

Taller:

Estadística para Data Science



Docente:

Melanie Oyarzún W.

Phd (c) in Social Complexity Sciences (UDD)
Mg. Economía (PUC)
Ing. Comercial (PUCV)
moyarzunw@udd.cl

Outline



Introducción:
La estadística y la
ciencia de datos



Fundamentos
estadísticos clave

Fenómenos y variables
aleatorias
Muestras y Estadígrafos



Análisis Exploratorio de
datos

- Analisis univariados
 - Tendencia central
 - Dispersión
 - Distrubuciones
- Analisis multivariados
 - Correlación y covarianza
 - Gráficos de dispersion
 - Causalidad



Inferencia estadística

Pruebas de hipótesis
Intervalos de confianza
Experimentos y pruebas A/B



Cierre y Consultas



Objetivos de aprendizaje del taller

- Abstraer un problema mediante modelamiento estadístico, identificando las variables aleatorias clave.
- Describir las principales variables de un problema, gráfica y numéricamente.
- Realiza inferencia a partir de una muestra de parámetros poblacionales como la esperanza.
- Compara poblaciones y grupos mediante pruebas de hipótesis.
- Desarrolla el análisis estadístico involucrado en un experimento aleatorio.



Metodología del taller

- A cada sesión del taller le acompaña un notebook de trabajo, que está a su disposición en el github del taller.
- Además, hay un notebook de ejercitación, que iremos desarrollando en ambas sesiones, el cual será la única evaluación.
 - 50% entrega completa
 - 50% Evaluación del problema de aplicación:
 - Pauta: de evaluación
 - Planteamiento
 - Justificación
 - Resultados
 - Análisis

O'REILLY®

Practical Statistics for Data Scientists

50+ Essential Concepts Using R and Python



Peter Bruce, Andrew Bruce
& Peter Gedeck

Second
Edition

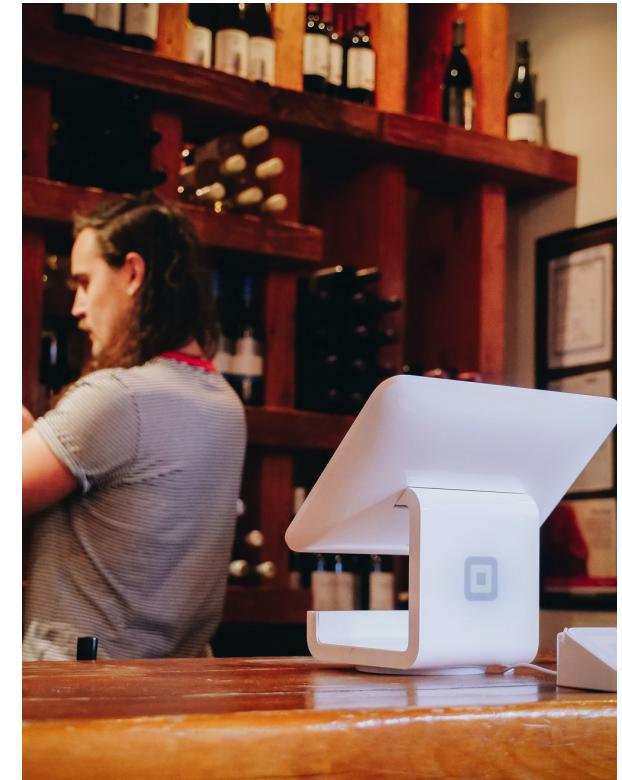
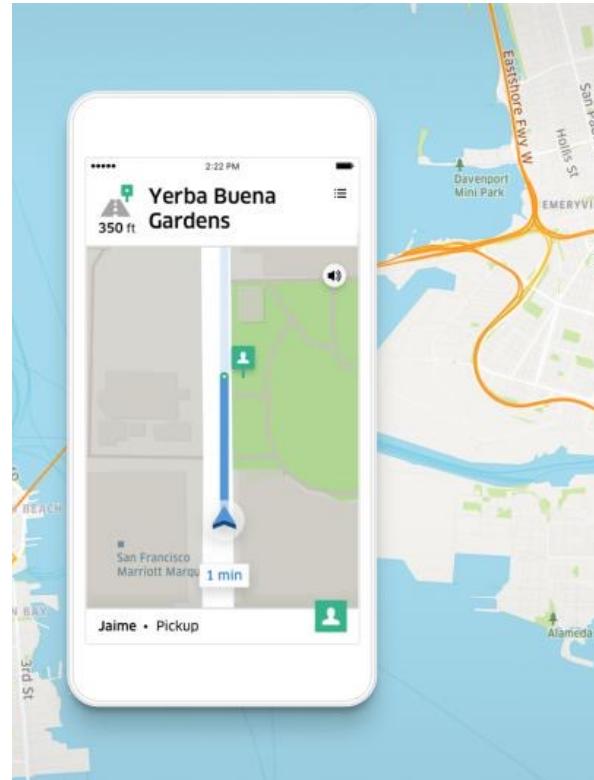
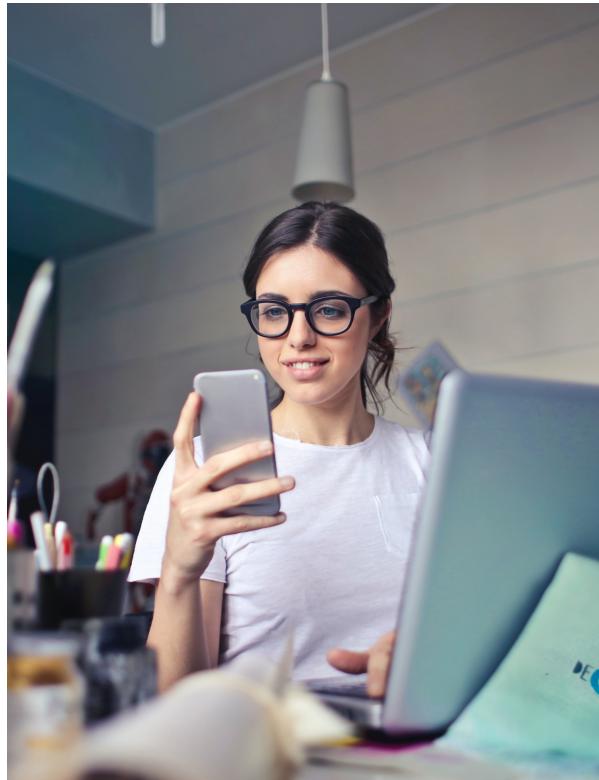
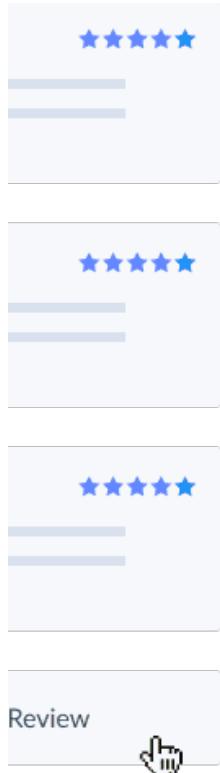
Bibliografía recomendada

