Clase 2

Fundamentos de la estadística para las CSC

Aproximación a las políticas públicas desde los datos | UC | 09 de junio, 2023

José D. Conejeros | jdconejeros@uc.cl

Guía

1. El uso de la estadística en Políticas Públicas
2. Estadística Descriptiva
3. Estadística Inferencial

1. El uso de la estadística en Políticas Públicas

Modelos de predicción matemáticos contradicen encuestas y dan por ganadora a la opción Apruebo

Por : Gustavo Espinoza
Periodista de Datos de El Mostrador

VER MÁS +

En climas políticos de polarización, hay estudios que han ido ganando terreno y prestigio en la predicción política: los análisis de modelamiento matemático, que han tenido muy buenos resultados en las elecciones presidenciales de EE.UU. y Canadá, como también en referendos en Europa. Se trata de estudios inicialmente utilizados para la predicción financiera, que ahora se usan también para la política. Información extraída de Google y otros espacios abiertos, como las redes sociales, es sometida a inteligencia artificial para procesarla y llegar a conclusiones. Dos organizaciones dedicadas a este quehacer dan por ganadora a la opción Apruebo este 4 de septiembre en Chile, con un 55,1% (en el caso de Espacio Político) y un 56% (en el de Daoura), contradiciendo así los pronósticos de las encuestas de opinión pública. "Nosotros somos científicos, estamos constantemente expuestos al fracaso, no estamos prisioneros de la verdad, pero lo que pasa es que hemos sido muy precisos antes", sostiene Francisco Vergara, experto en Big Data.

Estadística y Políticas Públicas

Reporte 20 agosto 2022 Plebiscito Salida

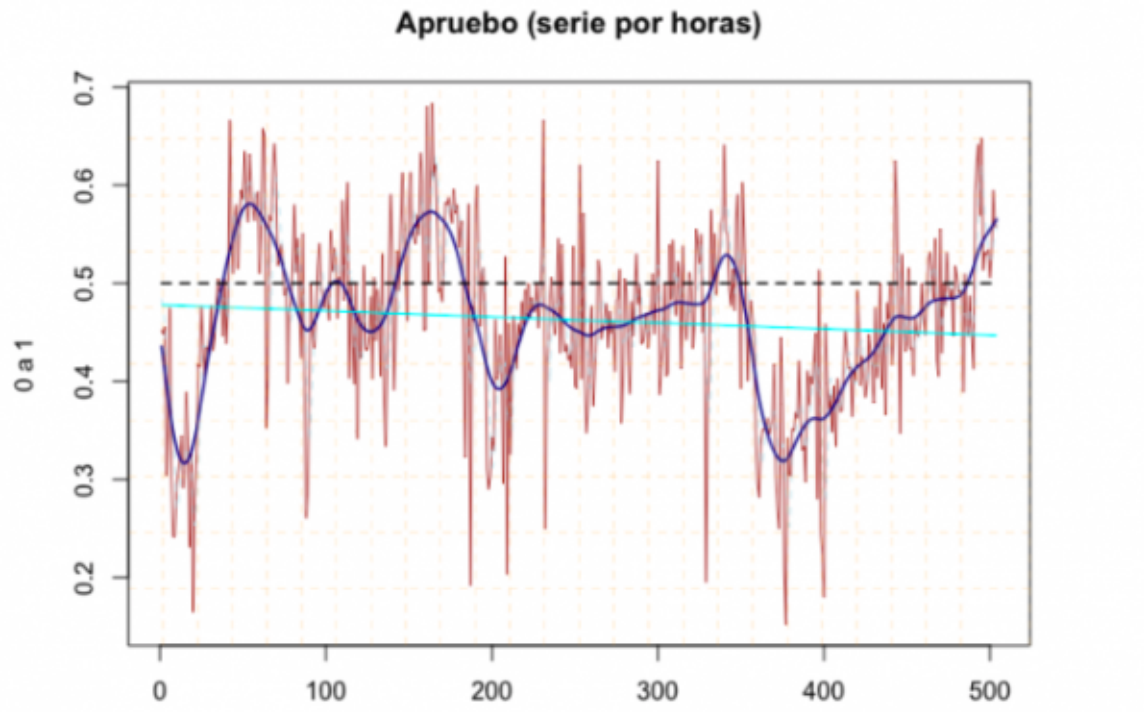


Fig 1. La figura indica la variación del peso del apruebo sobre rechazo entre el 30 de julio 2022 y el 20 de agosto 2022. En rojo las interacciones, en celeste la línea de tendencia por regresión lineal.

2. Estadística Descriptiva

Una definición

La estadística descriptiva nos entrega información sobre la **distribución** de los datos en una muestra. En otras palabras, es un **conjunto de procedimientos matemáticos** realizados para resumir y/o analizar información obtenida de manera sistemática.

Por lo tanto, podemos representar de forma apropiada cómo se distribuye una **variable** contabilizando tanto la frecuencia de ocurrencia de cada puntuación en una muestra, como la distribución y dispersión de las puntuaciones.

Para variables numéricas

Estadísticos de tendencia central

- Media:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- Mediana:

$$Mdn = \frac{n+1}{2}$$

- Moda: Valor con el mayor número de repeticiones.

Para variables numéricas

Dispersión

- Rango:

$$Rg = (X_{max} - X_{min}) + UR$$

- Varianza:

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

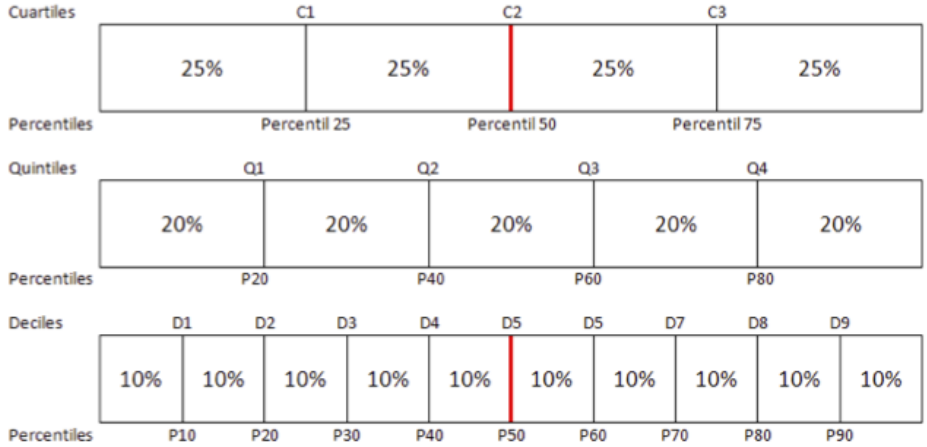
- Desviación estándar:

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Para variables numéricas

Posición

- Cuartiles: Fracionan la distribución en 4 partes con igual cantidad de casos. Corresponden a las puntuaciones (o valores) de una variable bajo los cuales queda ubicado un porcentaje determinado del total de las puntuaciones. En este caso, son las puntuaciones Q_1, Q_2 y Q_3 que cuentan con un 25% de los casos cada uno.



