad de sacar 2 luego de tirar el dados dos veces?

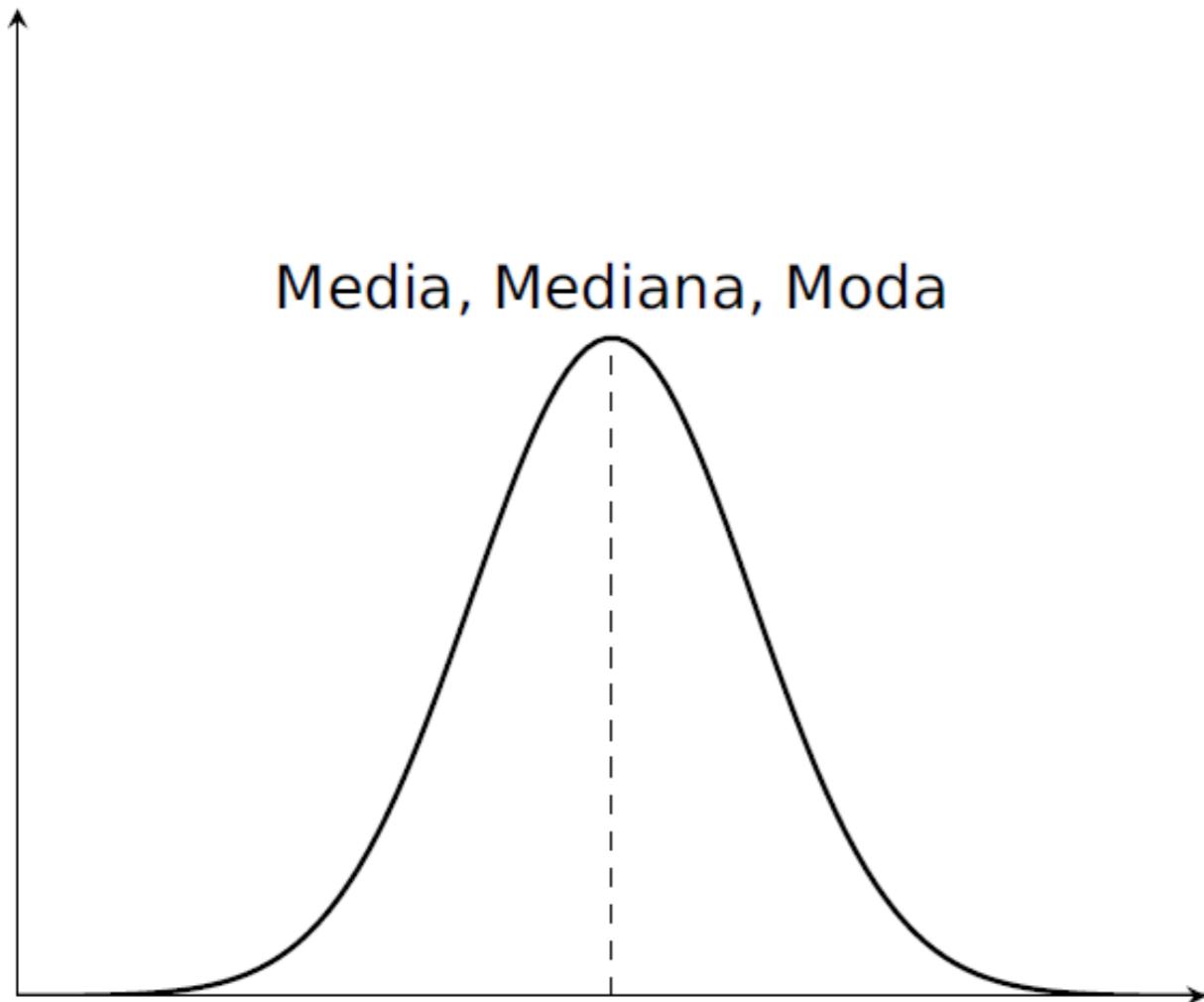
$$P(2 \text{ luego } 2) = P(2) * P(2) = \frac{1}{6} * \frac{1}{6} = \frac{1}{36}$$

- Es importante tener en cuenta si existe reemplazo o no.
- Por ejemplo, si un recipiente tiene 4 pelotas amarrillas y 2 azules. ¿Cuál es la probabilidad de sacar una amarilla y luego una azul sin reemplazo?

Breve intro a:

**Z-Score**

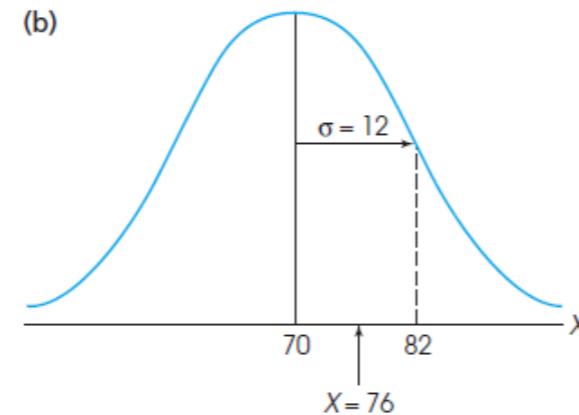
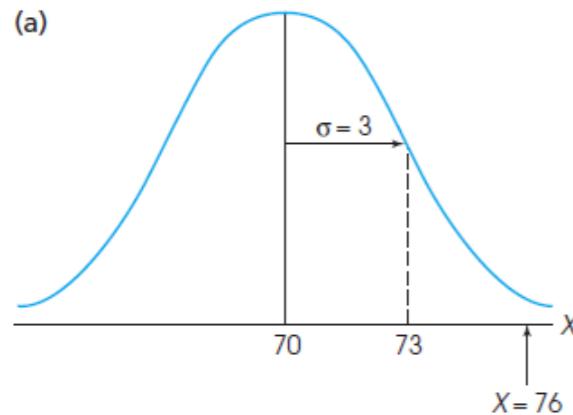
# Distribución normal



- La distribución normal es simétrica
- Cuando hablamos de distribución, piensen en esa curva suave.

# Z-Score

- Supongamos que en un examen cualquiera, la nota recibida es  $X = 76$ . ¿Es buena o mala la nota?
- Si la media es  $\mu = 70$ , sé que estoy 6 puntos por encima de la media.



- Aún teniendo información de la media, no es posible saber dónde está localizado el valor de la nota.

# Z-Score

- La ubicación relativa del valor dentro de una distribución depende tanto de la media como de la desviación estándar.
- Los valores Z tienen dos objetivos:
  1. Reportar la ubicación exacta de un valor dentro de una distribución.
  2. Permitir la comparación entre dos distribuciones estandarizadas.

# Z-Score

- ¿Se puede comparar un puntaje de 64 en la sección de matemáticas de Saber Pro con un puntaje de 66 directamente, si se sabe que cada prueba fue realizada en períodos diferentes?
- Si la media de ambos exámenes fue 62, además  $\sigma_1 = 1$  y  $\sigma_2 = 4$  ¿A quién le fue mejor?

# Z-Score

**Un valor Z reporta la ubicación precisa de cada valor X dentro de la distribución. Su signo señala si el valor está por encima o por debajo de la media. Además, el valor numérico especifica la distancia de la media al contar el número de desviaciones estándar entre el valor y la media.**

- La fórmula para calcular el valor Z es la siguiente:

$$Z_X = \frac{X - \bar{X}}{s_X}$$

- La unidad de medida del valor Z son el número de desviaciones estándar (SD).

# Z-Score

- Volvamos al ejemplo de la prueba Saber Pro.
- Para el puntaje de 64, donde  $\bar{X} = 62$  y  $\sigma_1 = 1$ :

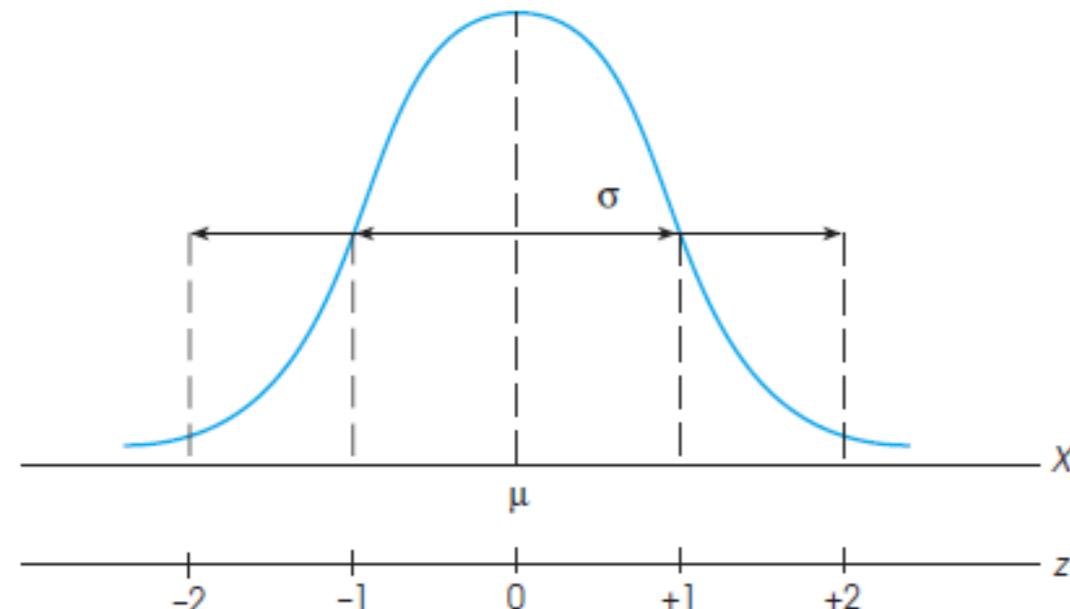
$$Z_X = \frac{64 - 62}{1} = \frac{2}{1} = +2SD$$