Capítulos: 1- 2- 3-4-5

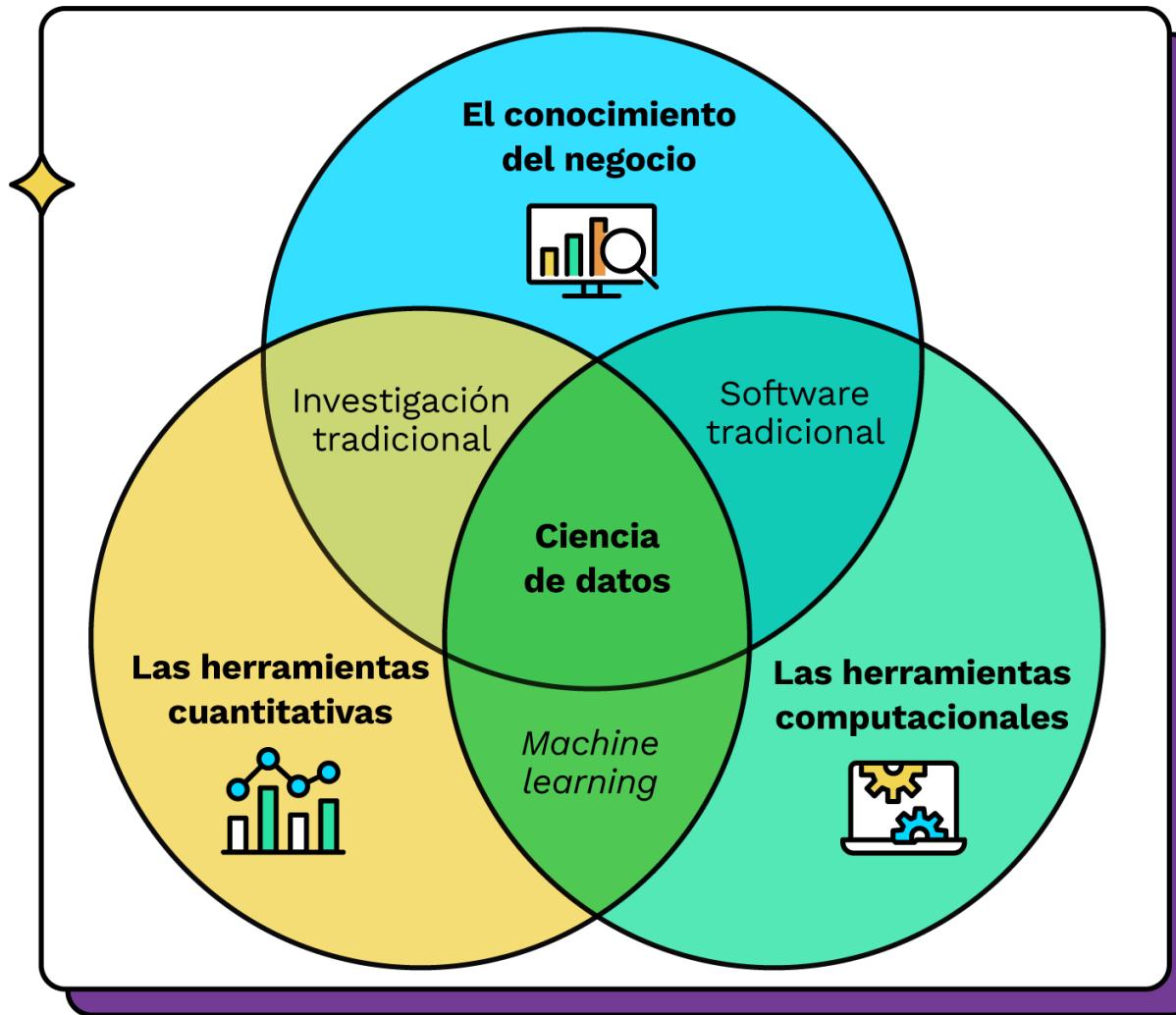
Introducción

Estadística en la ciencia de datos

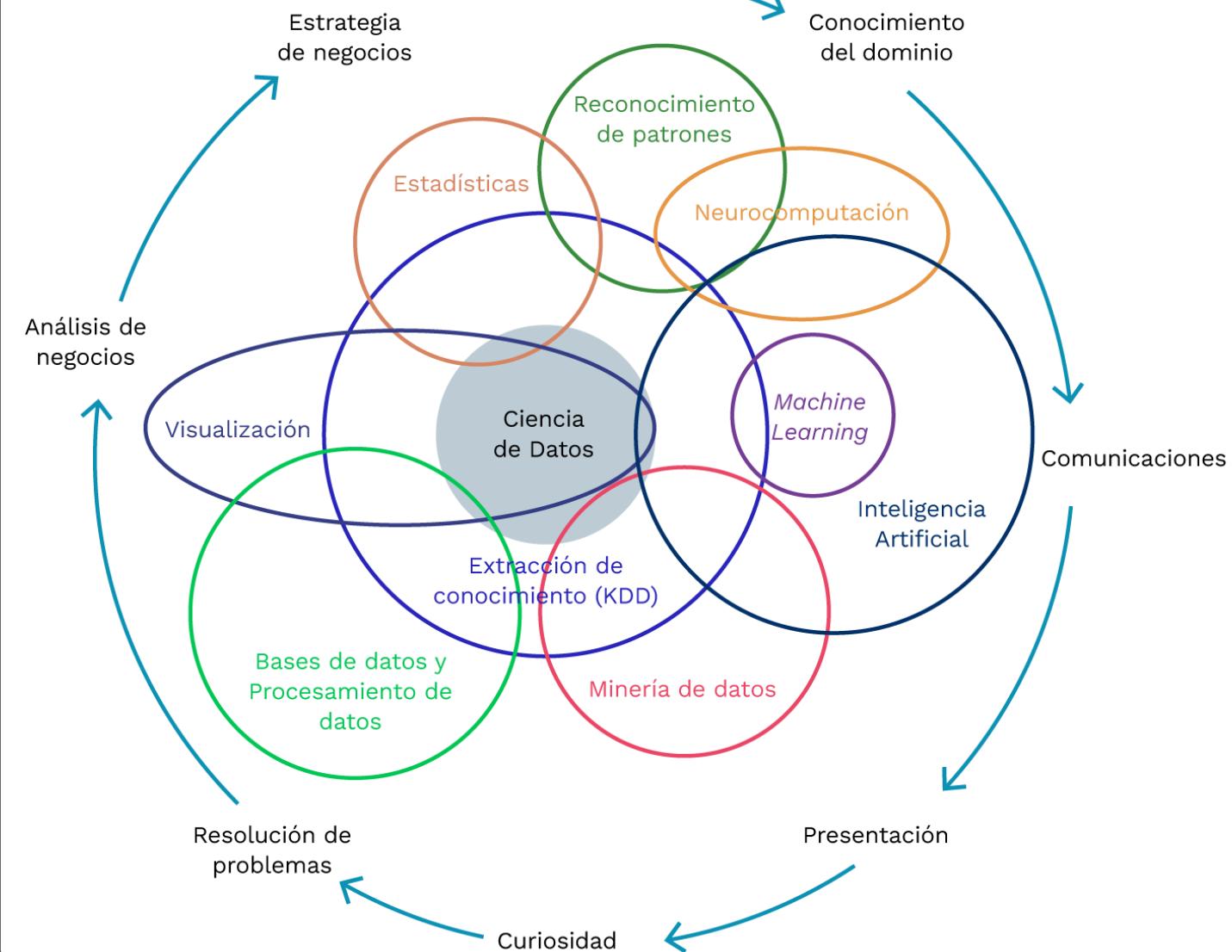
Trazas digitales están en todos lados



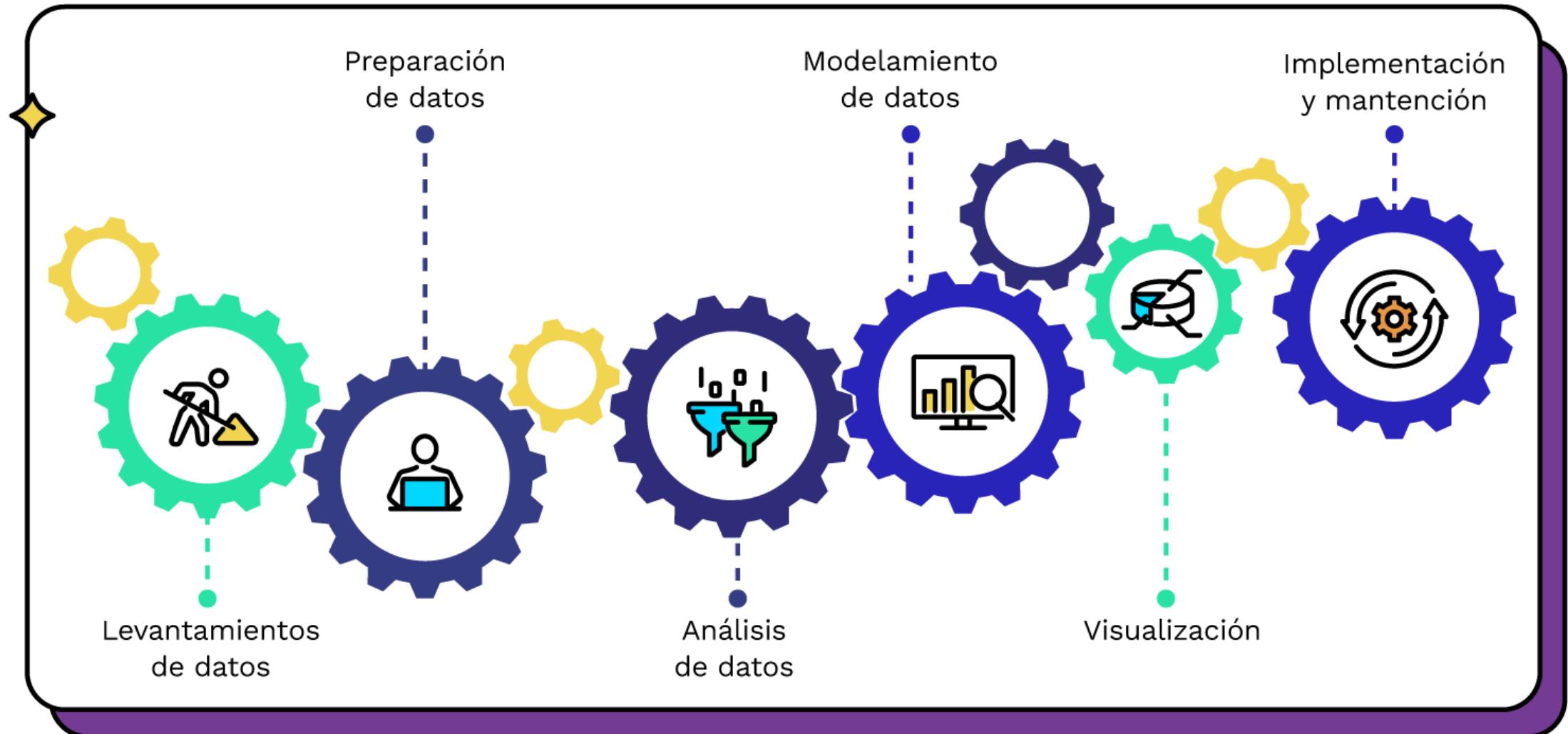
Ciencia de Datos



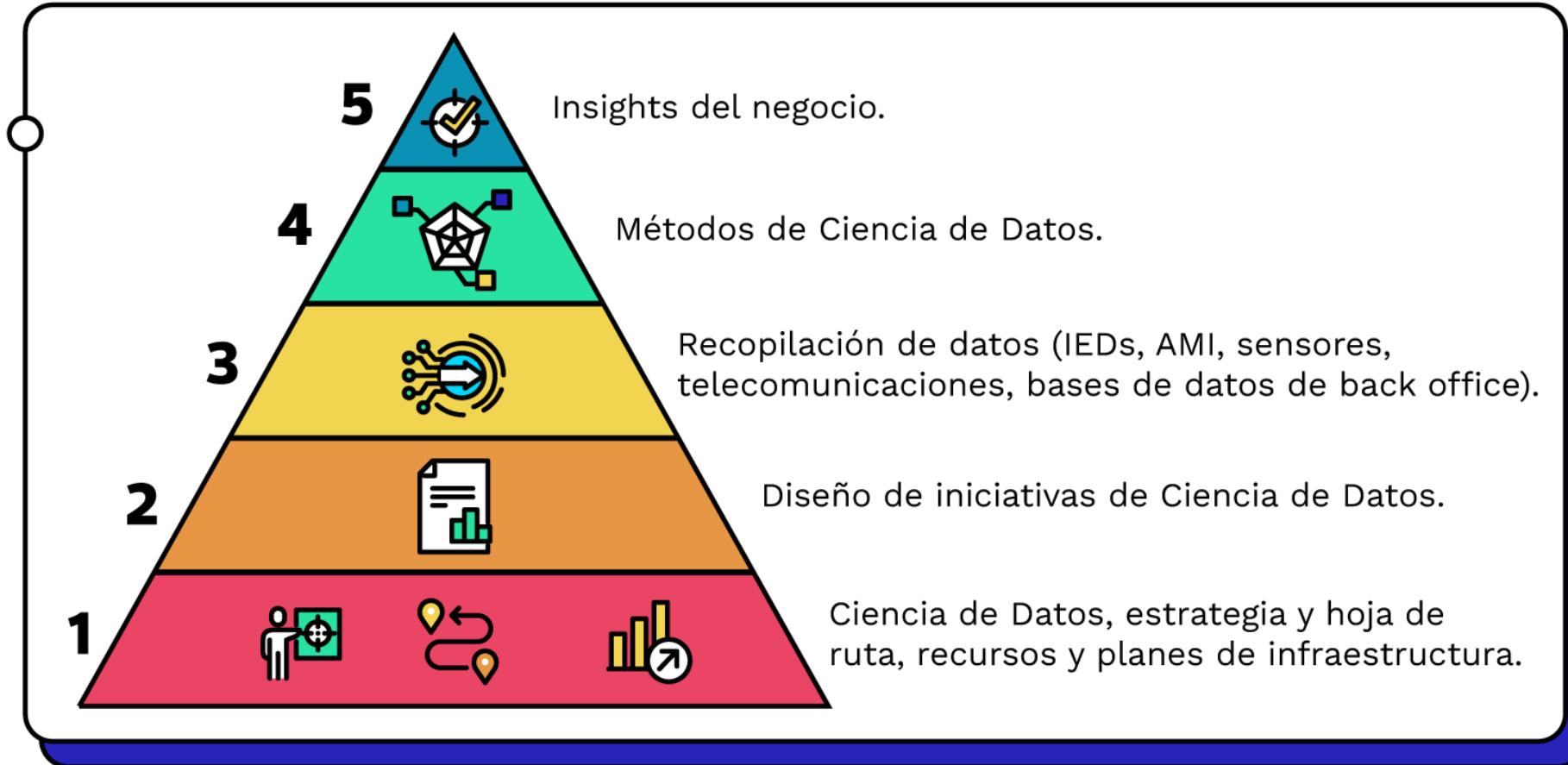
La ciencia de datos es multidisciplinaria