Para variables categóricas/discretas

Distribución

- Frecuencias: Nº de veces que es observada cada puntuación (valor o atributo) de una variable en la muestra.
- Proporciones:

$$p = \frac{freq\ categoria}{n}$$

- Porcentajes:

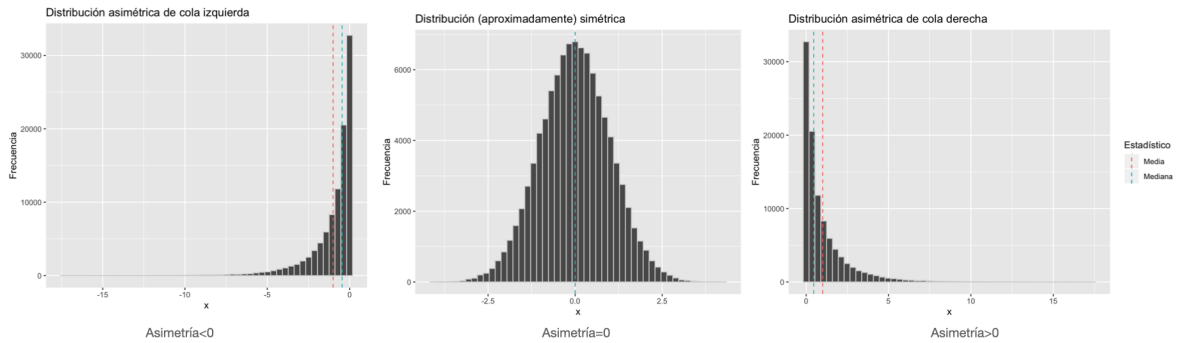
$$Porc = \frac{freq\ categoria}{n} * 100 = p * 100$$

- Tasas:

$$tasa = \frac{freq\ grupo\ 1}{freq\ grupo\ 2} * 100$$

Medidas asociadas a la forma de la distribución

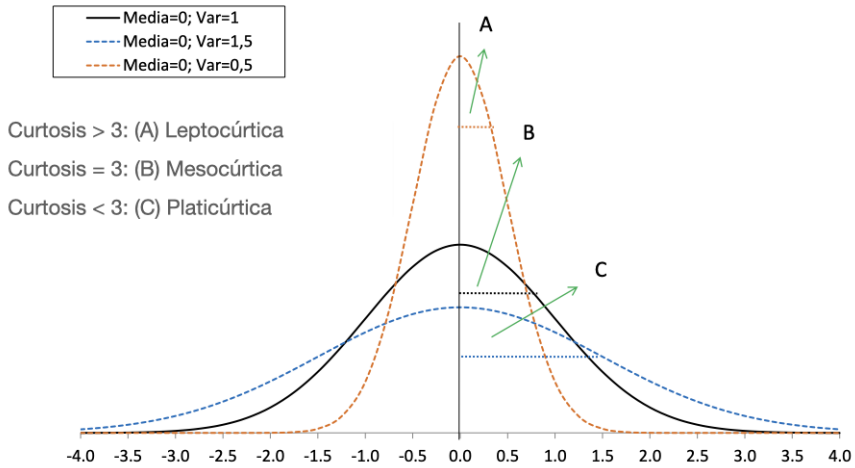
Sesgo o Asimetría: mide la forma en un sentido horizontal.



- **Distribución simétrica** (asimetría = 0): es una distribución que no presenta sesgo y sus indicadores de $\overline{X}$, Mdn y $Moda$ son equivalentes.
- **Asimetría hacia la izquierda** (asimetría < 0) cuando una distribución tiene una asimetría negativa significa que tiene una cola más larga a la izquierda. Esto es lo que se entiende por un sesgo a la izquierda de la distribución.
- **Asimetría hacia la derecha** (asimetría > 0): cuando una distribución tiene una asimetría positiva significa que tiene una cola más largo a la derecha. Esto es lo que se entiende por un sesgo a la derecha de la distribución.

Medidas asociadas a la forma de la distribución

Curtosis: mide la forma en un sentido vertical. Esta nos ayuda a identificar la dispersión de nuestros datos y la existencia de valores extremos.



- Si la **curtosis>3** entonces podemos observar que hay una importante concentración de los datos en torno al promedio.
- Si la **curtosis=3** la distribución representa una distribución normal de los datos.
- Si la **curtosis>3** entonces la distribución presenta una mayor dispersión.

Eventos aleatorios

Un **evento aleatorio** se entenderá se entenderá como el resultado de un fenómeno que no puede determinarse previamente. Y el conjunto de todos los posibles resultados de un evento aleatorio se entiende por **espacio muestral**: Ω.

- **Experimento**: Procedimiento que se puede repetir en muchas oportunidades y en el cual se conocen todos los resultados posibles.

  > Lanzar una moneda.

- **Evento aleatorio**: se entenderá como el resultado de un fenómeno que no puede determinarse previamente.

  > Resultado del lanzamiento de la moneda.

- **Espacio muestral**: el conjunto de todos los resultados de un evento aleatorio.

  > cara/sello.

- **Variable aleatoria**: el registro de los resultados de un experimento.

  > La variable con nuestros resultados.

3. Estadística Inferencial

Probabilidades

Los resultados de un experimento se pueden escribir como una variable aleatoria (categórica onumérica), lo cual facilita la expresión de los sucesos y sus probabilidades.

$\Omega = \{0,1,2,3,4,5,6,7,8,9,10\} \Rightarrow X = 1,2,\ldots,10$

$\Omega = \{0, A, B, AB\} \Rightarrow X = 1,2,3,4$

$\Omega = \{cara, sello\} \Rightarrow X = 0,1$

Ejemplo. Se lanza una moneda al aire hasta que aparezca la primera "cara".