- Para el puntaje de 66, donde  $\bar{X} = 62$  y  $\sigma_1 = 4$ :

$$Z_X = \frac{66 - 62}{4} = \frac{4}{4} = +1SD$$

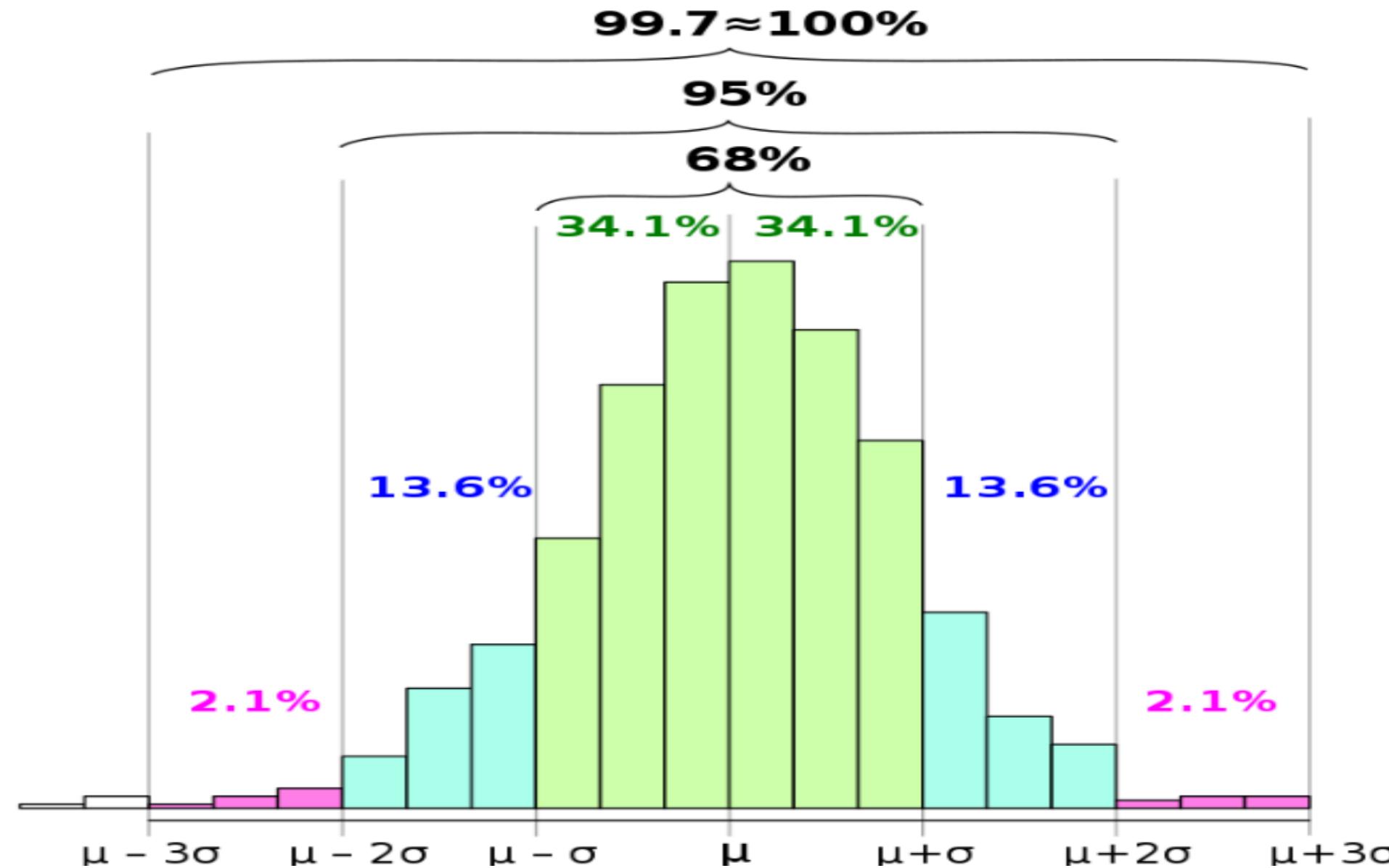
- La fórmula en muchas ocasiones no es necesaria. Si  $\bar{X} = 10$  y  $s_x = 2$ , cuál es el valor de Z para un  $X=8$ ?

# Z-Score



- Al estandarizar una distribución:
  1. su forma se mantiene.
  2. La media se convierte en  $\mu = 0$ .
  3. La desviación estándar será  $\sigma = 1$ .

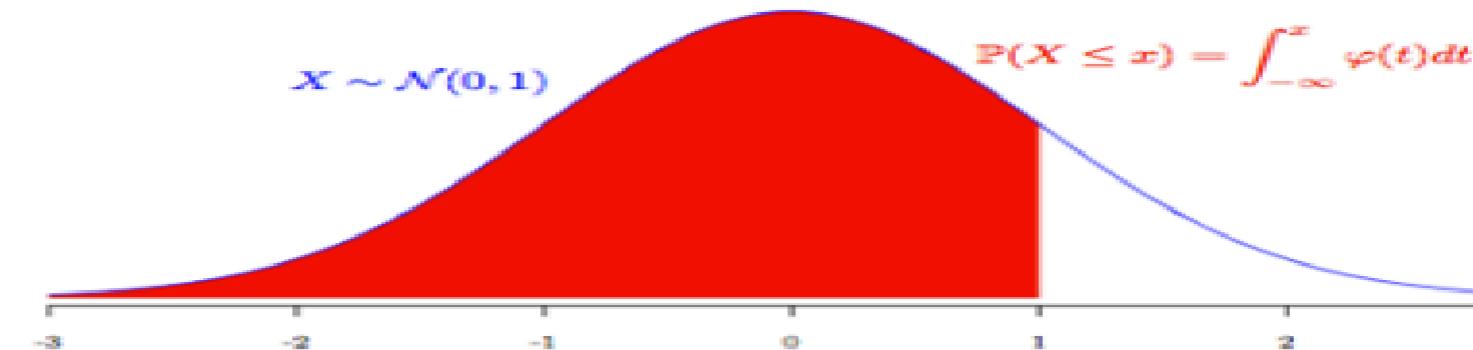
# La Regla 68-95-99



# La utilidad de la regla

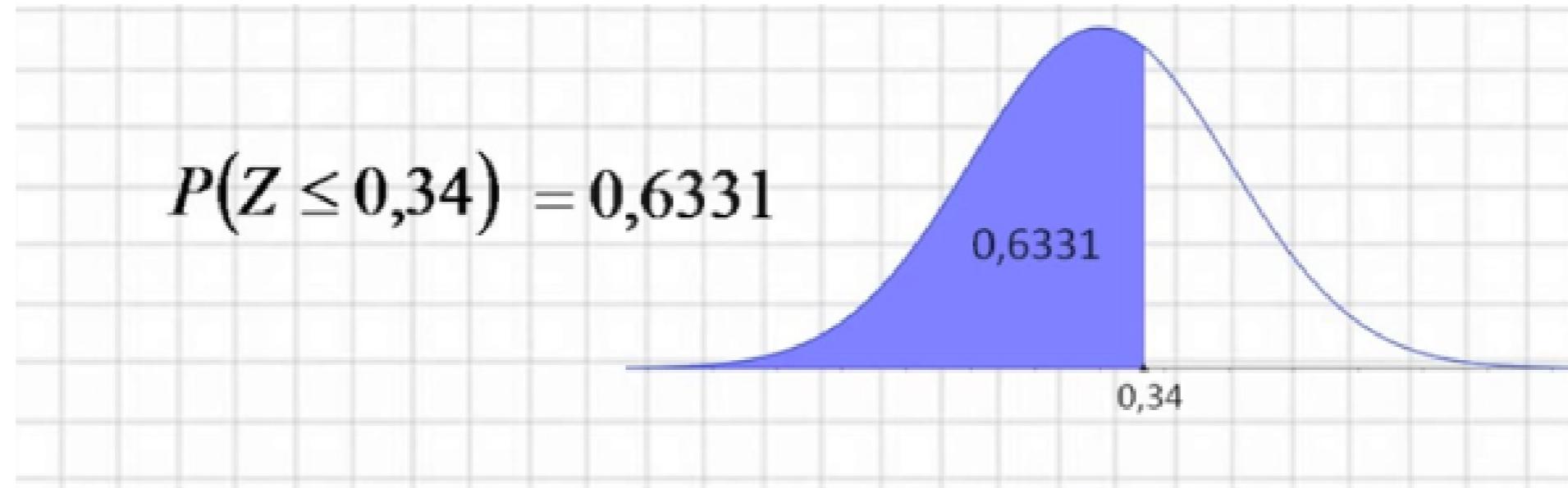
- Imaginemos una muestra de 1000 personas de la universidad.
- Su peso medio es de  $X = 60\text{kgs}$  y su desviación estándar es  $s_X = 5\text{kgs}$ .
- 500 personas pesan menos de 60 kgs.
- Cerca de 680 personas pesan entre 55 y 65 kgs.
- Alrededor de 950 personas pesan entre 50 y 70 kgs.
- Aproximadamente 3 pesan menos de 45 y más de 75 kgs.