El proceso de la ciencia de datos



Objetivo último: Insights



Abstrayendo y modelando

Bases estadísticas





Herramientas estadísticas



Describir

Fenomeno

Numerico

Grafico



Inferir

Muestra & Población

Estadígrafos

Pruebas de hipótesis

Experimentos



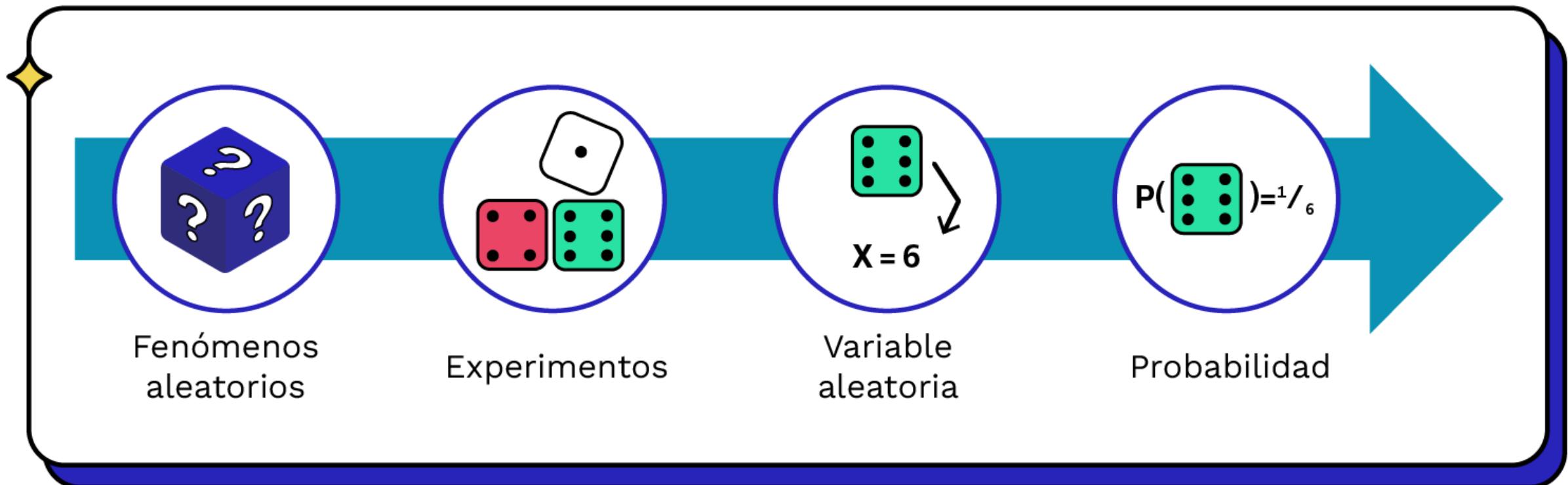
Modelar

Explicar

Predecir

- Análisis de redes
- Análisis de texto

Definiendo fenómenos



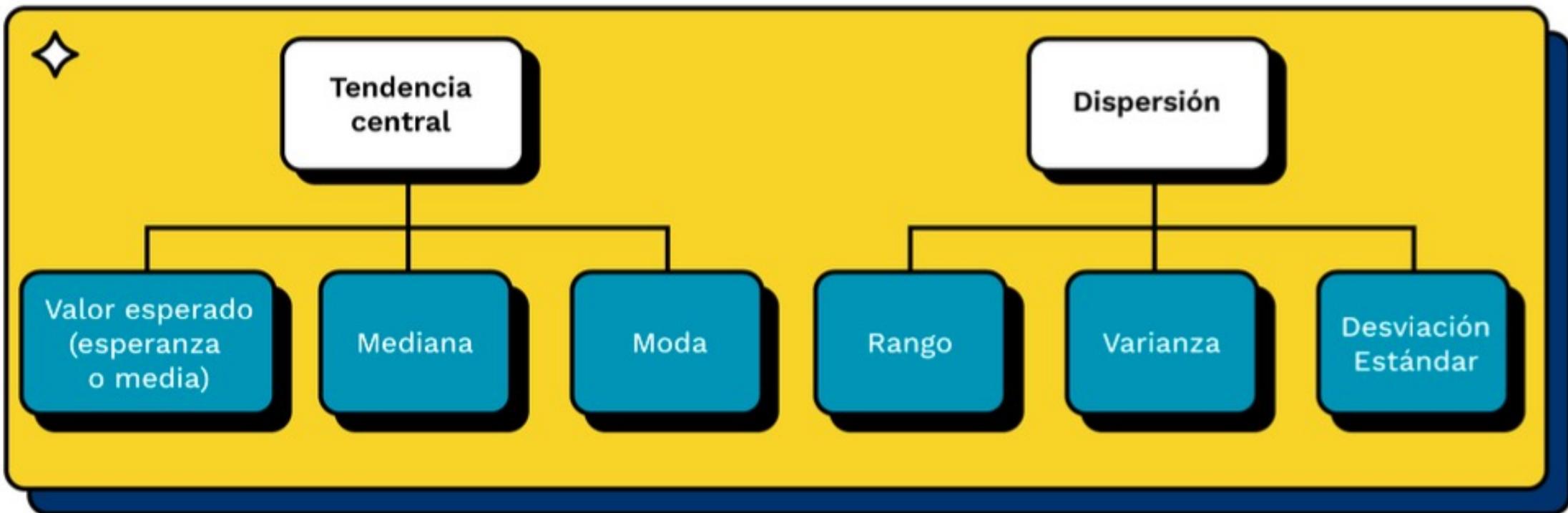
Ejemplo:

Resultado	X	f(x)
	1	$\frac{1}{6}$
	2	$\frac{1}{6}$
	3	$\frac{1}{6}$
	4	$\frac{1}{6}$
	5	$\frac{1}{6}$
	6	$\frac{1}{6}$

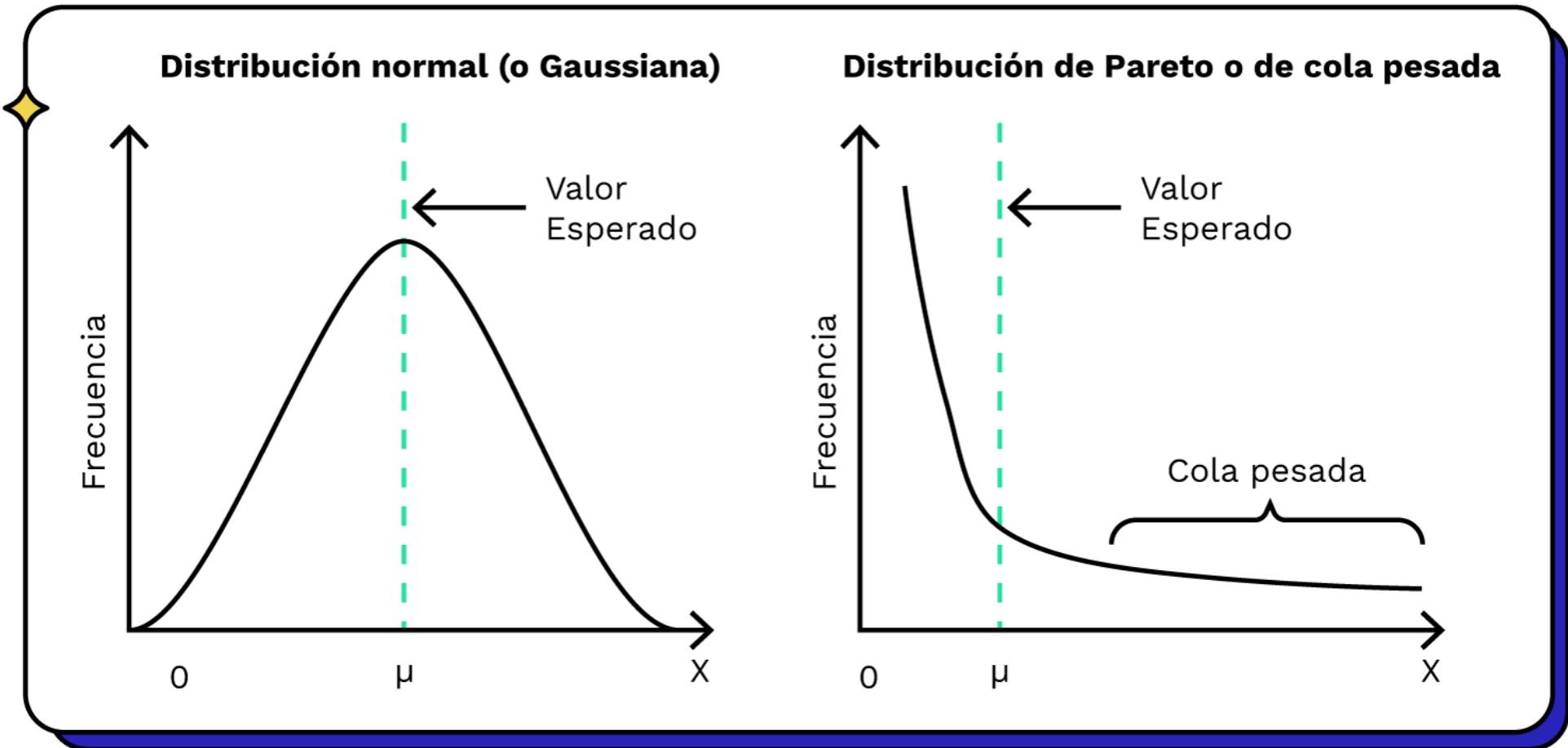
Describiendo variables aleatorias

- **Resúmenes numéricos:** Podemos resumir información sobre valores clave a través de números estos son los parámetros
 - Tendencia central: Valores que se tienden a repetir con mayor frecuencia
 - Dispersión: Variabilidad de los posibles resultados
- **Distribución:** Muestra a los valores obtenidos, para encontrar patrones de estos