$\Omega = \{c, sc, ssc, sssc, \ldots.\} \Rightarrow X = 1,2,\ldots,\infty$

Probabilidades

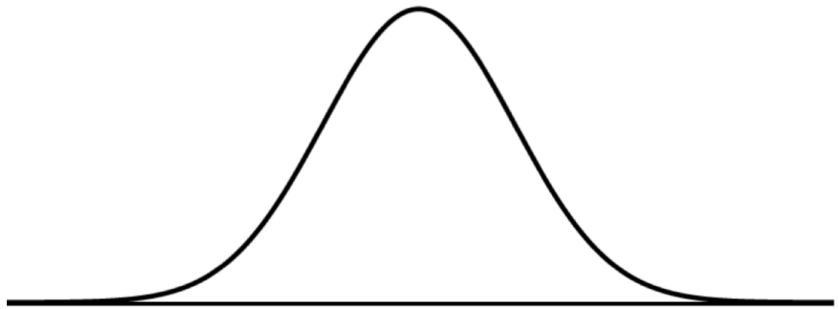
La probabilidad es la herramienta que tenemos para estudiar eventos aleatorios. Es una frecuencia relativa con que puede esperarse que ocurra un evento sabiendo todos los resultados posibles a obtener. En otras palabras nos permiten estudiar la incertidumbre sobre la ocurrencia de eventos aleatorios:

$$P(x) = \frac{\text{N° de formas en que el evento x puede ocurrir}}{\text{N° de casos posibles}}$$

Esta definición de probabilidad, denominada concepto frecuentista, considera la repetición del experimento infinitas veces, siempre bajo las mismas condiciones (i.i.d.).

Probabilidades

La **probabilidad de un resultado** es la proporción de veces que ese resultado ocurriría si pudiésemos observar un evento aleatorio un número infinito de veces.



Dos parámetros: media μ, varianza σ^2

$$X \sim N(\mu, \sigma^2)$$

Simétrica respecto a su media μ

Media = moda = mediana = μ

Inflexión en $X = \mu - \sigma; X = \mu + \sigma$

Donde $\hat{p}_n$ corresponde a la proporción de ocurrencias de un resultado específico.

Probabilidades

Rango de variación: $0 \leq P(x) \leq 1$

Suceso seguro: $P(x) = 1$

Suceso nulo: $P(x) = 0$

Suma de probabilidades: $P(A \cup B) = P(A) + P(B);\ si\ A \cap B = \emptyset$

Suceso complementario: $P(A^c) = 1 - P(A);\ con\ A^c\ complemento\ de\ A$

Eventos mutuamente excluyentes: dos eventos no pueden ocurrir simultáneamente.

$$P(Cara) = \frac{1}{2}$$

$$P(Sello) = \frac{1}{2}$$

$$P(Cara) \cup P(Sello) = P(Cara) + P(Sello) = 1$$

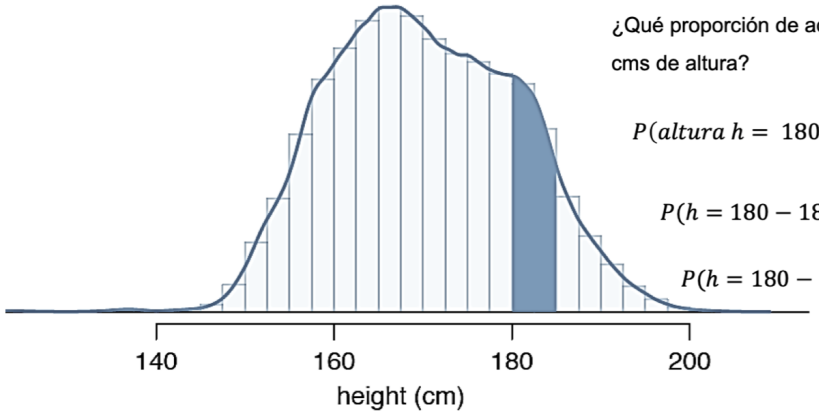
$$P(Cara) \cap P(Sello) = \emptyset$$

Distribuciones de probabilidad

Muchas veces interesa saber cuál es la probabilidad asociada a cada uno de los valores posibles de la variable aleatoria y compararlos. Para ello, se utilizan las distribuciones de probabilidad:

Distribución de probabilidad: Listado que relaciona cada valor de una variable aleatoria con su frecuencia relativa teórica, es decir, su probabilidad de ocurrencia en la población.

Figure 2.29: Density for heights in the US adult population



¿Qué proporción de adultos en EEUU mide entre 180 y 185 cms de altura?

$$P(altura\ h = 180 - 185cm) = \frac{personas\ entre\ 180 - 185}{tamaño\ total\ muestra}$$

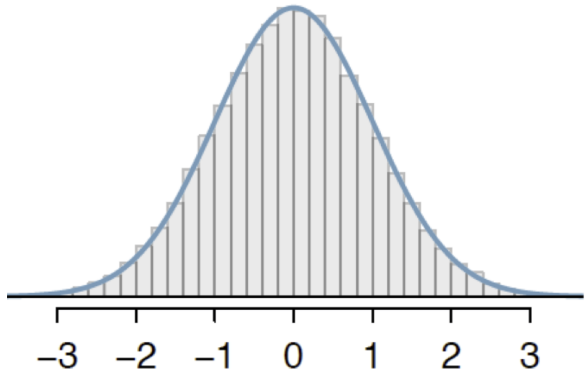
$$P(h = 180 - 185) = area\ bajo\ la\ curva\ histograma$$

$$P(h = 180 - 185) = 0.116$$

La belleza de los datos: La probabilidad de que un adulto elegido aleatoriamente mida entre 180y 185 cms es 0.116

Distribución de probabilidad normal

De gran utilidad: estandarización de variables aleatorias normales



$$X \sim N(\mu, \sigma^2)$$

$$Z = \frac{X - \mu}{\sigma}$$

$$Z \sim N(0,1)$$

Z score = número de desviaciones estándar sobre o bajo la media

Observaciones sobre la media > 0; bajo la media <0

Observaciones iguales a la media = 0

La distribución de probabilidad normal es una de las más comunes, ejemplos los constituyen la distribución del coeficiente intelectual, la distribución de notas en un curso, la distribución del peso, y la distribución de estatura de la población. Se caracteriza por su forma acampanada, su simetría y porque media, mediana y moda coinciden.

Regla empírica

Cuando una distribución está distribuida normalmente, un 68% de los datos está a menos de 1 desviación estándar de la media, el 95% de los datos está a menos de 2 desviaciones estándar de la media y 99,7% está a menos de 3 desviaciones estándar de la media.