# La distribución normal



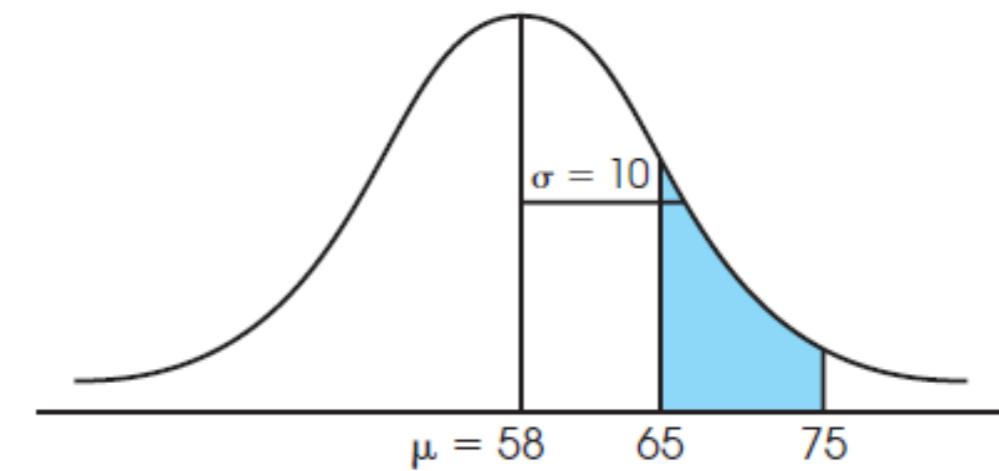
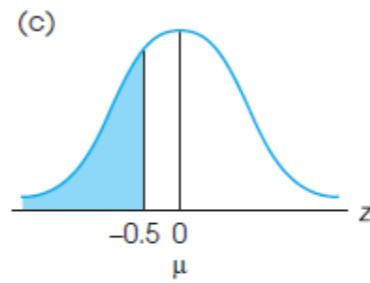
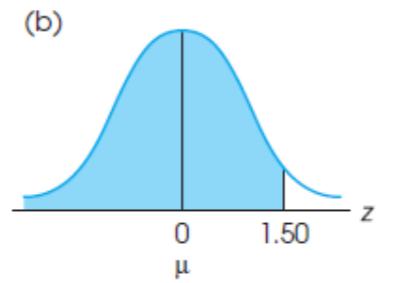
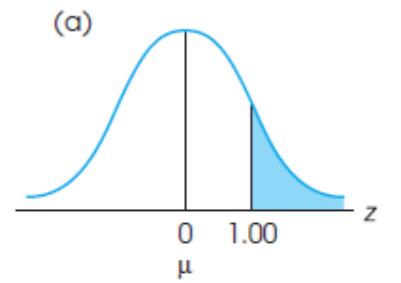
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

# La distribución normal



<b><i>z</i></b>	<b>,00</b>	<b>,01</b>	<b>,02</b>	<b>,03</b>	<b>,04</b>	<b>,05</b>	<b>,06</b>
<b>0,0</b>	0,5000	0,5040	0,5080	0,5120	<b>0,5160</b>	0,5199	0,5239
<b>0,1</b>	0,5398	0,5438	0,5478	0,5517	<b>0,5557</b>	0,5596	0,5636
<b>0,2</b>	0,5793	0,5832	0,5871	0,5910	<b>0,5949</b>	0,5987	0,6026
<b>0,3</b>	<b>0,6179</b>	<b>0,6217</b>	<b>0,6255</b>	<b>0,6293</b>	<b>0,6331</b>	0,6368	0,6406
<b>0,4</b>	0,6554	0,6591	0,6628	0,6664	<b>0,6700</b>	0,6736	0,6772

# La distribución normal



# Distribuciones Muestrales y Teorema del Límite Central

# Distribución Muestral

- Dos muestras separadas probablemente diferirán a pesar de ser tomadas de una misma población.
- Las muestras tienen diferentes individuos, diferentes valores, diferentes medias, etc.
- En muchos casos, es posible obtener infinitas muestras de una población.
- Por ejemplo, para Colombia existen más de 10000 muestras de 2 personas dados los 45 millones de habitantes.

# Distribución Muestral

- Aun cuando en muchas ocasiones es imposible obtener todas las muestras posibles de una población (e.g., más de 10000 muestras de 2 personas para Colombia) , existen ciertos patrones en el comportamiento de esas muestras.
- La habilidad de predecir características muestrales está basada en la distribución muestral de medias.

**La distribución muestral de medias es la colección de las medias muestrales para todas las posibles muestras aleatorias de un tamaño particular ( $n$ ) que pueden ser obtenidas de una población.**

# Distribución Muestral

- Como los estadísticos son obtenidos de muestras, la distribución de estadísticos es denominada como *distribución muestral*.

**Una distribución muestral es una distribución de estadísticos obtenidos al seleccionar todas las muestras posibles de un tamaño ( $n$ ) específico de una población.**

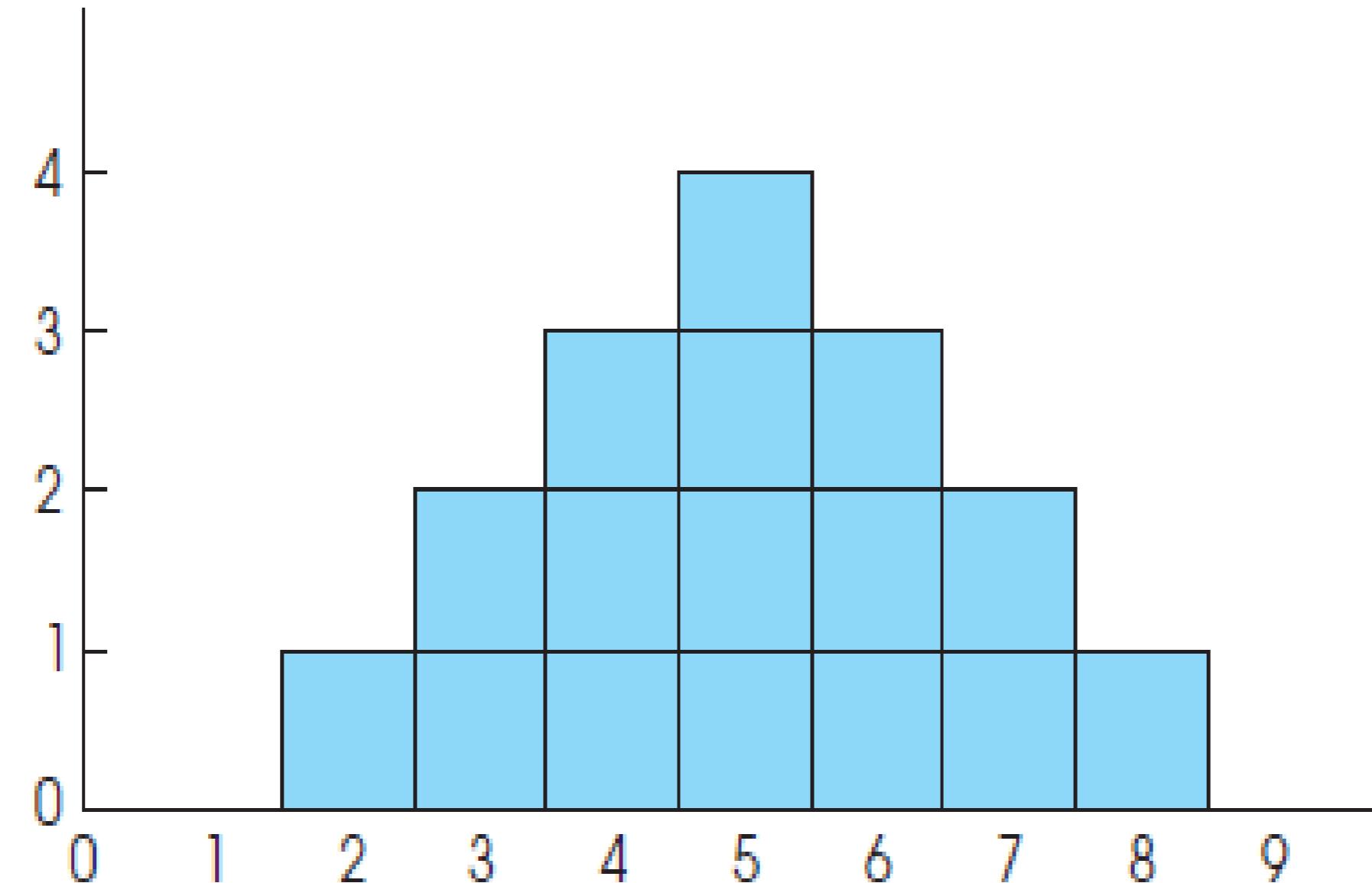
- De esta manera, la distribución muestral de medias es un ejemplo de una distribución muestral.