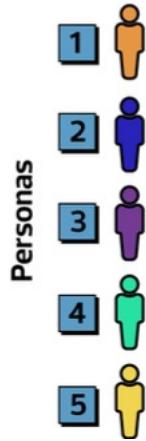
Parámetros



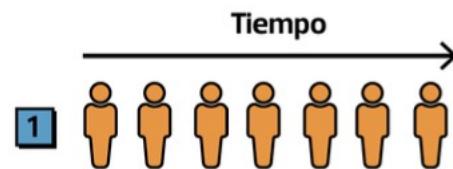
Distribución



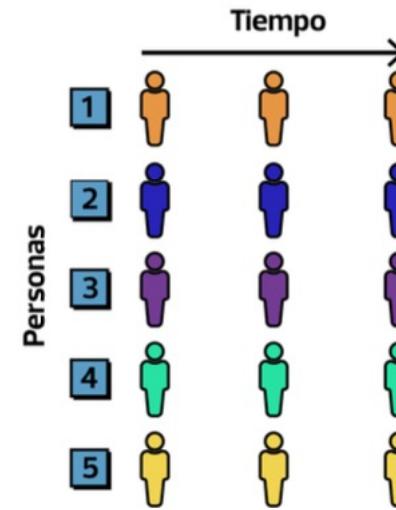
Tipos de datos



Corte
transversal

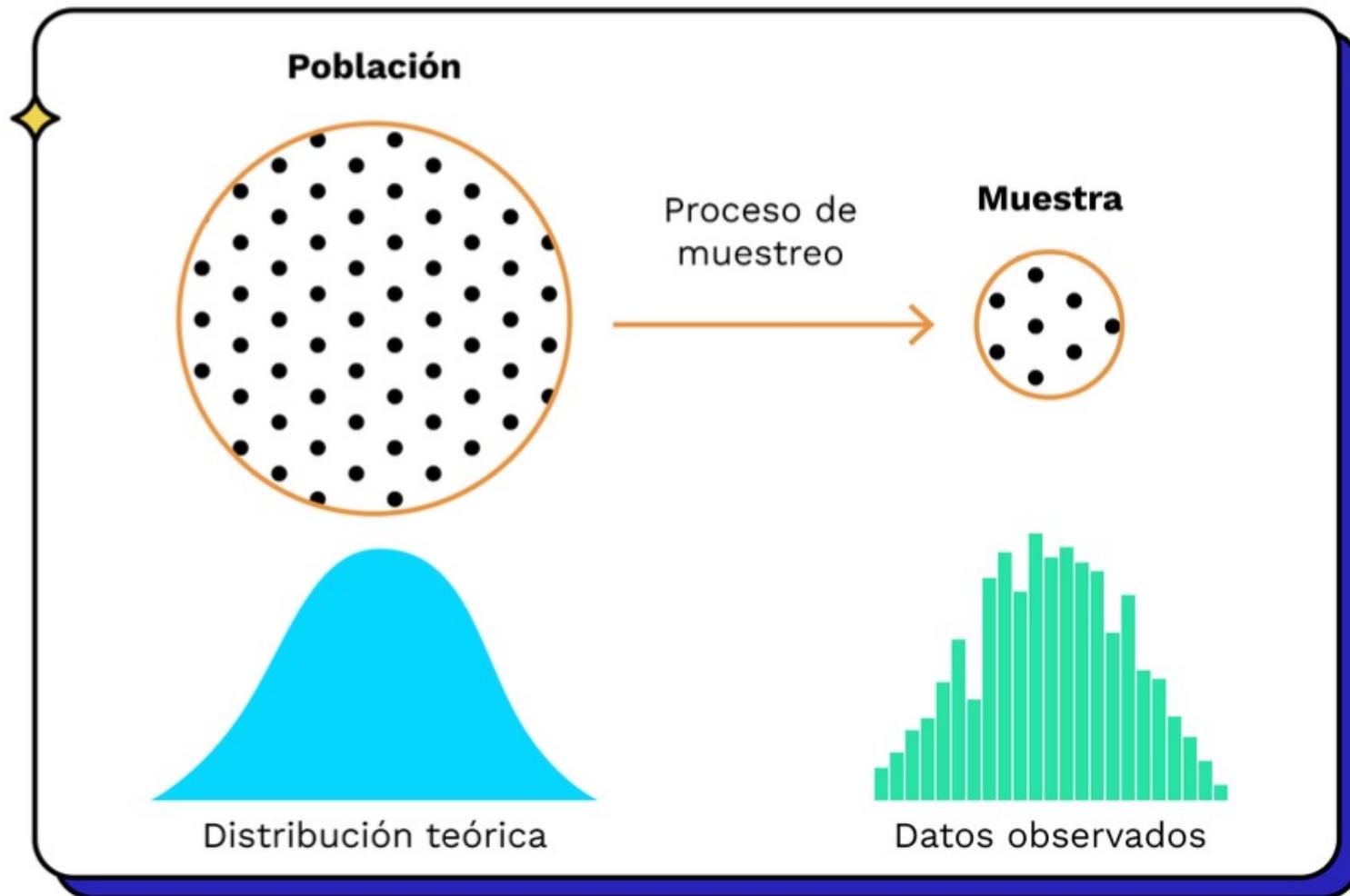


Series de
tiempo

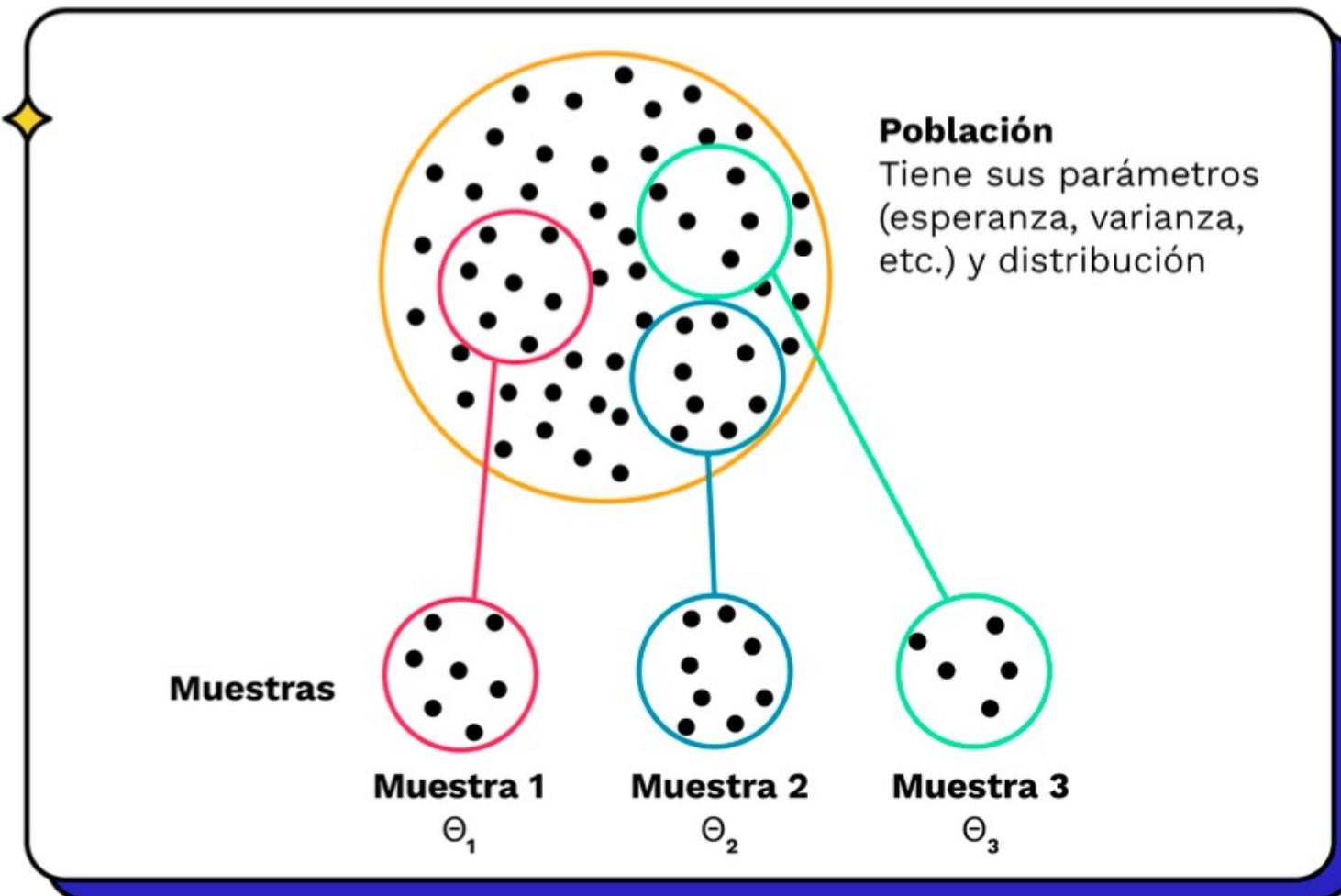


Panel o datos
longitudinales

Población y muestra



Estadígrafos



Estadígrafos comunes

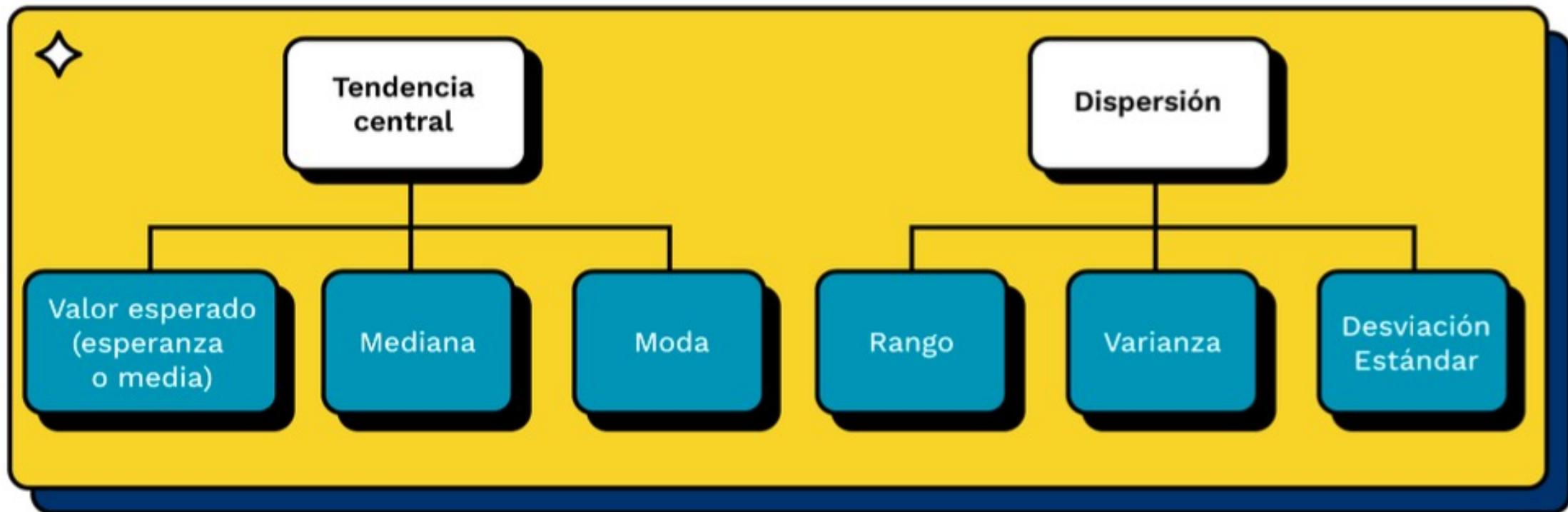
Parámetro poblacional	Definición	Estimador más usado	Fórmula estimador con muestra tamaño n
Esperanza	$E[X] = \int x_i p(X=x_i)$ (caso continuo)	Media muestral o promedio	$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$
Varianza	$V[X] = E[(X - E[X])^2]$	Varianza muestral	$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2$
Covarianza	$Cov[X,Y] = E[(X - E[X])(Y - E[Y])]$	Covarianza muestral	$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})$