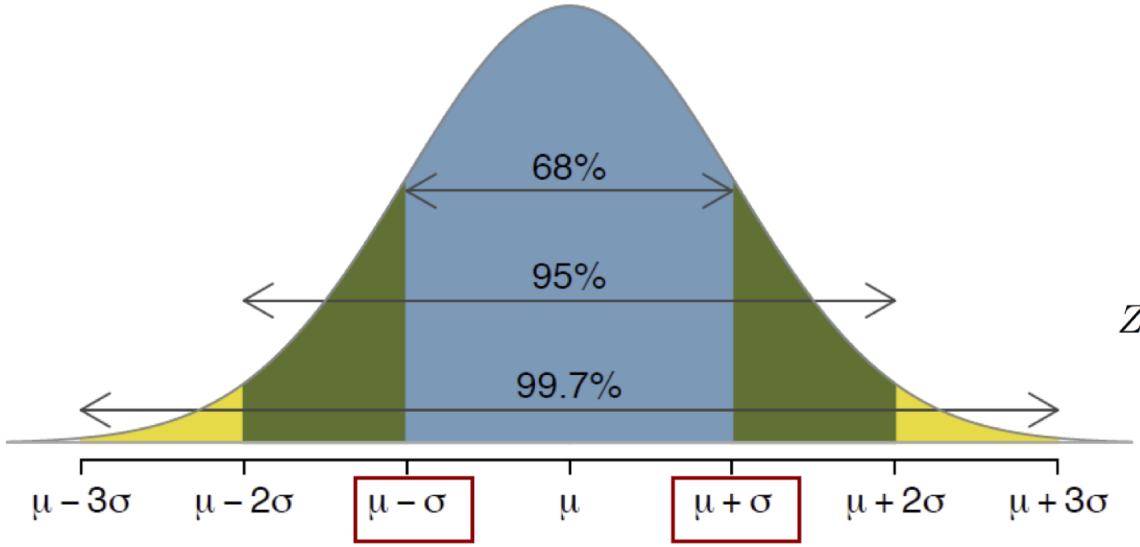


Figure 3.9: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

$$Z \sim N(0,1)$$

$$Z = \frac{x - \mu}{\sigma}$$

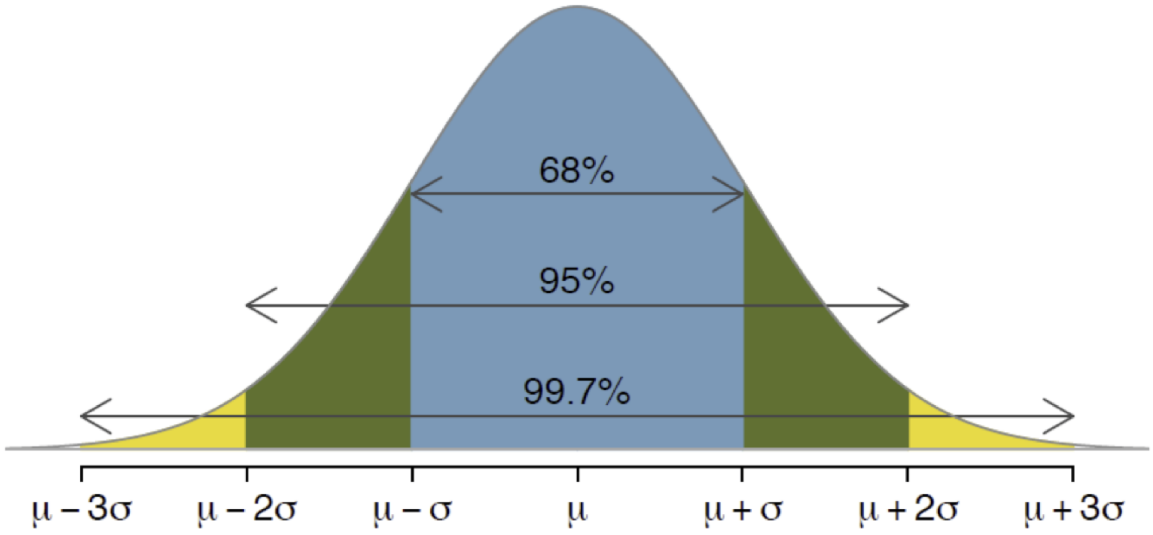
Regla empírica

Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	-2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	-1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	-1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	-1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	-1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	-1.5
0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793	0.0808	-1.4
0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951	0.0968	-1.3
0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151	-1.2
0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357	-1.1
0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587	-1.0

Regla empírica

	Second decimal place of Z									
Z	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

Regla empírica



1 - 0.1587 - (1-0.8413) = 0.6826

Normalidad como condición

Histograma con distribución normal

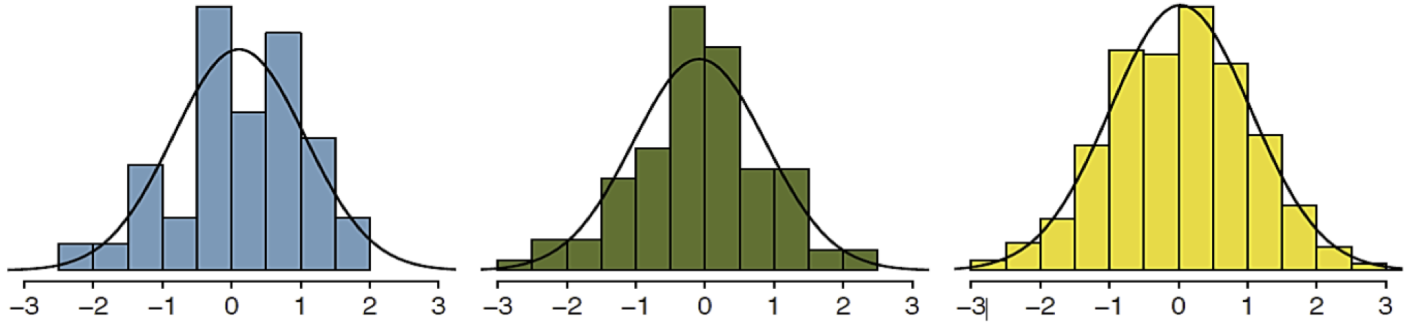
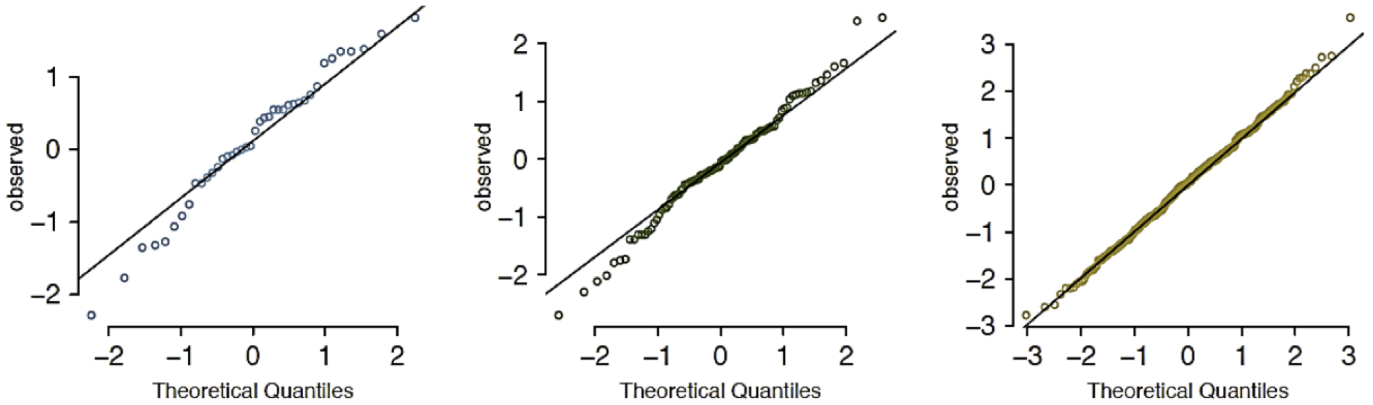