# Distribución Muestral

Muestra	Valores		Media Muestral ( $\bar{X}$ )
	Primer	Segundo	
1	2	2	2
2	2	4	3
3	2	6	4
4	2	8	5
5	4	2	3
6	4	4	4
7	4	6	5
8	4	8	6
9	6	2	4
10	6	4	5
11	6	6	6
12	6	8	7
13	8	2	5
14	8	4	6
15	8	6	7
16	8	8	8

- Consideren una población que consiste de sólo 4 valores: 2, 4, 6, 8.
- A partir de esta distribución, vamos a construir la distribución muestral de medias para  $n = 2$ .

# Distribución Muestral



# Distribución Muestral

- Noten que la distribución muestral de medias contienen *todas las muestras posibles*.
- Es necesario tener todos los posibles valores para calcular probabilidades.
- Por ejemplo, si el conjunto entero contiene 16 muestras, entonces la probabilidad de obtener cualquier muestra específica es 1 de 16:

$$p = \frac{1}{16}.$$

- Antes hablábamos de distribuciones de valores/puntajes; ahora los valores en la distribución no son puntajes sino estadísticos (medias muestrales).

# Distribución muestral

```
1 library(tidyverse)
2 ## Creamos el vector para el dado
3 dado <- c(1,2,3,4,5,6)
4
5 ## Tiremos el dado 5 veces
6 muestra_de_5 <- sample(dado, 5, replace=TRUE)
7
8 muestra_de_5
```

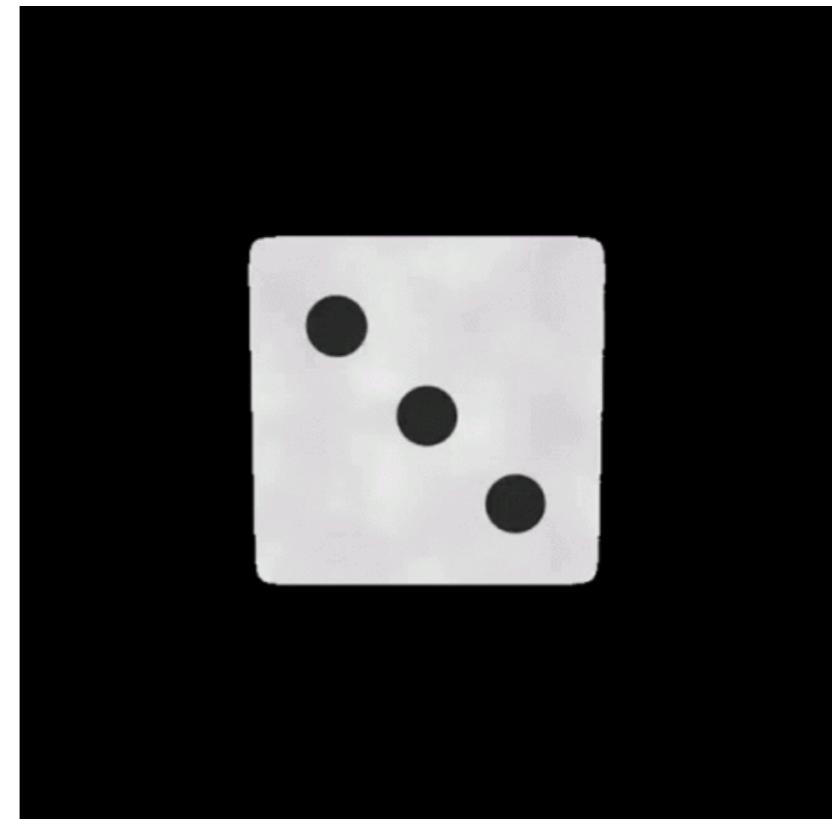
```
[1] 6 3 6 6 6
```

```
1 ## Y calculemos su media
2 mean(muestra_de_5)
```

```
[1] 5.4
```

```
1 ## Otra manera de calcular la media
2 ## de una muestra de 5
3 sample(dado, 5, replace = TRUE) %>% mean()
```

```
[1] 3.8
```



# Distribución Muestral

```
1 library(tidyverse)
2 ## Creamos el vector para el dado
3 dado <- c(1,2,3,4,5,6)
4
5 ## Tiremos el dado 5 veces
6 muestra_de_5 <- sample(dado, 5, replace=TRUE)
7
8 muestra_de_5
```

```
[1] 3 2 3 3 1
```

```
1 ## Y calculemos su media
2 mean(muestra_de_5)
```

```
[1] 2.4
```

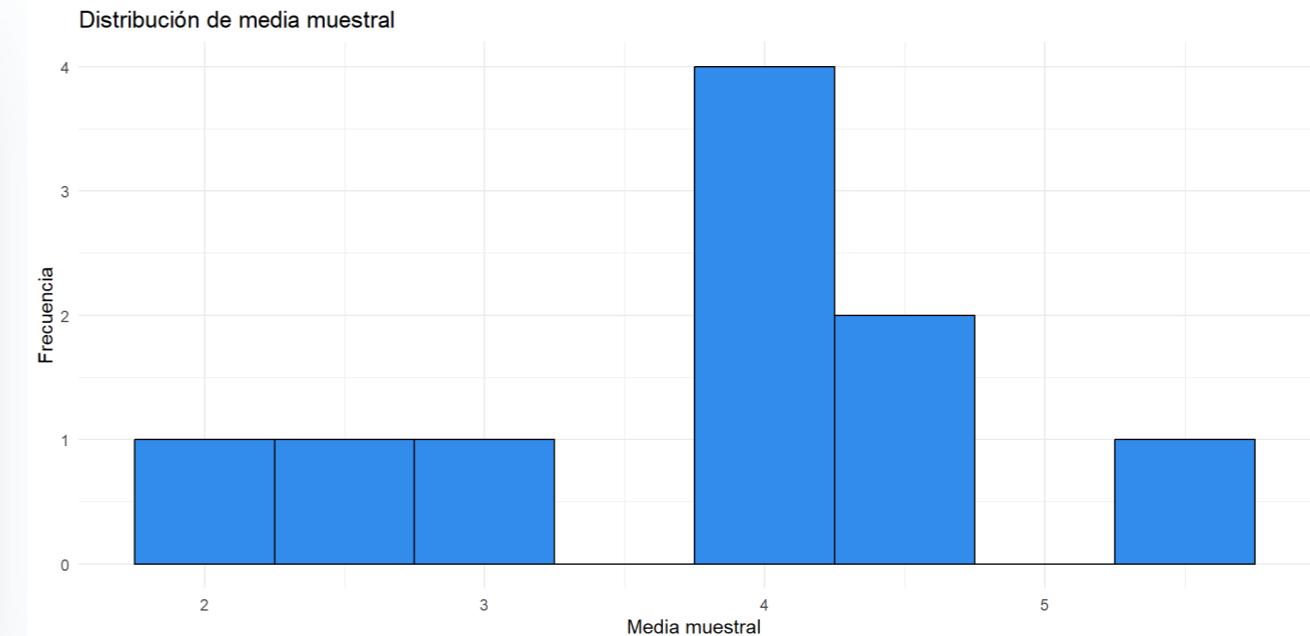
```
1 ## Otra manera de calcular la media de una muestra de 5
2 sample(dado, 5, replace = TRUE) %>% mean()
```

```
[1] 3.6
```

# Distribución Muestral

¿Cómo sería la distribución de medias si tomamos una muestra de 10?