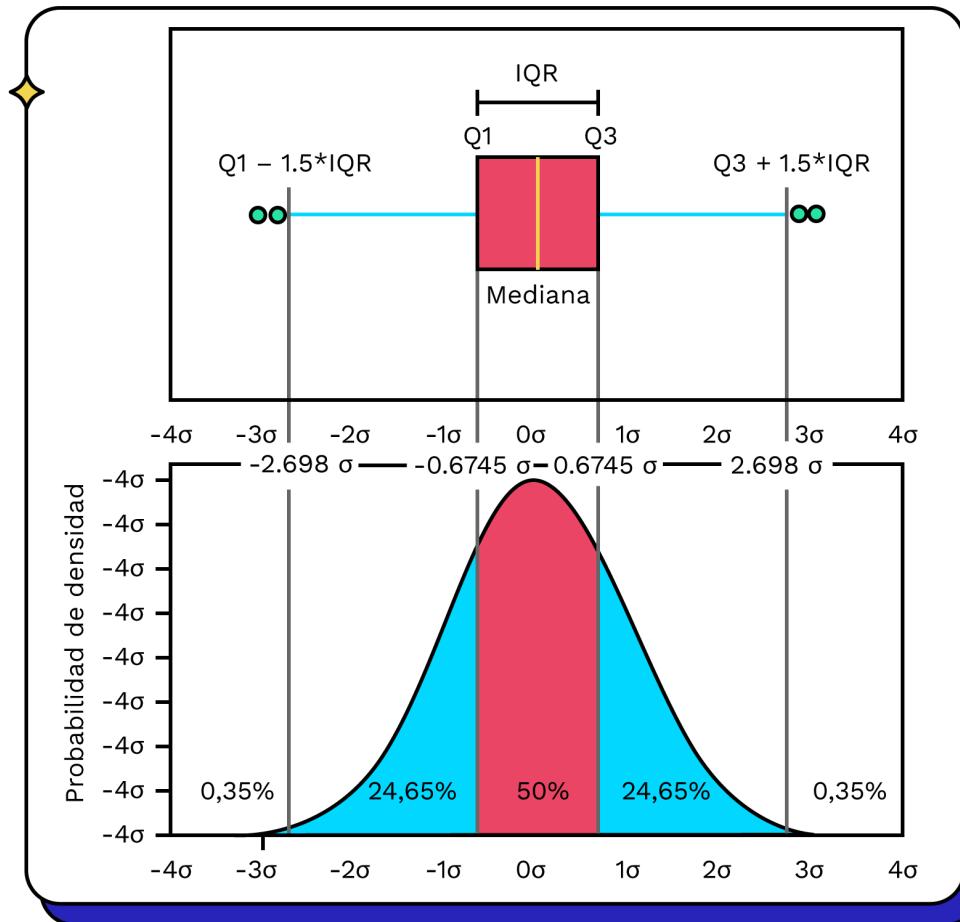
I. Análisis exploratorio de datos

Definición de variables, visualización

Resúmenes numéricos

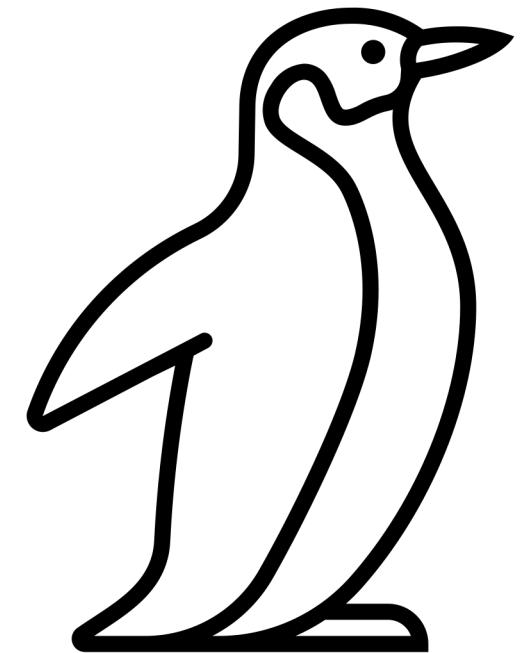


Visualización de datos individuales



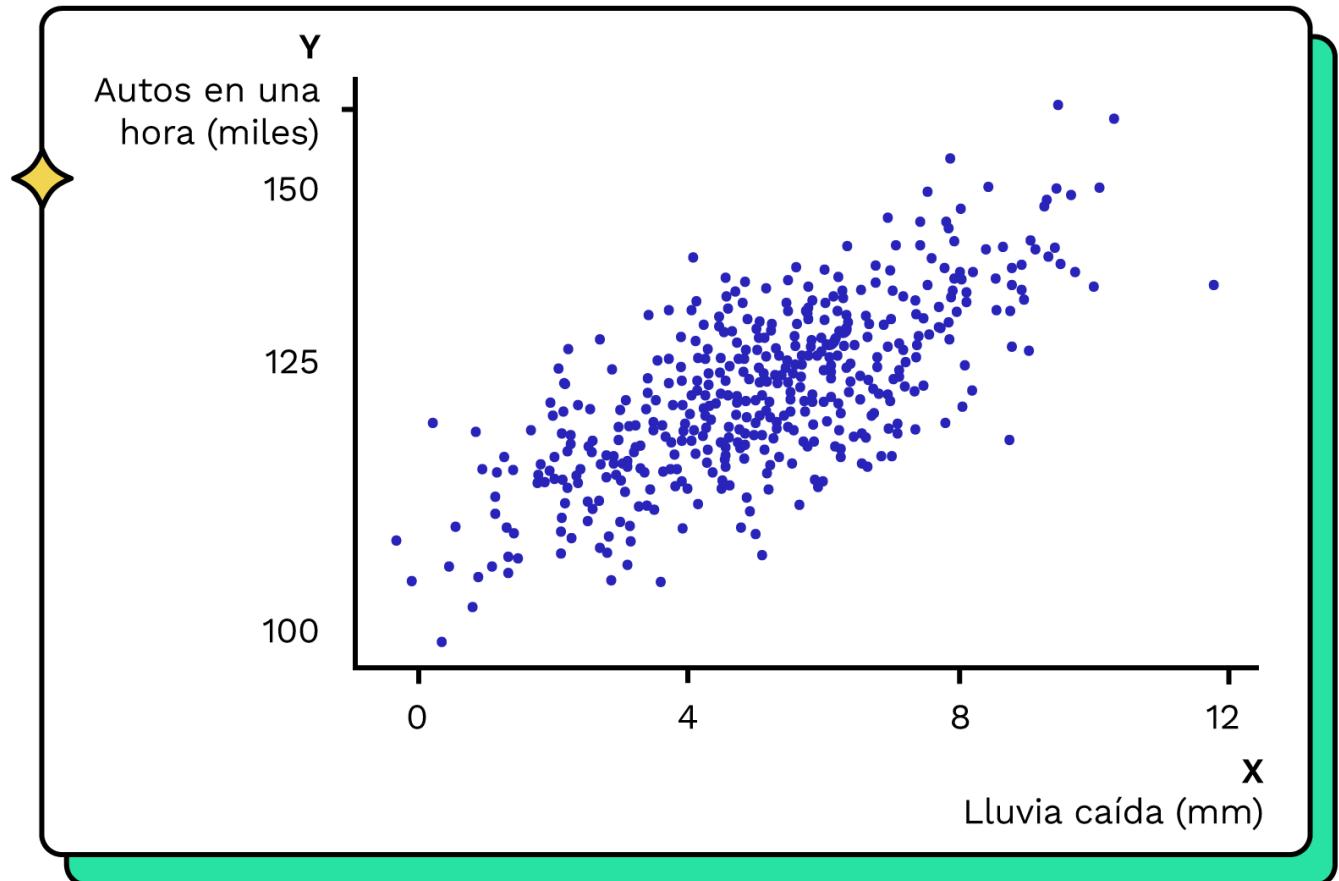
Ejemplo aplicación. Pingüinos Palmer

- Los datos "Palmer Penguins" son un conjunto que detalla medidas morfológicas y características de tres especies de pingüinos: Adelie, Gentoo y Chinstrap. Recopilados por el Dr. Bill Link y su equipo. (Horst AM, Hill AP, Gorman KB (2020). palmerpenguins: Palmer Archipelago (Antarctica) penguin data. doi:10.5281/zenodo.3960218, R package version 0.1.0, <https://allisonhorst.github.io/palmerpenguins/index.html>)



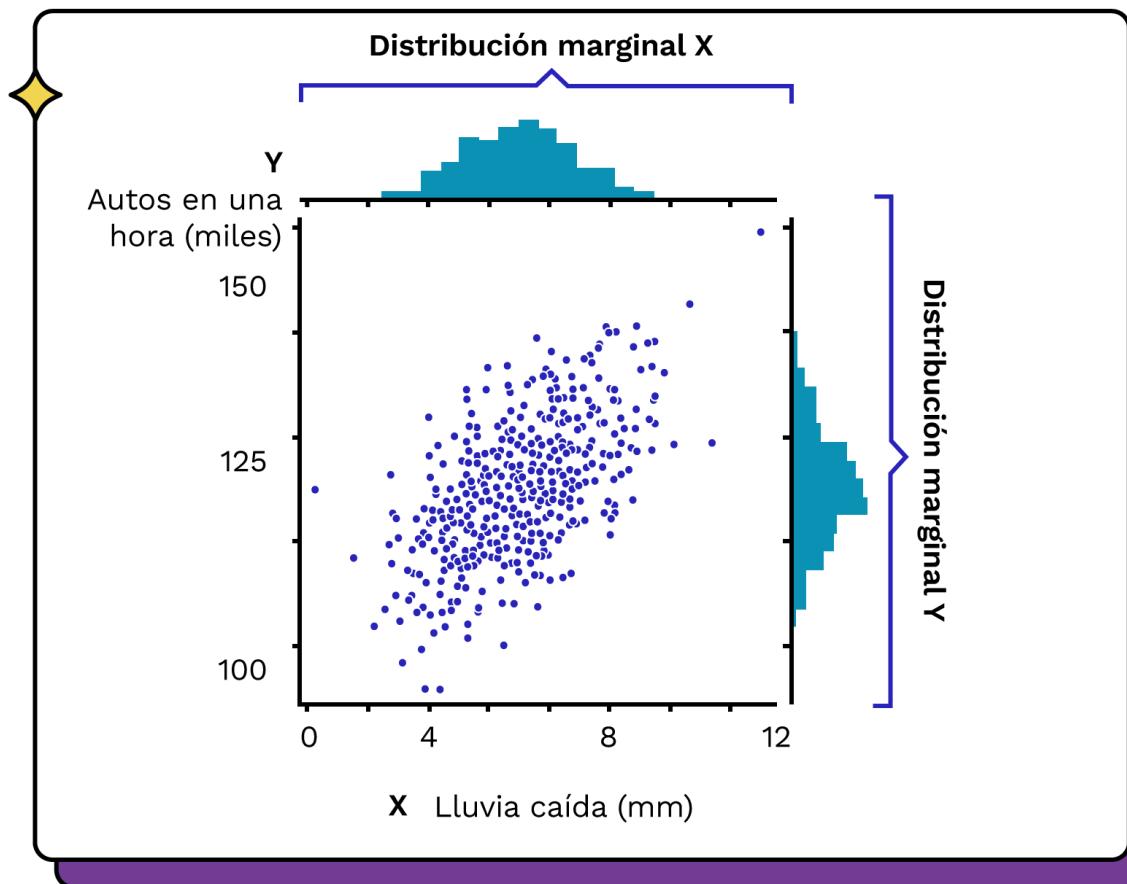
Relacionando variables aleatorias

Muchas veces queremos entender la relación entre variables:

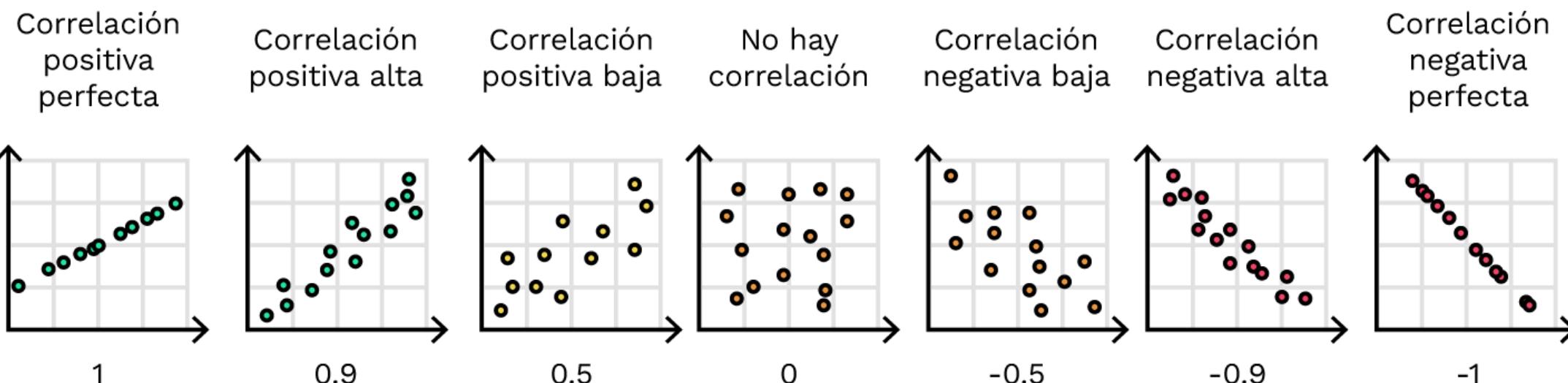


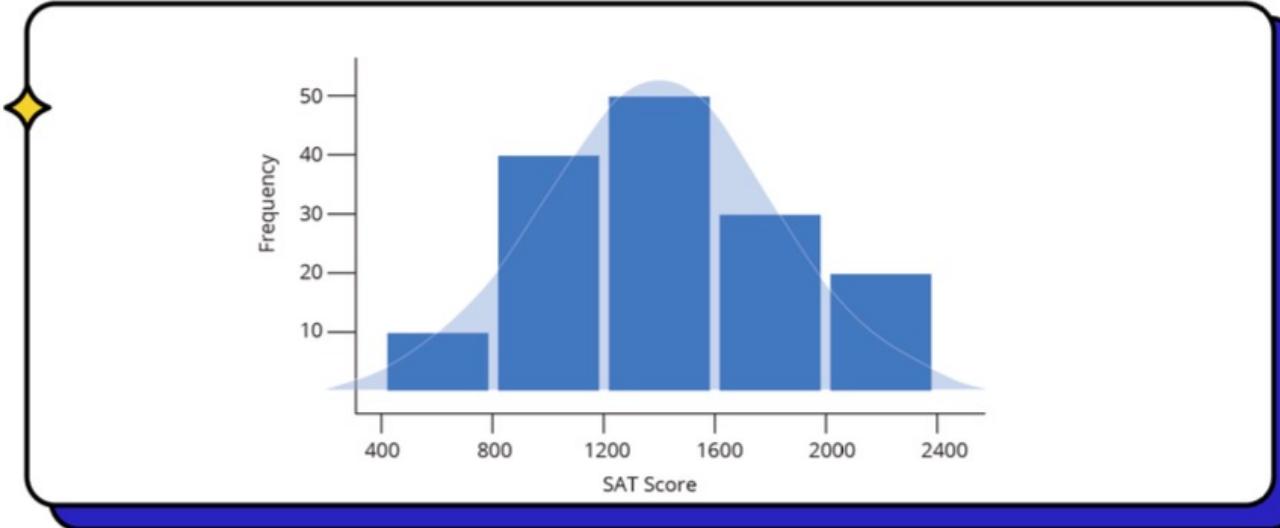
Relacionando variables aleatorias

- Muchas veces queremos entender la relación entre variables:

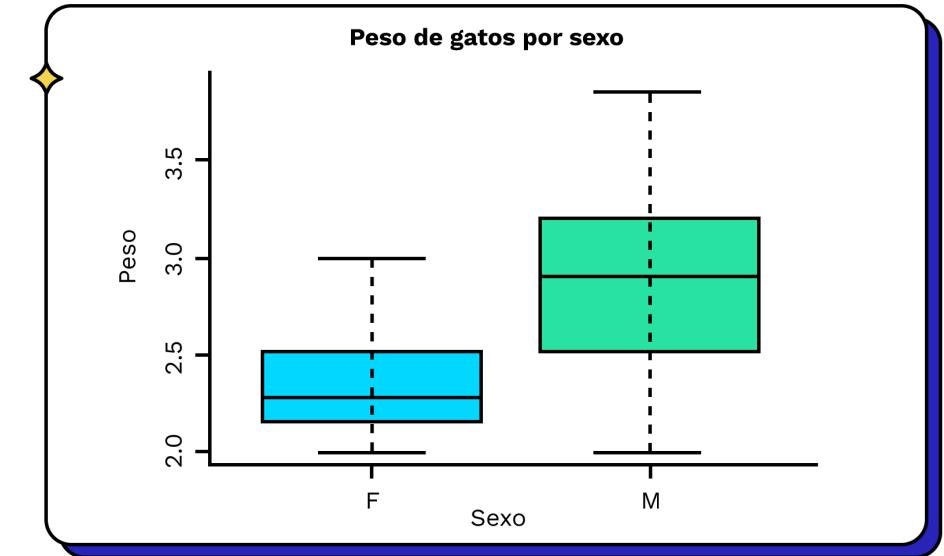


Midiendo la relación entre variables aleatorias



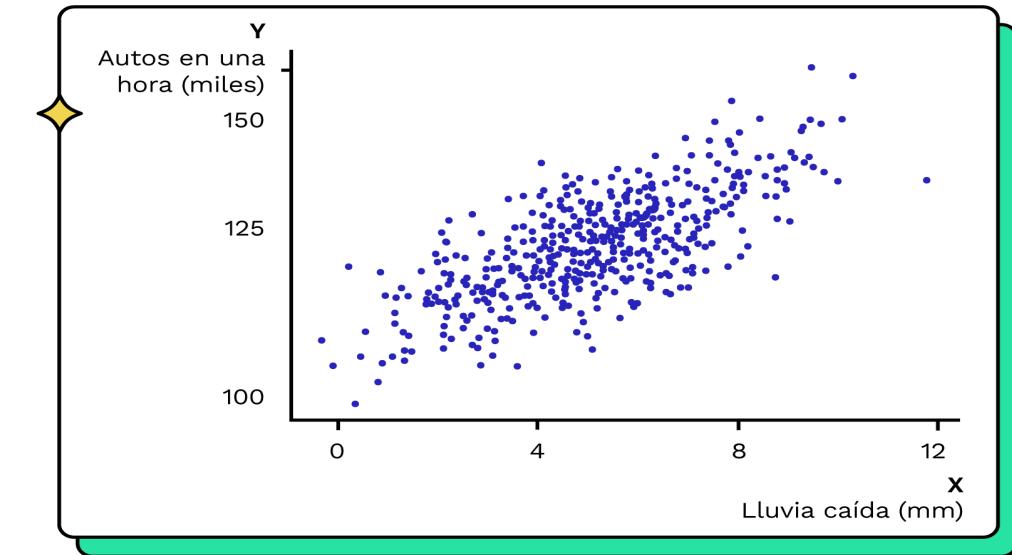
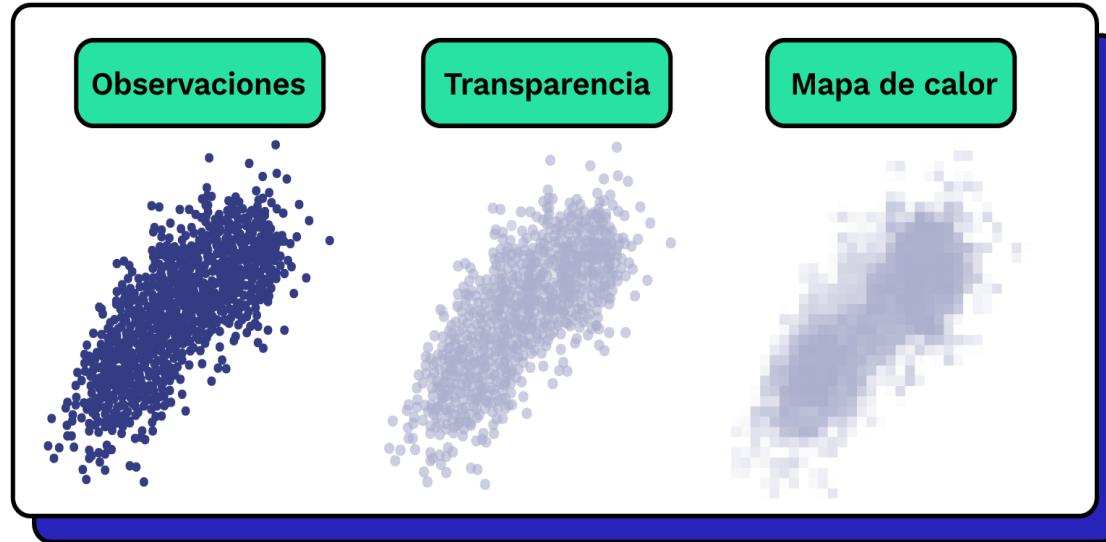