Gráfico de probabilidad normal



$n = 40$ $n = 100$ $n = 400$

Distribución muestral y teorema del límite central

Dado el comportamiento de la distribución *t* Student ante una cantidad importante de grados de libertad y la aplicación de la Ley de los Grandes Números, estamos en condiciones de enunciar el Teorema del Límite Central:

> Las **medias de las muestras aleatorias** simples, extraídas de una población que **se distribuye normalmente**, darán lugar a una distribución muestral que **también es normal**, aunque sean pequeñas.
>
> Si el **tamaño de cada muestra es suficientemente "grande"**, con independencia de la forma de la distribución de la población, las **medias de las muestras tenderán a distribuirse normalmente**.

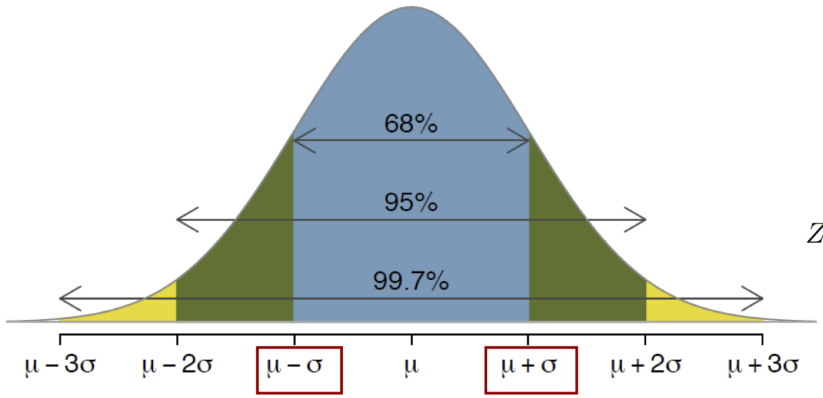


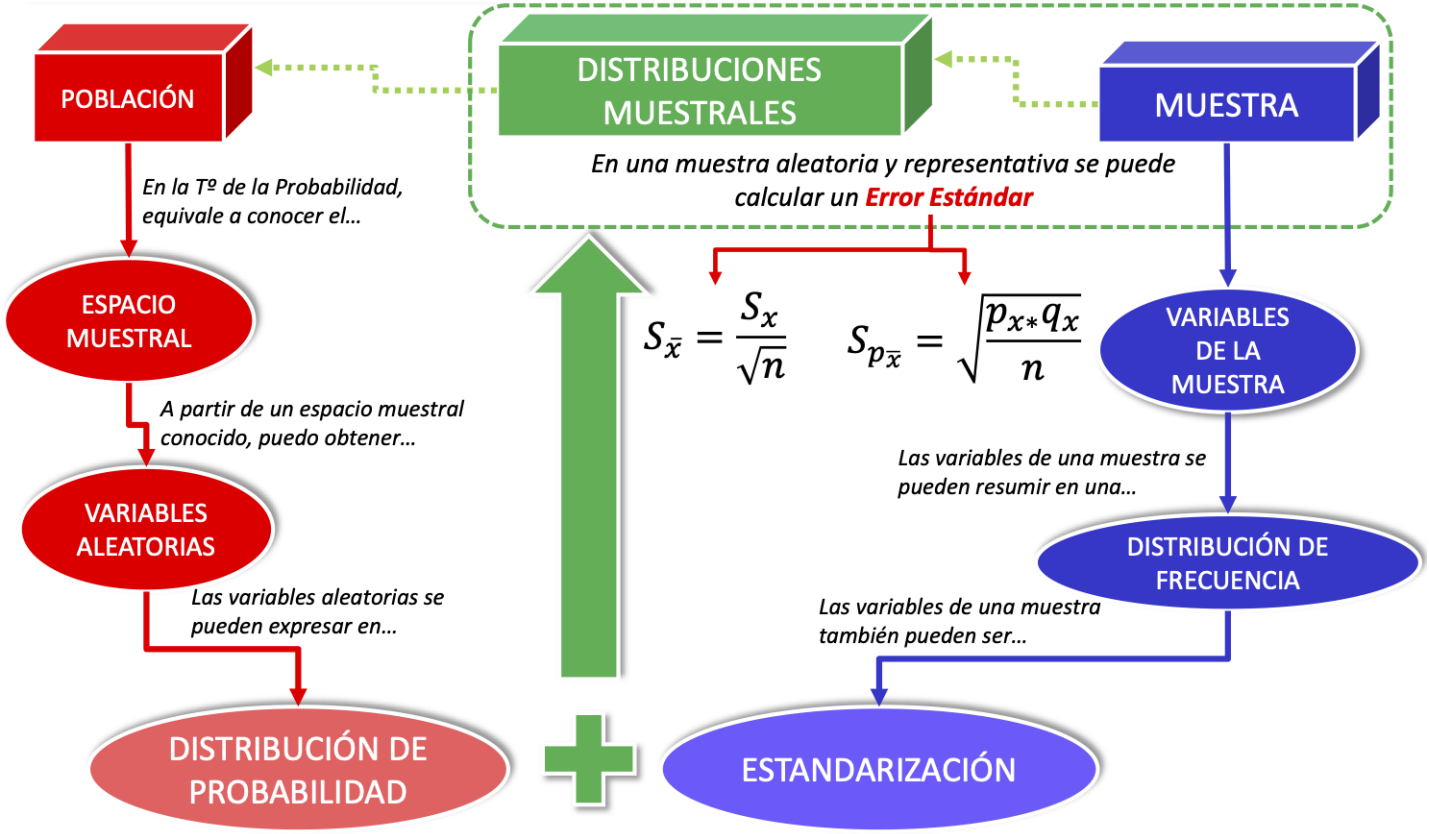
Figure 3.9: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