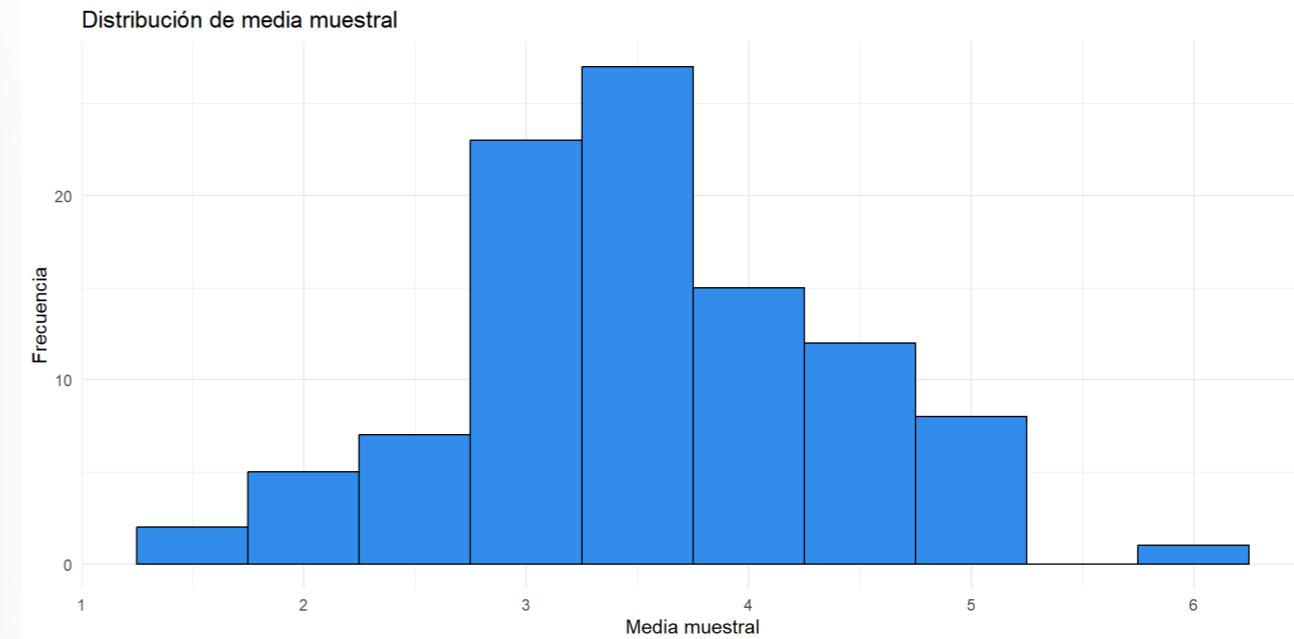
```
1 library(tidyverse)
2
3 sample_means <- replicate(10, sample(dado, 5, replace = TRUE))
4
5 # Convert to a data frame for ggplot2
6 df <- data.frame(sample_means = sample_means)
7
8 # Plot the histogram
9 ggplot(df, aes(x = sample_means)) +
10   geom_histogram(binwidth = 0.5, fill = "#348fe9")
11   labs(title = "Distribución de media muestral")
12   theme_minimal()
```



# Distribución Muestral

¿Cómo sería la distribución de medias si tomamos una muestra de 100?

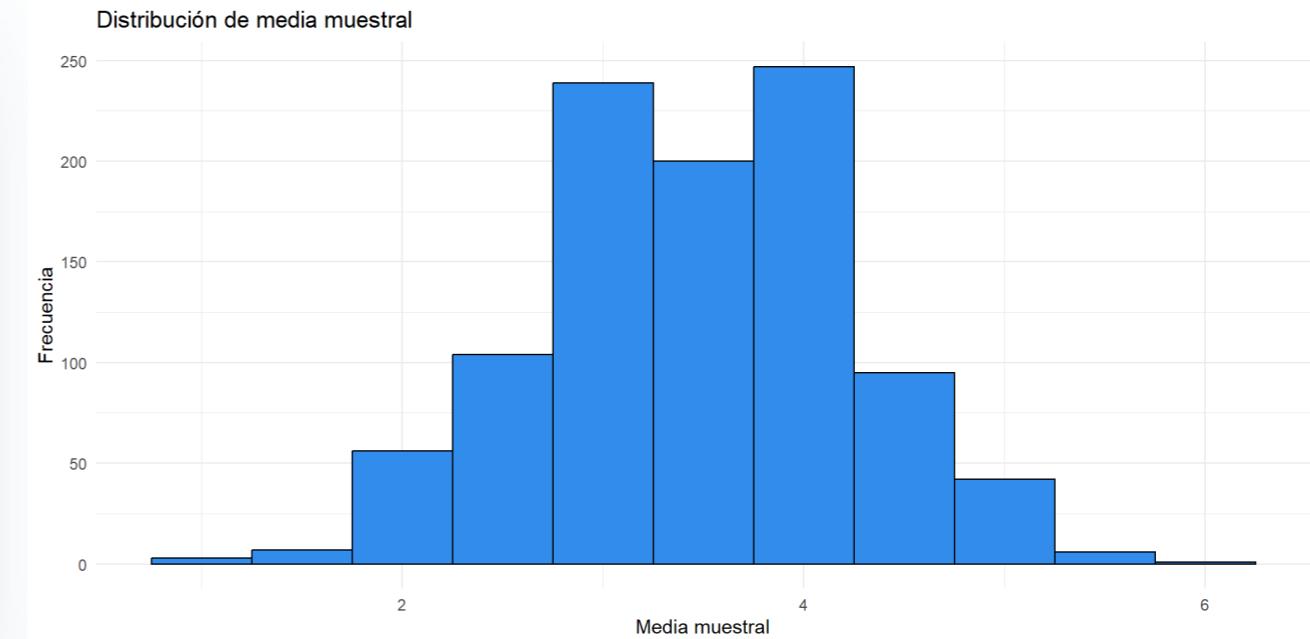
```
1 library(tidyverse)
2
3 sample_means <- replicate(100, sample(dado, 5,
4
5 # Convert to a data frame for ggplot2
6 df <- data.frame(sample_means = sample_means)
7
8 # Plot the histogram
9 ggplot(df, aes(x = sample_means)) +
10   geom_histogram(binwidth = 0.5, fill = "#348fe9")
11   labs(title = "Distribución de media muestral"
12     theme_minimal()
```



# Distribución Muestral

¿Cómo sería la distribución de medias si tomamos una muestra de 1000?

```
1 library(tidyverse)
2
3 sample_means <- replicate(1000, sample(dado, 5))
4
5 # Convert to a data frame for ggplot2
6 df <- data.frame(sample_means = sample_means)
7
8 # Plot the histogram
9 ggplot(df, aes(x = sample_means)) +
10   geom_histogram(binwidth = 0.5, fill = "#348fe9")
11   labs(title = "Distribución de media muestral")
12   theme_minimal()
```



# Distribución Muestral

- Dos características se destacan del histograma de la distribución muestral de medias:
  1. Las medias muestrales se mueven alrededor de la media.
  2. La distribución muestral de medias se aproxima a una curva normal.

# Distribución Muestral

# Distribución Muestral

- En situaciones más reales, con poblaciones y muestras mucho más grandes, el número de muestras posibles aumenta drásticamente.
- Por lo tanto, es imposible tener cada muestra posible.
- A pesar de esto, el teorema del límite central provee una descripción precisa de la distribución resultante si se seleccionan todas las muestras posibles.

# Teorema del Límite Central

Para cualquier población con media  $\mu$  y desviación estándar  $\sigma$ , la distribución muestral de medias para un tamaño de muestra  $n$  tendrá una media igual a  $\mu$  y una desviación estándar de  $\frac{\sigma}{\sqrt{n}}$ . Además, se aproximará a una normal a medida que  $n$  tiende a infinito.