Histogramas para aproximar distribuciones



Box plot resumen distribuciones

Visualización de datos muestrales



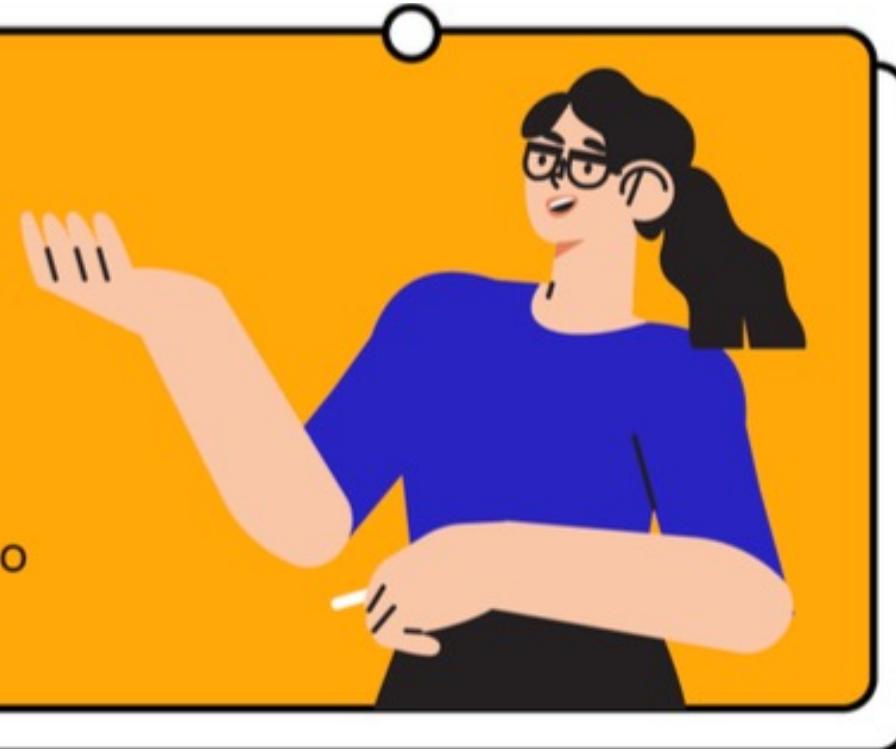
Mapas de calor

Graficos de dispersión

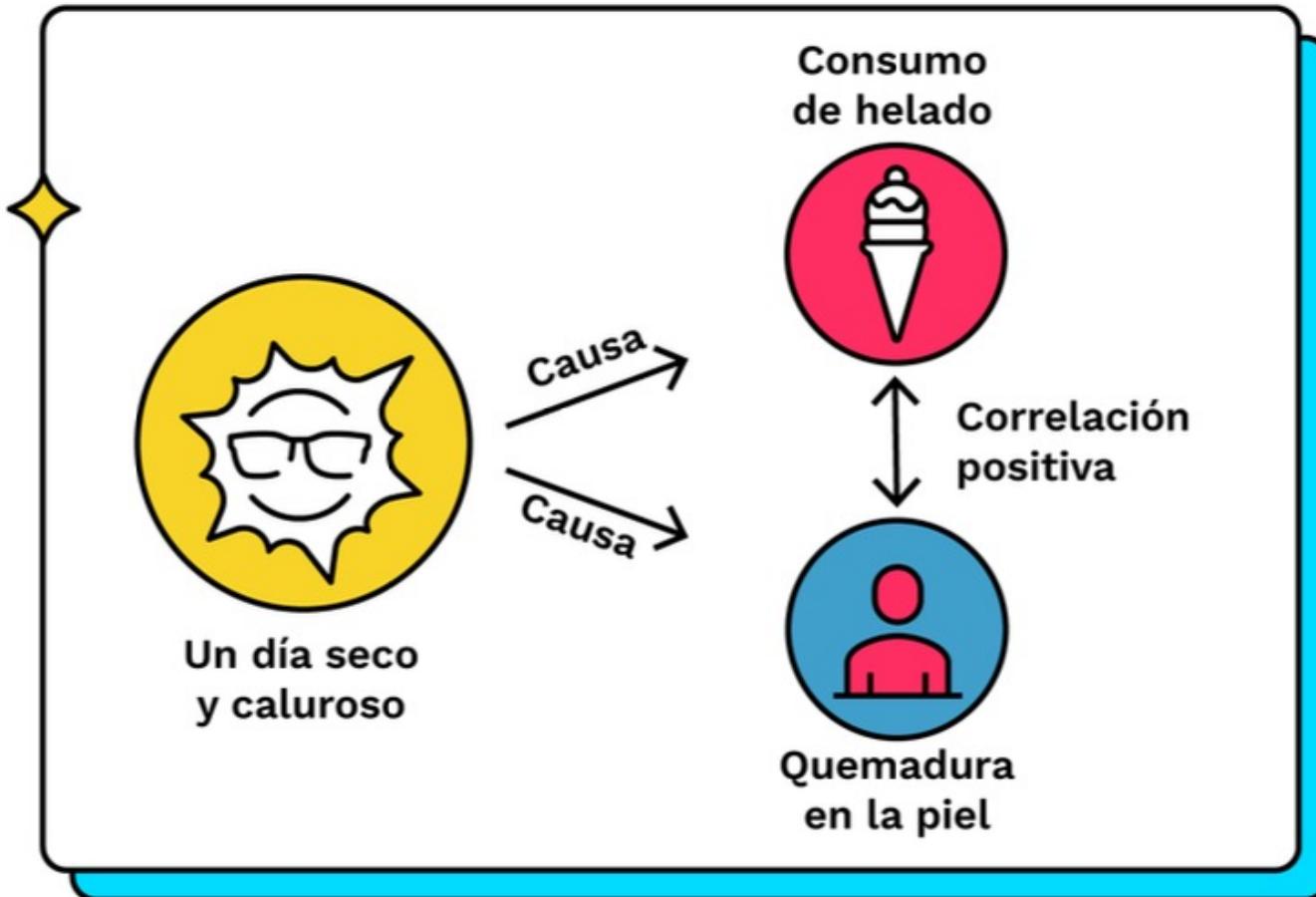
Visualización de datos muestrales

Correlación vs Causalidad

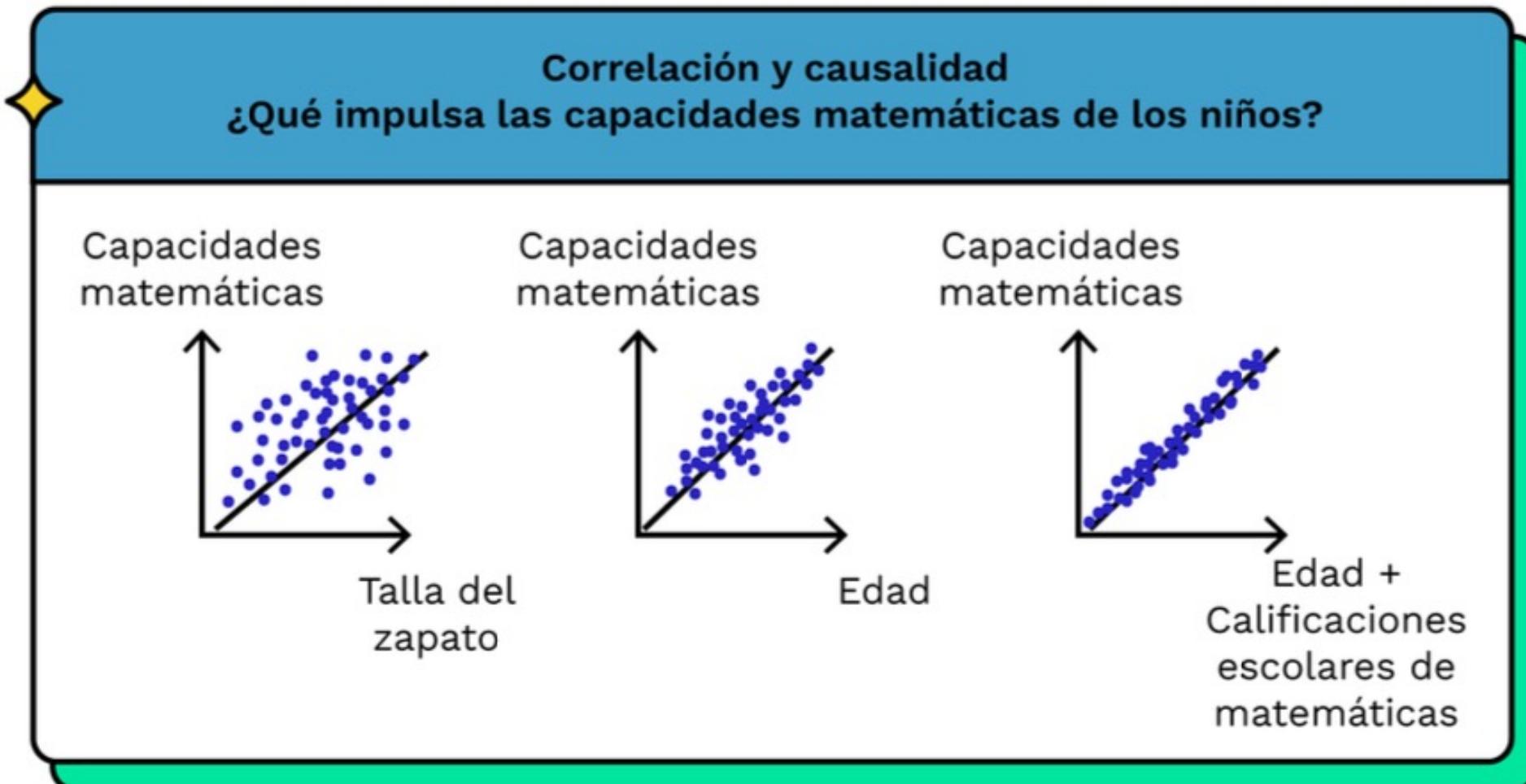
Que la asociación no implica causalidad es quizás la lección más importante que se aprende en una clase de estadística. Hay muchas razones por las que una variable X puede correlacionarse con una variable Y sin tener ningún efecto directo sobre Y. Dos variables pueden moverse en conjunto y con una relación fuerte, pero eso no implica que una esté causando la otra.



Correlación vs Causalidad



Correlación vs Causalidad



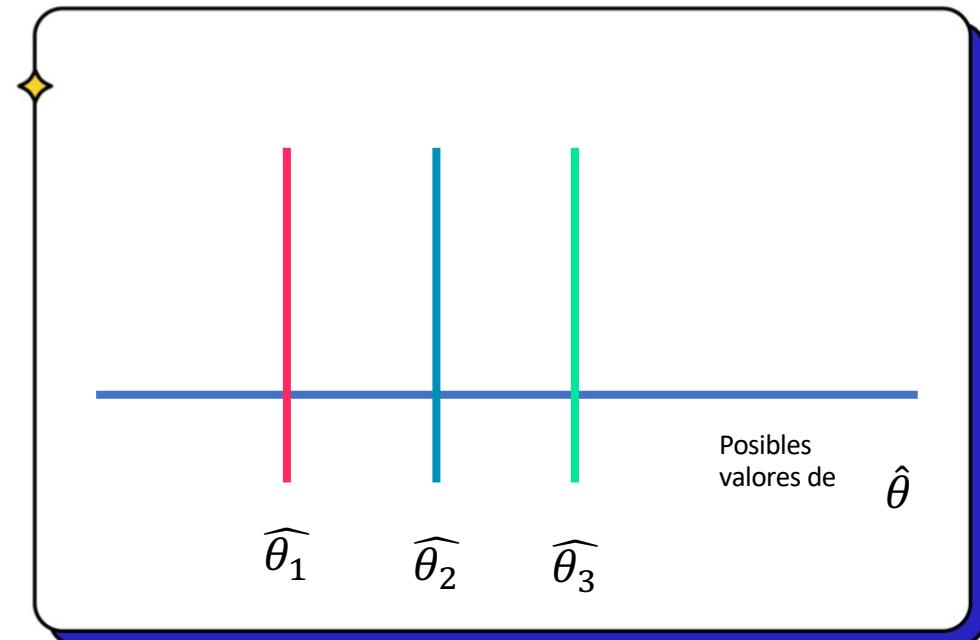
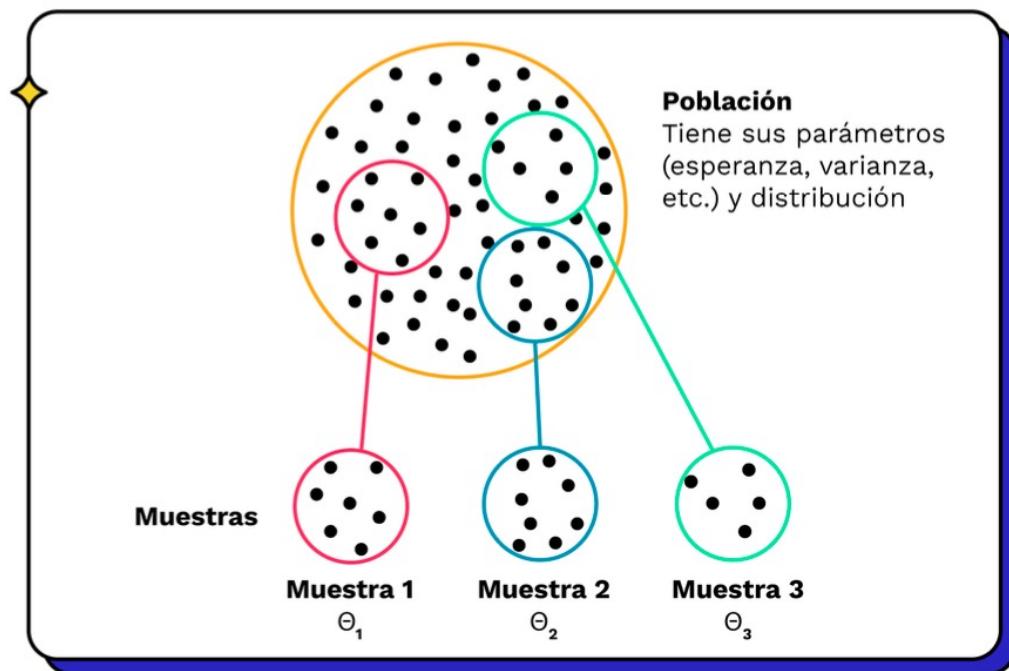
A petri dish containing a DNA gel electrophoresis pattern, with a pipette being used to transfer liquid onto the gel.

II. Inferencia Estadística

Inferencia estadística & Experimentos aleatorios, A/B testing

Estimadores son variables aleatorias

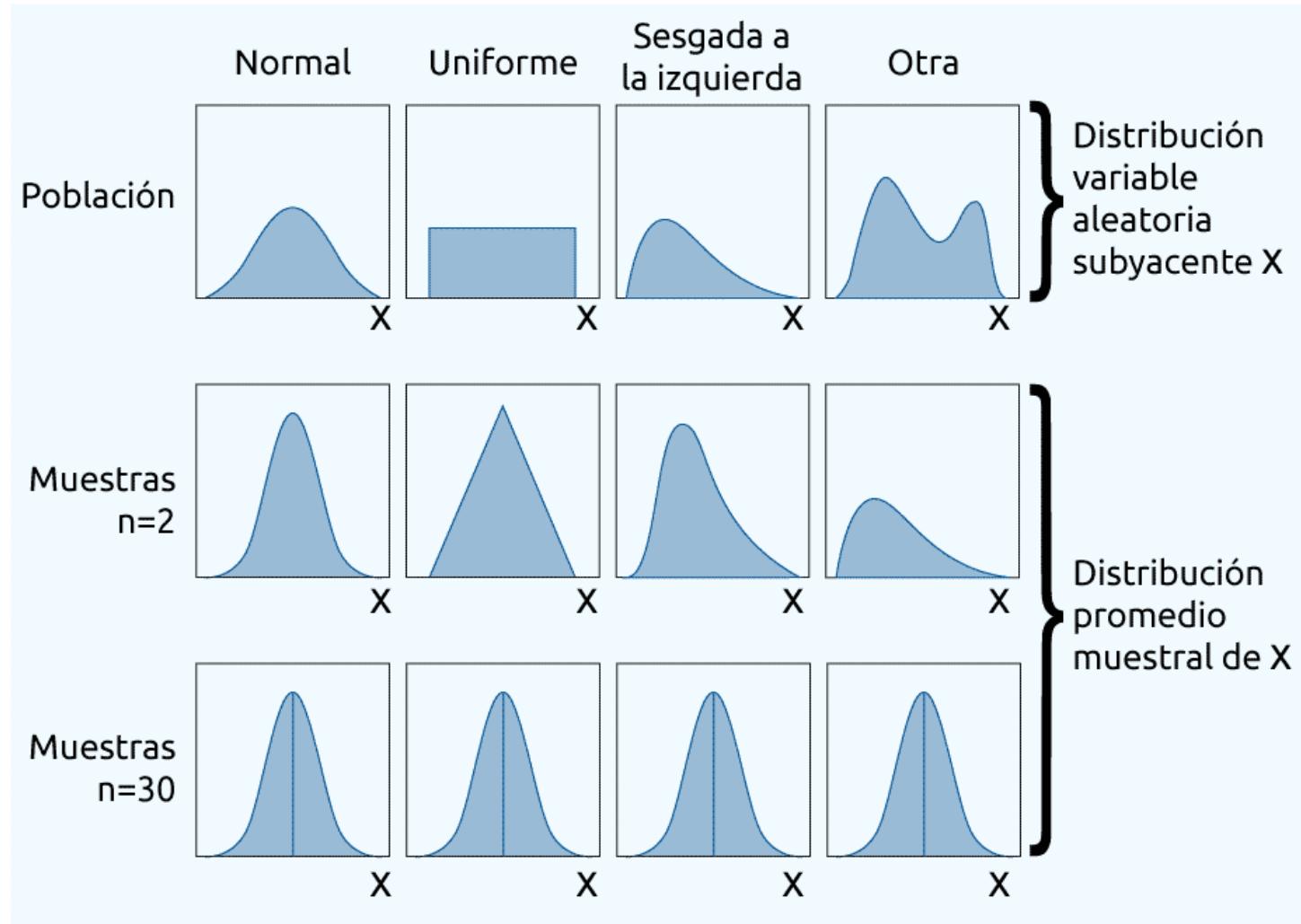
- Tienen esperanza
- Tienen dispersión
- Tienen distribución