$$Z \sim N(0,1)$$

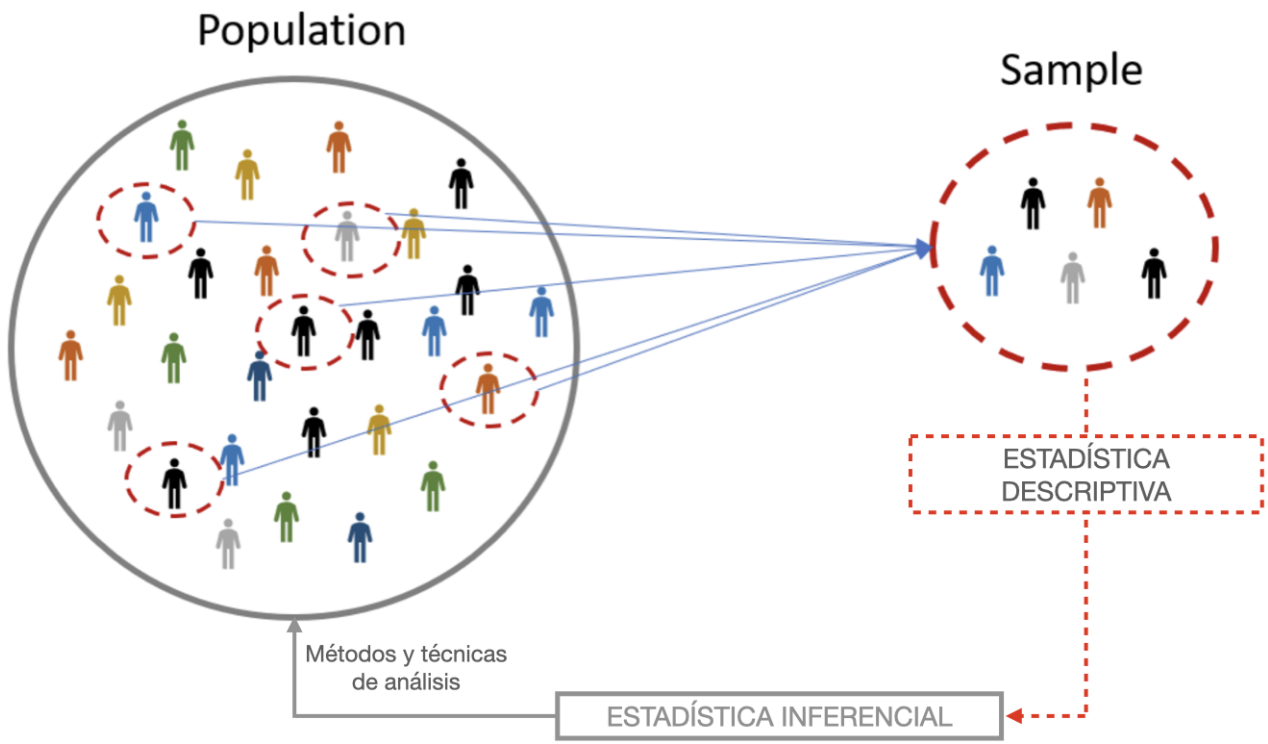
$$Z = \frac{x - \mu}{\sigma}$$

Inferencia estadística

La inferencia estadística proporciona métodos que permiten sacar conclusiones de una población a partir de los datos de una muestra.

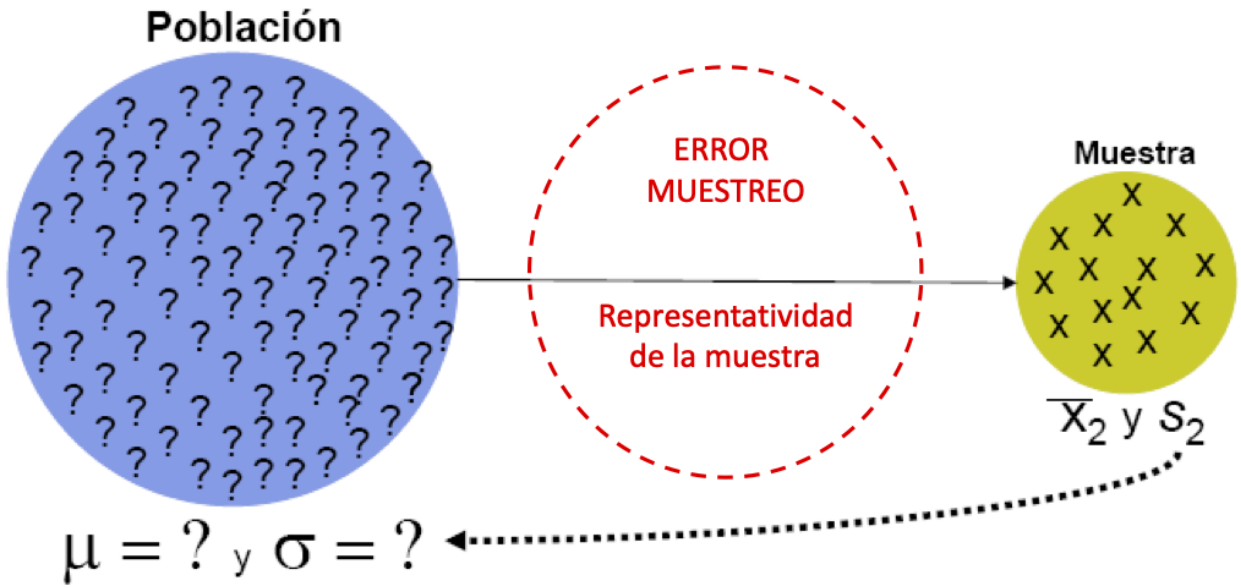


Inferencia estadística



¿Qué es la inferencia estadística?