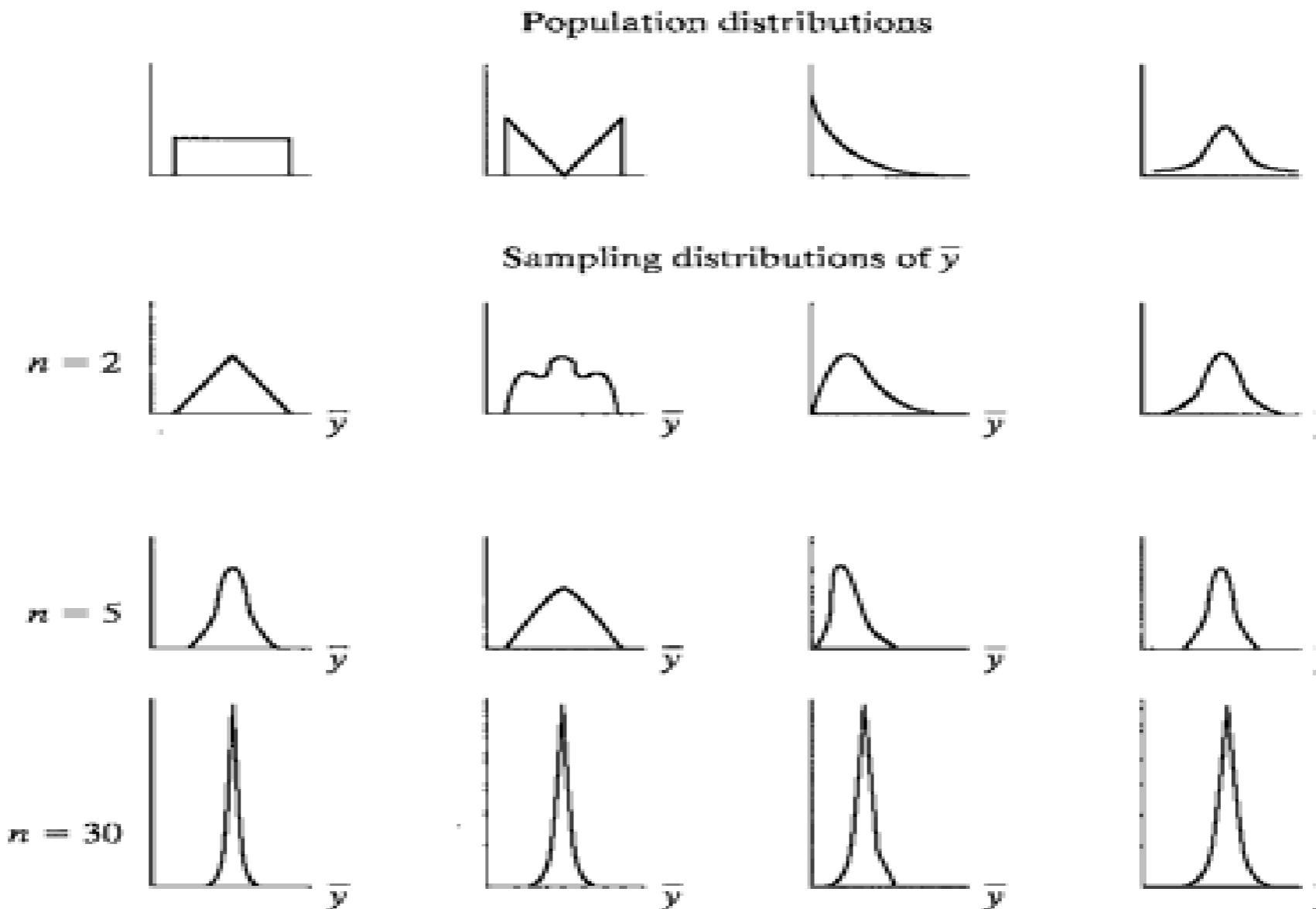
# Teorema del Límite Central

- El valor de este teorema recae en dos hechos:
  1. Describe la distribución muestral de medias para *cualquier población*, sin importar su forma, media o desviación estándar.
  2. La distribución muestral de medias se aproxima a una normal de manera rápida. Cuando la muestra alcanza un  $n = 30$ , la distribución es muy cercana a una normal.

# Teorema del Límite Central

- En resumen, el teorema del límite central identifica las tres características básicas de una distribución:
  1. Forma → Normal (Si la distribución poblacional es normal o si  $n > 30$ )
  2. Tendencia central →  $\mu$
  3. Dispersión →  $\frac{\sigma}{\sqrt{n}}$

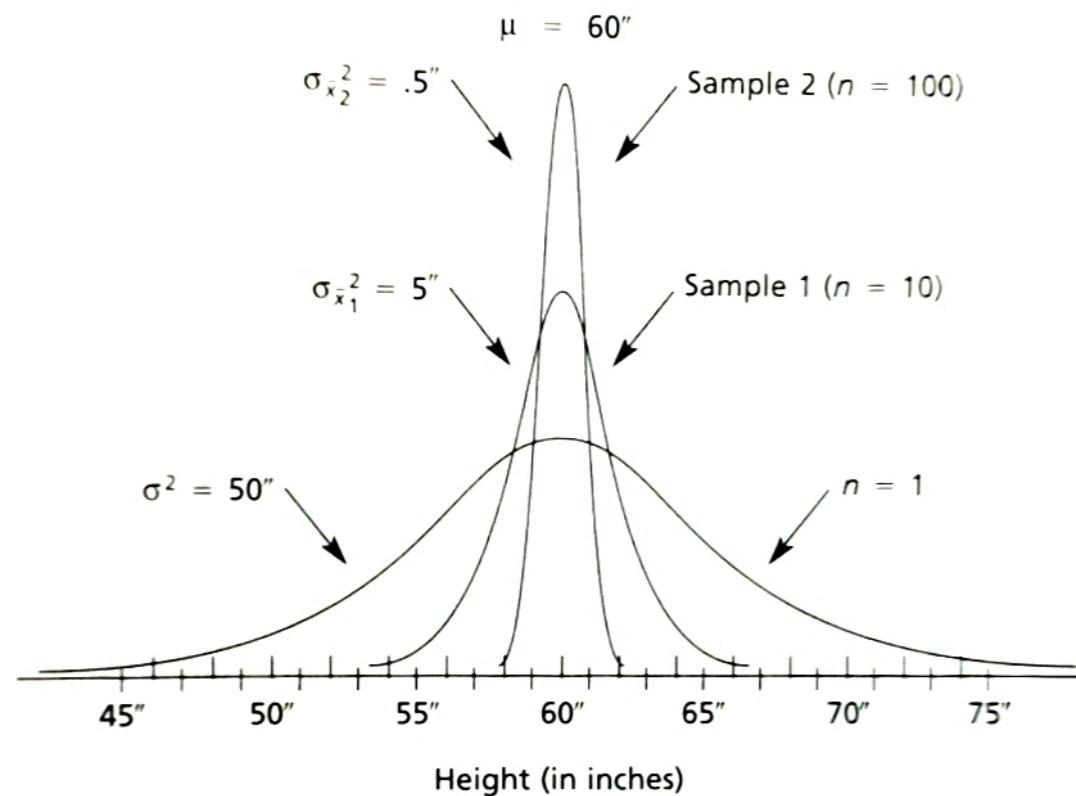
# Distribución Muestral



# Error Estándar

- La desviación estándar de una distribución muestral de medias se denomina *error estándar* y se identifica con el símbolo  $\sigma_{\bar{X}}$ .
- El error estándar cumple los dos propósitos de una desviación estándar:
  1. Describe la distribución al decir si las medias muestrales están agrupadas o se dispersan a lo largo de un intervalo amplio.
  2. Mide qué tan bien a una media muestral representa a toda la distribución de medias.

# Error Estándar



- Por tanto, un error estándar muy grande significa que existen grandes diferencias entre las muestras.
- Dado que la media de la distribución es  $\mu$ , el error estándar provee un estimado de la distancia entre una media muestral  $\bar{X}$  y la media poblacional  $\mu$ .

# Probabilidad y Distribución Muestral

- El principal uso de la distribución muestral de medias es encontrar la probabilidad asociada a cualquier muestra específica.
- Recuerden que la probabilidad es equivalente a una proporción!
- Dado a que la distribución muestral de medias presenta el conjunto de todas las posibles medias muestrales, podemos utilizar proporciones de esta distribución para determinar las probabilidades.
- Además, gracias al teorema del límite central podemos utilizar la tabla de la distribución normal.

# Probabilidad y Distribución Muestral

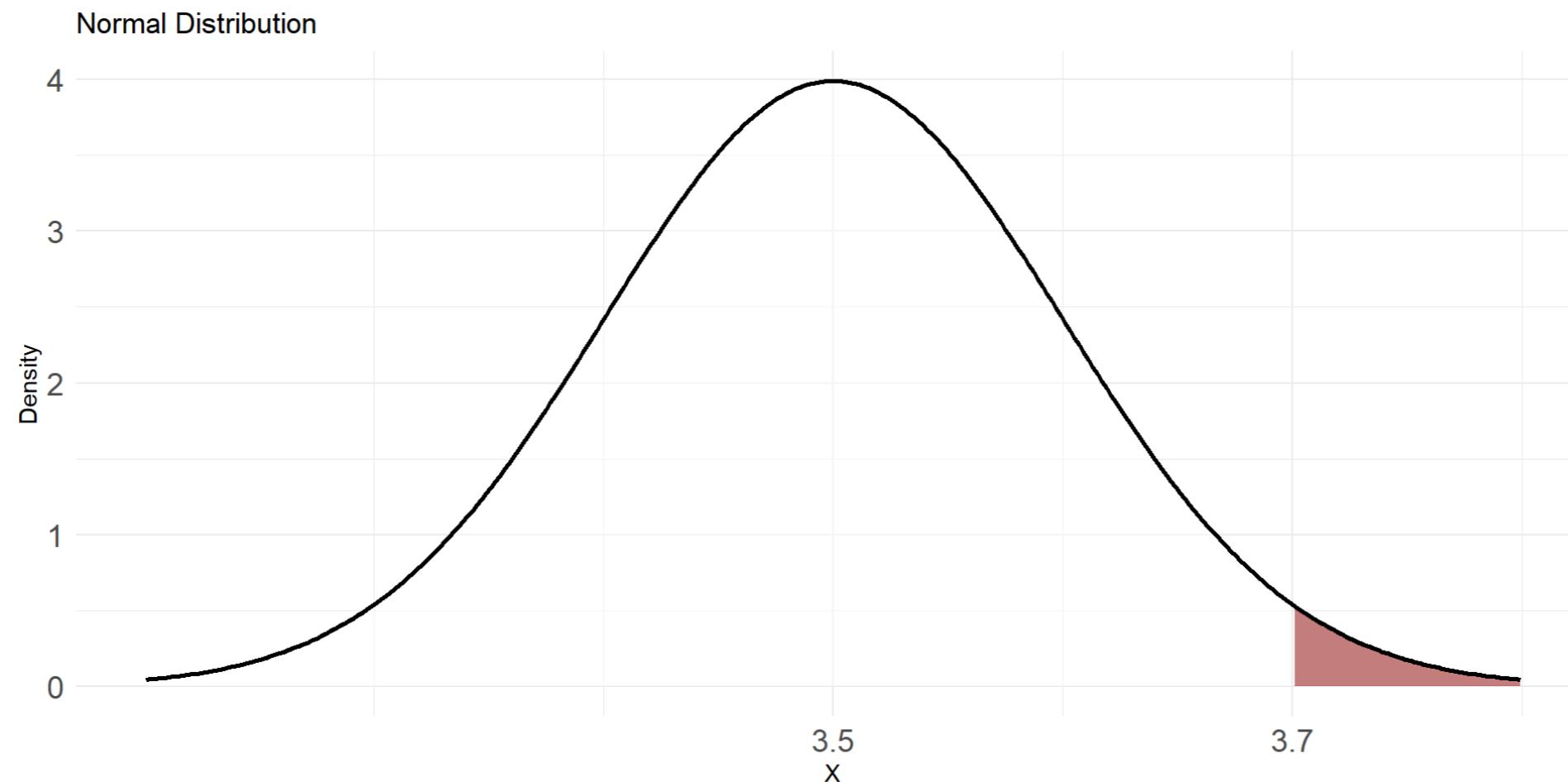
- Asumamos que se realiza un examen a todos los estudiantes de la universidad. La media de la distribución es  $\mu = 3.5$  y la desviación estándar es  $\sigma = 0.5$ . La distribución de la nota del examen es normal.
- Si se toma una muestra de  $n = 25$  estudiantes, ¿cuál es la probabilidad que la media muestral sea mayor a  $\bar{X} = 3.7$ ?

# Probabilidad y Distribución Muestral

- ¿Qué sabemos?
  1. La distribución muestral es normal porque la distribución del examen es normal.
  2. La distribución muestral tiene una media de 3.5 dado que la media poblacional es  $\mu = 3.5$
  3. La distribución muestral tiene una error estándar  $\sigma_{\bar{X}} = 0.1$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{0.5}{\sqrt{25}} = \frac{0.5}{5} = 0.1$$

# Probabilidad y Distribución Muestral



$$z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{3.7 - 3.5}{0.1} = \frac{0.2}{0.1} = 2$$

# Probabilidad y Distribución Muestral

- Por lo tanto, la probabilidad o proporción de encontrar una muestra con una media mayor a 3.7 es  $p(z > 2) = 0.0228=2.8\%$ .
- Considerando la misma distribución del ejemplo anterior, ahora encontremos el rango de valores que son esperados para la media muestral el 80% de las veces.
- Ya sabemos que la distribución es normal con una media esperada de  $\mu = 3.5$  y desviación estándar  $\sigma_{\bar{X}} = 0.1$ .

# Probabilidad y Distribución Muestral

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015

- En la tabla encontramos que para una proporción o probabilidad de 0.9, el z-score es 1.28
- De esta manera, los límites del 80% medio de la distribución corresponden a  $z = -1.28$  y  $z = +1.28$

# Probabilidad y Distribución Muestral

- Con un error estándar de 0.1, la distancia a la media es de  $1.28 * 0.1 = 0.128$
- La media es  $\mu = 3.5$ , por lo cual, una distancia de 0.128 en ambas direcciones produce un rango de valores entre 3.372 y 3.628
- De esta manera, 80% de todas las posibles medias muestrales se encuentran dentro de un intervalo entre 3.372 y 3.628
- Otra interpretación es que si seleccionamos una muestra  $n = 25$ , estamos 80% seguros que la media de la muestra va a encontrarse en ese intervalo.

# **Intervalos de Confianza**

# Intervalos de Confianza

- Es claro que sólo en contadas ocasiones se tendrán los valores para la media y la desviación estándar poblacional.
- Al trabajar con una de las posibles muestras, es necesario acercarnos a los parámetros poblacionales a partir de los estadísticos muestrales que se tienen disponibles.
- Un **intervalo de confianza** es un rango de valores posibles de un parámetro expresado en un grado o nivel específico de confianza.
- Con los intervalos de confianza tomamos una estimación puntual de la muestra y la acoplamos con el conocimiento que tenemos sobre las distribuciones muestrales.

# Intervalos de Confianza

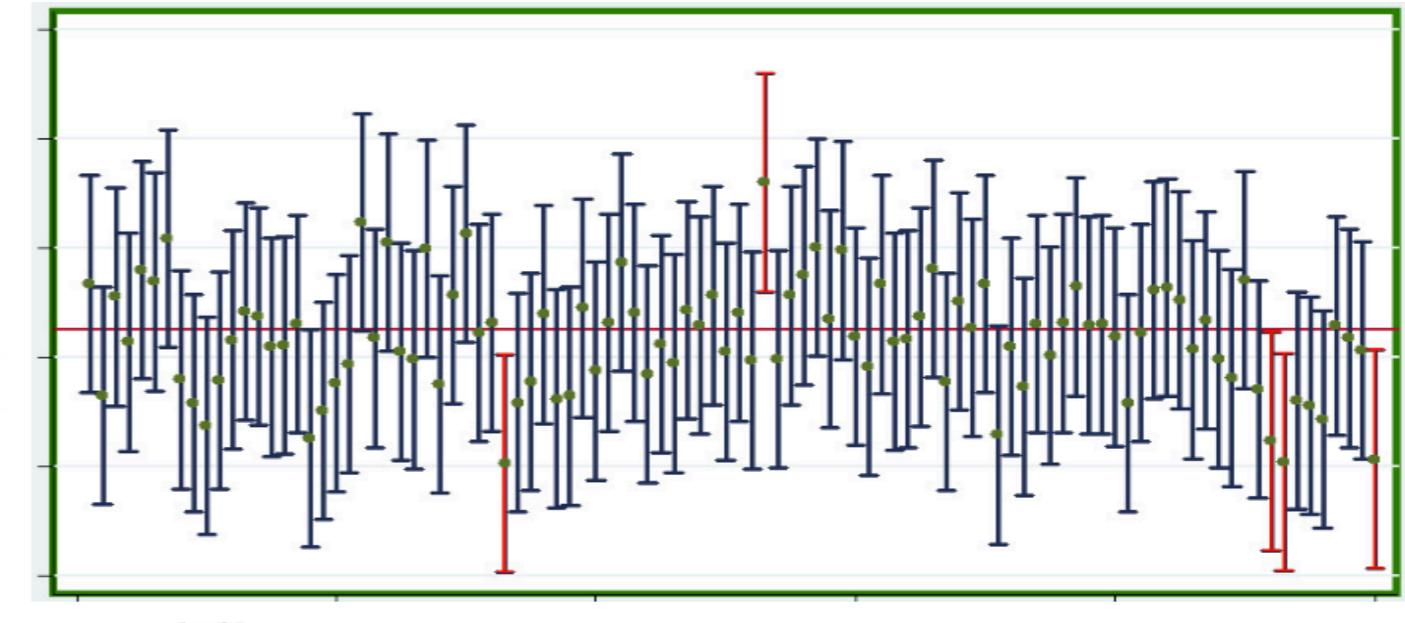
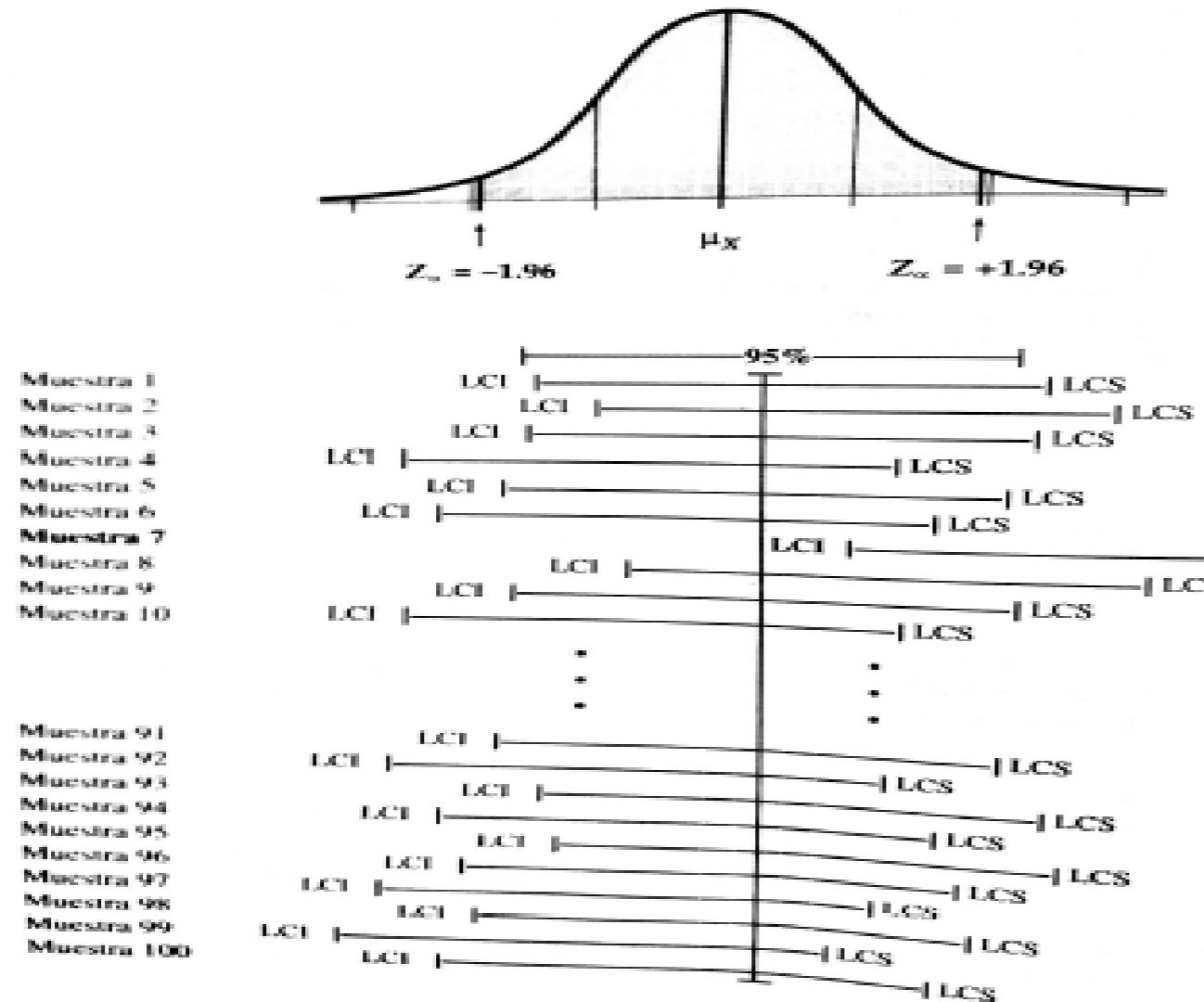
$$IC \text{ de } \mu = \bar{X} \pm Z_{\left(\frac{\alpha}{2}\right)} \times \sigma_{\bar{X}}$$