La estadística inferencial busca elaborar conclusiones de la población a partir de una muestra:



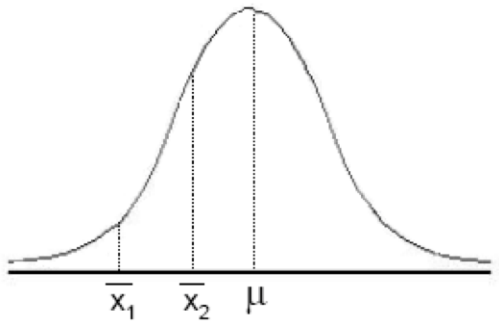
> La muestra no es un fiel espejo de la población (hay un error por el hecho de trabajar con muestras).
>
> Las distribuciones muestrales contienen probabilidades de que el estadístico muestral sea diferente al parámetro poblacional.

Estimación de parámetros

- **Estimación puntual**: es la estimación que utiliza el valor del estadístico muestral como indicador del valor del parámetro.

¿Cuál es el problema de la estimación puntual?

La estimación puntual no incorpora la variabilidad del estadístico entre distintas muestras: esta variabilidad es la probabilidad de lo que sucede en mi muestra no sea lo que ocurre en la población.



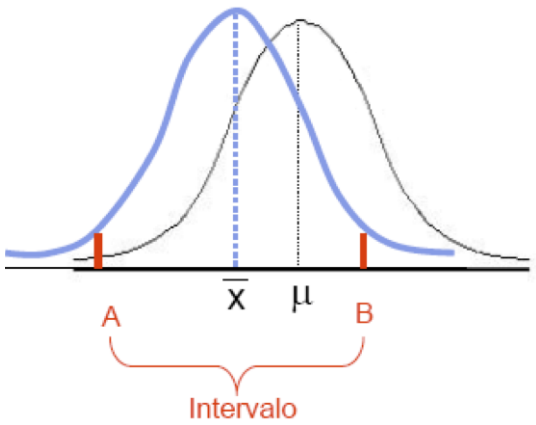
De todas las muestras posibles, puede ser que la media sea: $\bar{x}_1$ ó $\bar{x}_2$.

Ambas medias son diferentes al parámetro, pero hay una que está más próxima.

Para poder estimar qué pasa en la población, se debe incorporar la variabilidad del estadístico muestral.

Estimación por intervalo

La estimación por intervalo del parámetro poblacional es aquella que identifica un rango de valores entre los que puede situarse el parámetro poblacional. Estos valores se determinan a partir de la muestra.



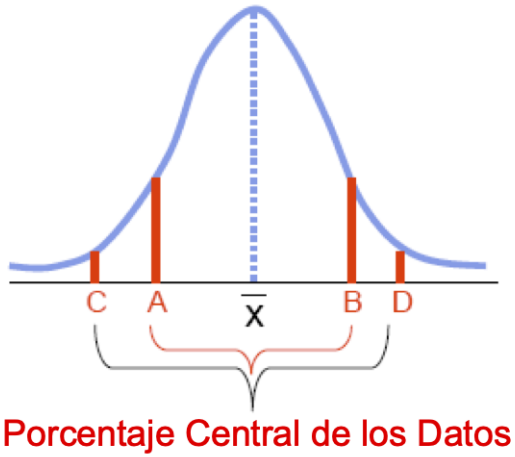
El estadístico muestral **puede estar en cualquier punto de la distribución muestral**.

El intervalo es un rango determinado a partir de valores de la muestra (tamaño, desviación estándar y media) que tiene una **probabilidad conocida de comprender el parámetro**.

La probabilidad de que el intervalo contenga el parámetro se calcula por el área bajo la curva.

Nivel de confianza y significancia

¿Cómo se determina entre qué rangos de valores está el parámetro?. La respuesta depende de cuán confiados deseemos estar de que el parámetro está en ese rango. De esta forma:

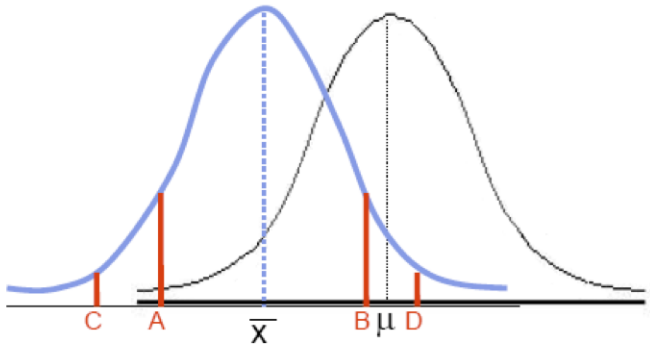


Nivel de confianza ($1-\alpha$): Es la probabilidad de que la muestra elegida produzca un intervalo que incluya el parámetro que se está estimando.

Nivel de significancia (α): Probabilidad de que el intervalo calculado a partir de la muestra no comprenda el parámetro.

Nivel de confianza y significancia

Mientras más grande el nivel de confianza, el rango se amplía y por lo tanto aumenta la certeza (la probabilidad de incluir el parámetro).



Entre A y B está el 95% de los datos.

Entre C y D está el 99,7% de los datos.

Intervalos de confianza

Estimador puntual es una valor posible para un parámetro poblacional

¿Y si usamos un rango de valores?

El intervalo de confianza es una estimación intervalar que busca capturar un paramétro poblacional (desconocido) bajo una cierta probabilidad en muestras repetidas o nivel de confianza. Si consideramos un nivel de confianza ($1 - \alpha$) del 95% podemos obtener una medida de posibles valores para el parámetro poblacional de la siguiente manera:

$$Estimador\ Puntual\ \pm\ 1.96 * EE$$

Donde EP es el estimador puntual y EE el error estándar. En este caso el estimador está a $\pm$ 1.96 desviaciones estándar del parámetro en un 95% de las veces.