- Para calcular el intervalo de confianza necesitamos:
  1. la media muestral
  2. el  $Z$  al nivel de significancia  $\alpha$
  3. el error estándar

# Nivel de Confianza

- El **nivel de confianza** nos dice nuestra tasa de éxito, es decir, con qué frecuencia el parámetro poblacional se encuentra en el rango del intervalo de confianza.
- Usualmente, los grados de confianza más utilizados son el 95 y 99%.
- Al confiar en una muestra, sabemos que podemos fallar en la predicción debido a la existencia del error de muestreo.

# ¿Cuál es la interpretación?



# Nivel de Significancia

- La única manera de tener total certeza sobre nuestras conclusiones es reunir datos de la población.
- Cabe destacar que la cantidad de error es conocida. El nivel de error esperado es la diferencia entre el nivel de confianza y la **confianza perfecta** del 100%.
- En otras palabras, si estamos 95% seguros acerca de nuestro resultado, estamos 5% inseguros acerca de este.

$$Nivel\ de\ confianza = 95\%$$

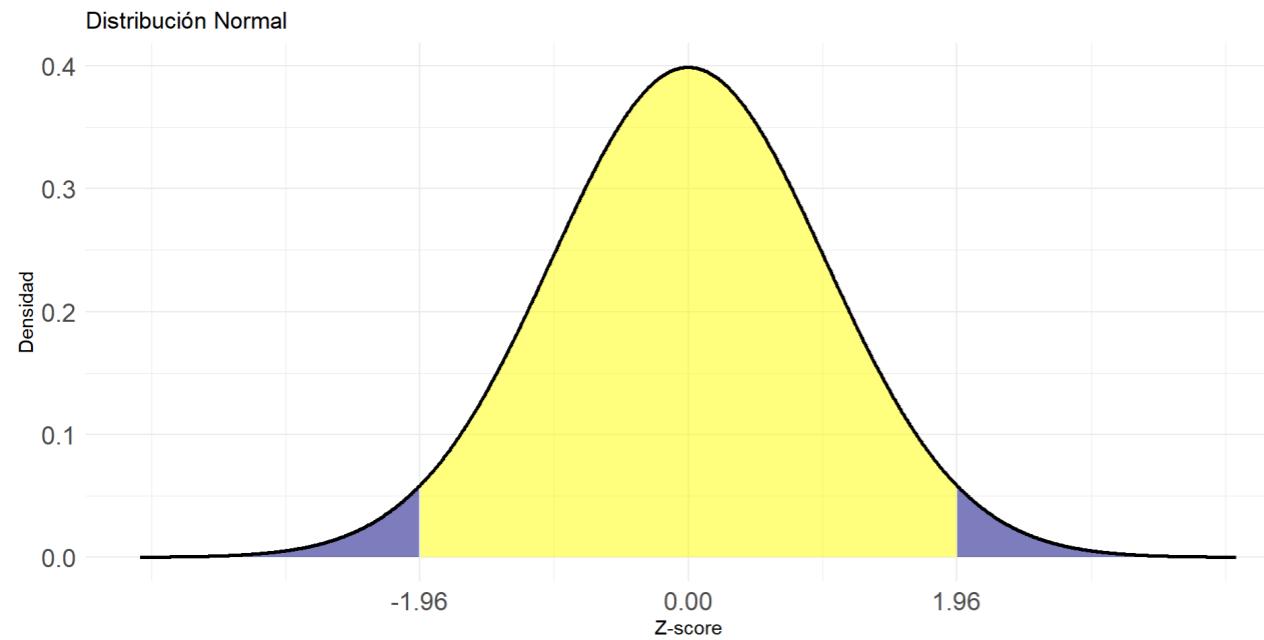
$$Nivel\ de\ significancia = \alpha = 100\% - 95\% = 5\%$$

# Nivel de Significancia

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767

- ¿A qué distancia del parámetro poblacional se encuentra nuestra media muestral para un nivel de confianza del 95%?
- A partir dela tabla de la normal podemos encontrar los valores Z críticos para el nivel de significancia  $\alpha$  ( $Z_{\left(\frac{\alpha}{2}\right)}$ ).

# Nivel de Significancia



- El área amarilla representa el 95% de las medias muestrales, mientras que, las dos áreas moradas representan el 5% de las medias que están fuera de los dos valores críticos.
- 95% de las observaciones de una normal están dentro de 1.96 SD.

# Intervalos de Confianza

- Supongamos que tomamos una muestra de  $n = 300$  estudiantes de la universidad con un promedio de edad de  $\bar{X} = 20.5$  años y  $s_X = 1.5$ .
- Sabemos que la distribución muestral de medias toma forma de curva normal cuando  $n > 30$ .
- La edad promedio de la muestra debe estar cerca del parámetro poblacional real (la edad media de todos los estudiantes de la universidad).
- Dado lo anterior, podemos decir con seguridad que el 95% de las muestras caen dentro de casi 2 errores estándar del parámetro real (Regla 68-95-99).

# Margen/Término de Error

- Ya sabemos que el  $Z_{\left(\frac{\alpha}{2}\right)} = 1.96$  y que  $s_x = 1.5$ . Por lo tanto:

$$\sigma_{\bar{X}} = \frac{1.5}{\sqrt{300}} = \frac{1.5}{17.3} = 0.08$$

- El margen de error será igual a  $Z_{\left(\frac{\alpha}{2}\right)} \times \sigma_{\bar{X}} = 1.96 \times 0.08 = 0.169$

- El intervalo de confianza estará entre:
  - Límite Inferior =  $20.5 - 0.169 = 20.3$
  - Límite Superior =  $20.5 + 0.169 = 20.6$

# ¿Cuál es la interpretación?

- Estoy 95% seguro de que la edad promedio de los estudiantes de la universidad se ubica entre 20.3 y 20.6 años.
- En otras palabras, si se realizan los mismos procedimientos muestrales 100 veces, el parámetro poblacional  $\mu$  estará entre los intervalos calculados el 95 de esas veces.

# Grado de Precisión

- Entre mayor sea el nivel de confianza estipulado, mayor será el margen de error y por lo tanto será menos preciso el intervalo de confianza.

$$Z_{0.05} = 1.96 \quad vs \quad Z_{0.01} = 2.58$$

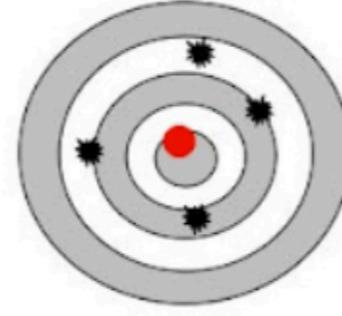
- Entre mayor sea el tamaño de la muestra, más preciso será el intervalo de confianza.

$$\sigma_{\bar{X}} = \frac{s_X}{\sqrt{n}} \text{ entonces si } \uparrow n \rightarrow \downarrow \sigma_{\bar{X}}$$

# Precisión y Exactitud



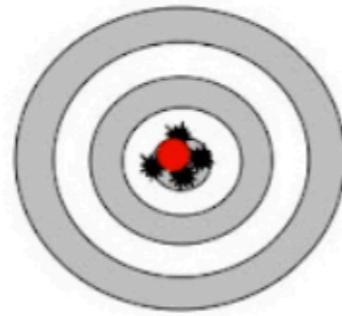
Ni preciso  
ni exacto



Impreciso  
pero exacto



Preciso  
pero inexacto



Preciso y exacto

- A medida que los intervalos de confianza se hacen más estrechos, se vuelven más precisos y ofrecen menos variabilidad.
- A medida que los intervalos de confianza se hacen más amplios, se vuelven más exactos.