SUPUESTO:

1. Los datos provienen de una muestra aleatoria de la población.
2. Normalidad en la distribución o tamaños muestrales grandes (por Teorema del Límite Central)

Intervalos de confianza

Error estándar nos da una guía del tamaño del intervalo de confianza

Podemos tener (aprox.) un 95% de confianza en que hemos encontrado el parámetro poblacional si:

$$Estimador\ Puntual\ \pm\ 1.96 * EE$$

Notas:

> Aproximadamente verdad, se cumple muy bien para distribuciones normales y muestras grandes
>
> El estimador está a menos de 1.96 desviaciones estándar del parámetro poblacional un 95% de las veces

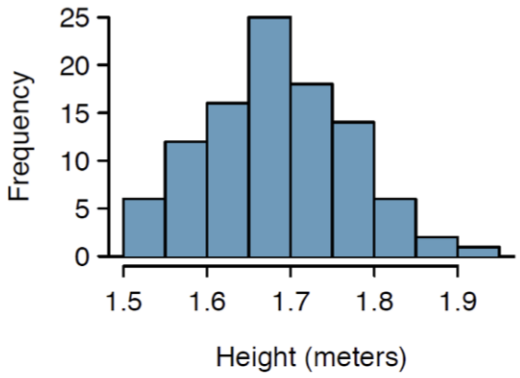
Intervalos de confianza: Altura a un 95%

$$\bar{x}_{height} = \frac{1.50 + 1.78 + \cdots + 1.70}{100} = 1.697$$

Estimador puntual de la media

$$s_{height} = 0.088 \text{ meters}$$

Estimador puntual desviación estándar



$$SE_{\bar{x}} = \frac{0.088}{\sqrt{100}} = 0.0088\, metros$$

$$1.697 \pm 2 \times 0.0088 \rightarrow (1.6794, 1.7146)$$

Intervalos de confianza: Altura a un 95%

	Second decimal place of Z									
Z	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

Intervalos de confianza: Días físicamente activo a un 95%

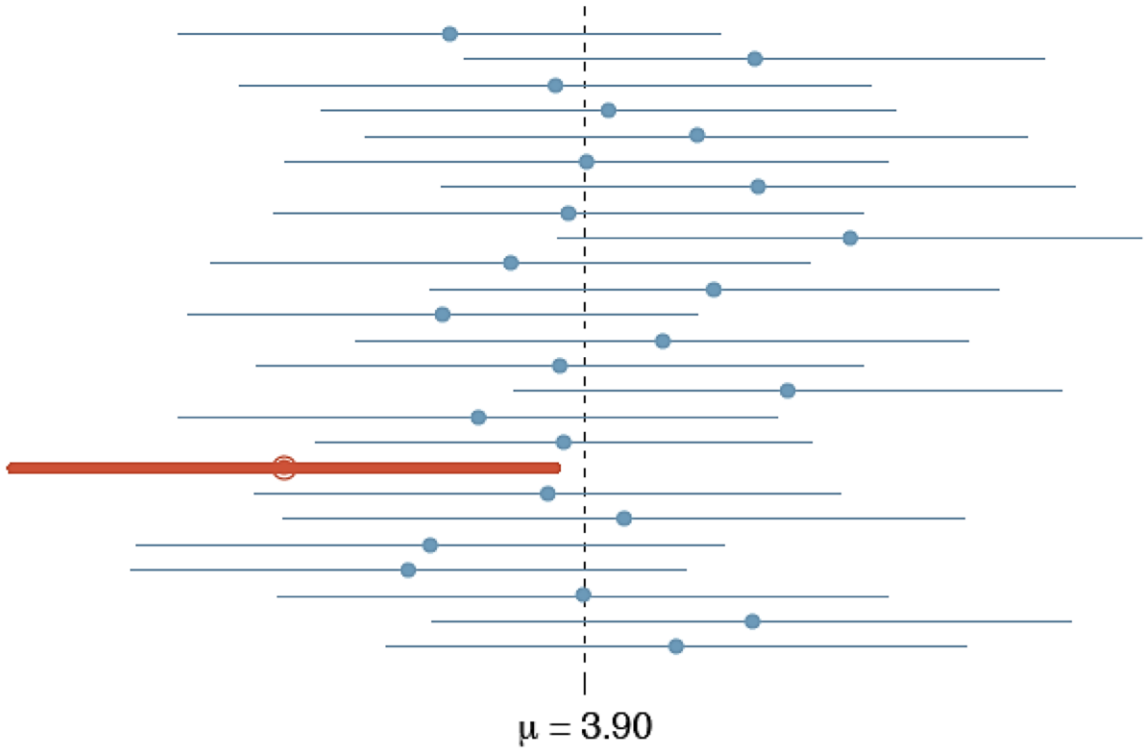


Figure 4.8: Twenty-five samples of size $n = 100$ were taken from `yrbss`. For each sample, a confidence interval was created to try to capture the average number of days per week that students are physically active. Only 1 of these 25 intervals did not capture the true mean, $\mu = 3.90$ days.

Intervalos de confianza: Días físicamente activo a un 95%

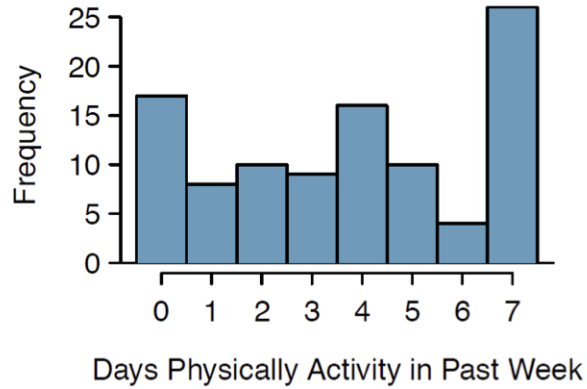
active	estimate	parameter
mean	3.75	3.90
median	4.00	4.00
st. dev.	2.556	2.564

Table 4.5: Point estimates and parameter values for the `active` variable. The parameters were obtained by computing the mean, median, and SD for all YRBSS respondents.

$$\bar{x}_{active} = \frac{0 + 7 + \cdots + 1}{100} = 3.75$$

$$SE = \frac{2.6}{\sqrt{100}} = 0.26 \text{ days}$$

$$3.75 \pm 2 \times 0.26 \quad \rightarrow \quad (3.23, 4.27)$$



Referencias

Diez D., D Barr C., and Çetinkaya-Rundel M. (2019) Open Intro Statistics, fourth edition. OpenIntro, Inc.

> Capítulo 4. "Distribution of random variables". pp. 131-144
>
> Capítulo 5. "Foundations for inference". pp. 168-189

Moore, D. S. (2005). Estadística aplicada básica. Antoni Bosch editor.

> Capítulo 1. "Análisis de distribuciones". Tema 1.4 Distribuciones normales pp. 54-78
>
> Capítulo 4. "Distribuciones muestrales y probabilidad". Tema 4.4 Distribuciones de la media muestra pp. 298-313

Clase 2

Fundamentos de la estadística para las CSC

09 de junio, 2023

José D. Conejeros | jdconejeros@uc.cl | JDConejeros