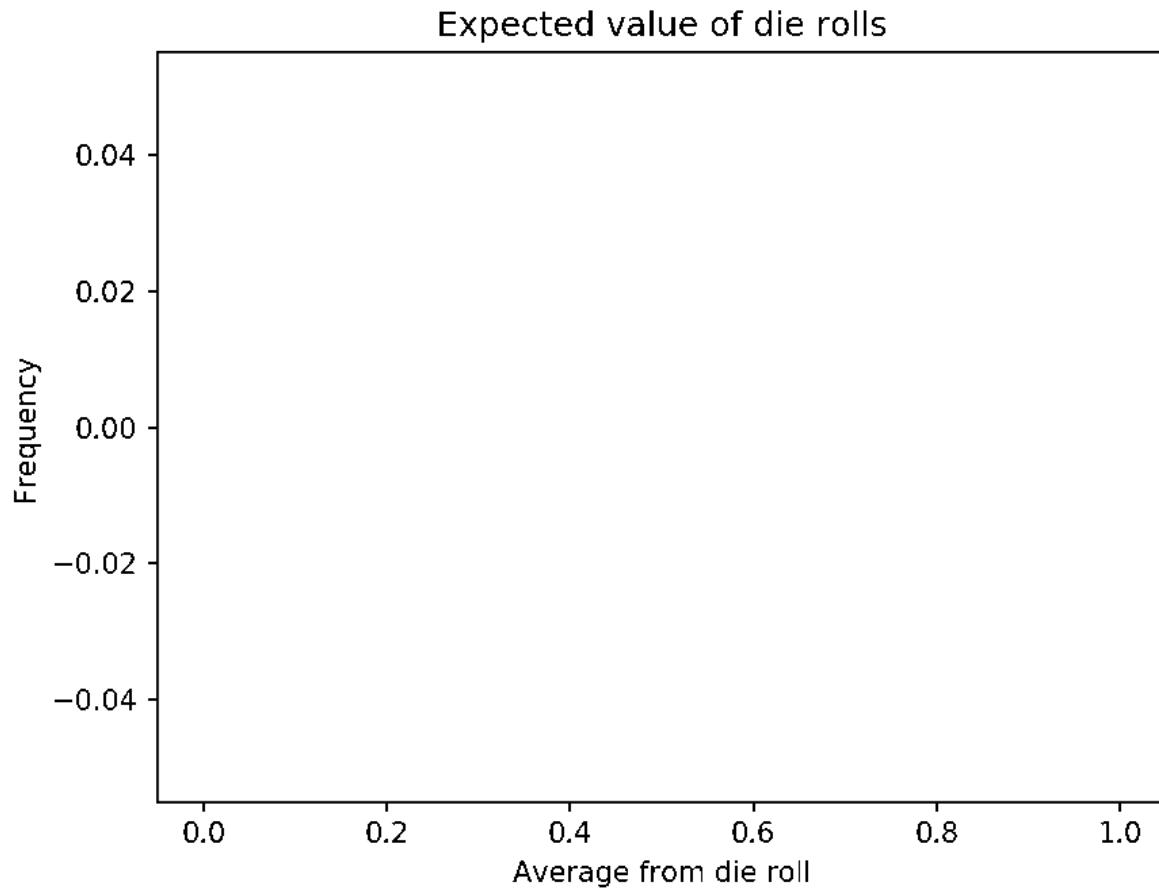


Estadígrafos comunes

Parámetro poblacional	Definición	Estimador más usado	Fórmula estimador con muestra tamaño n
Esperanza	$E[X] = \int x_i p(X=x_i)$ (caso continuo)	Media muestral o promedio	$\bar{X} = \frac{1}{n} \sum_{i=1}^n x_i$
Varianza	$V[X] = E[(X - E[X])^2]$	Varianza muestral	$s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2$
Covarianza	$Cov[X,Y] = E[(X - E[X])(Y - E[Y])]$	Covarianza muestral	$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})(y_i - \bar{Y})$

El promedio es un estimador especial



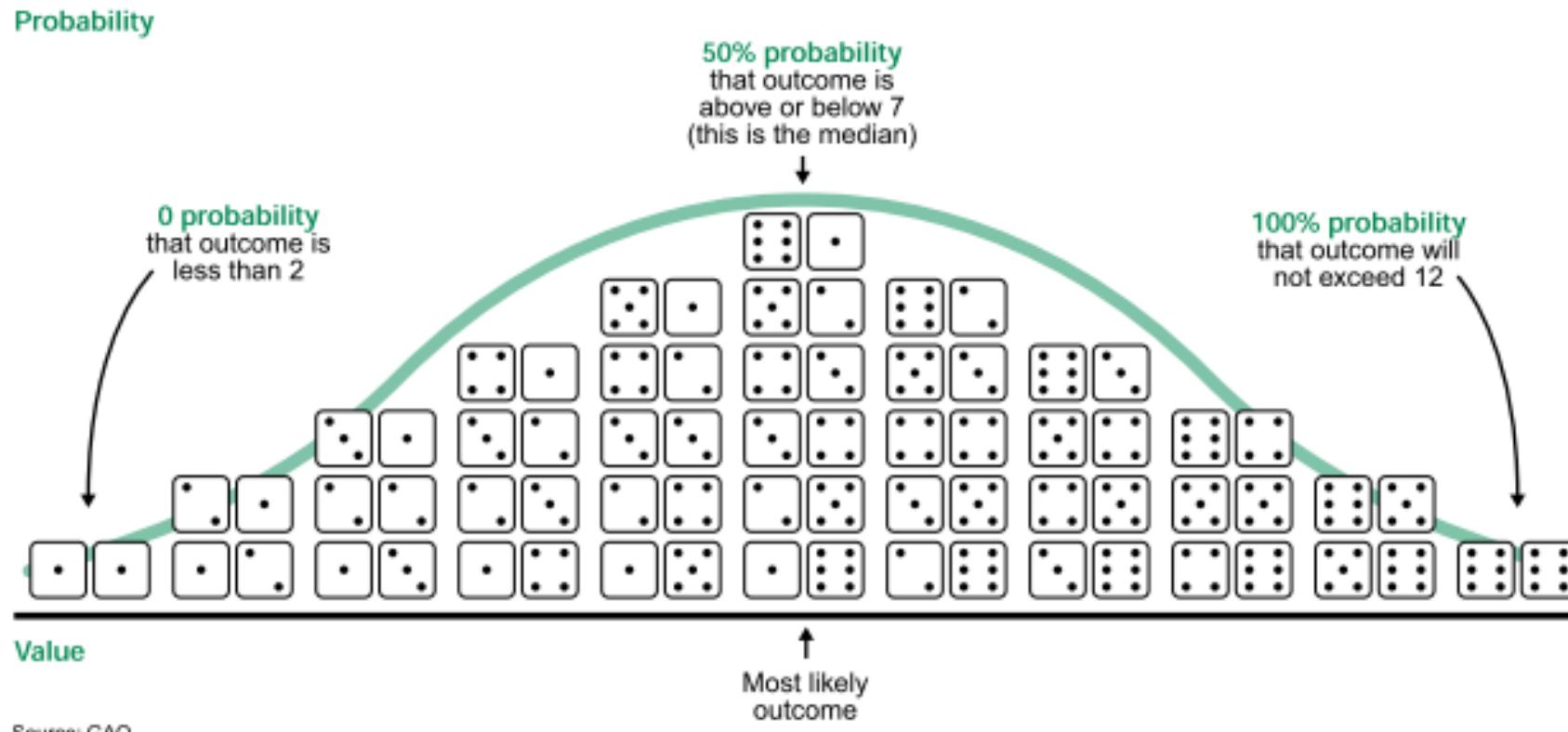


Ejemplo dados

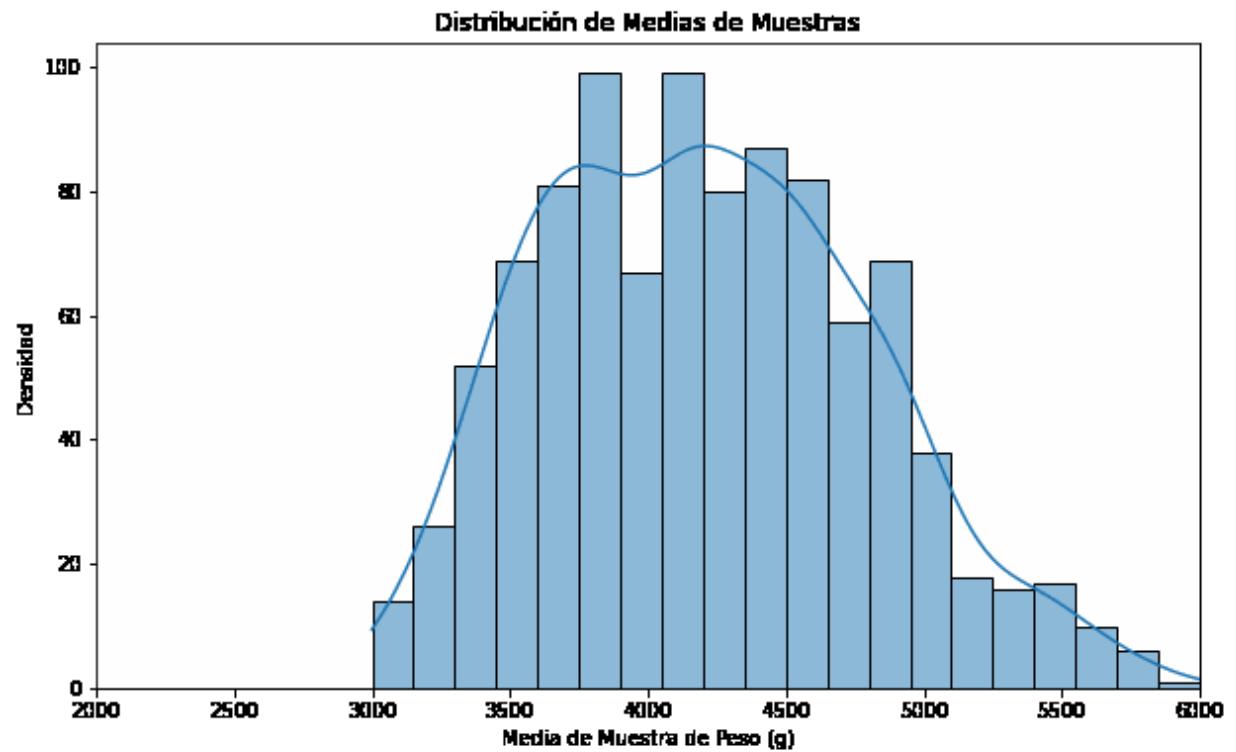
Pruebas de hipótesis y significancia



El promedio es un estimador especial

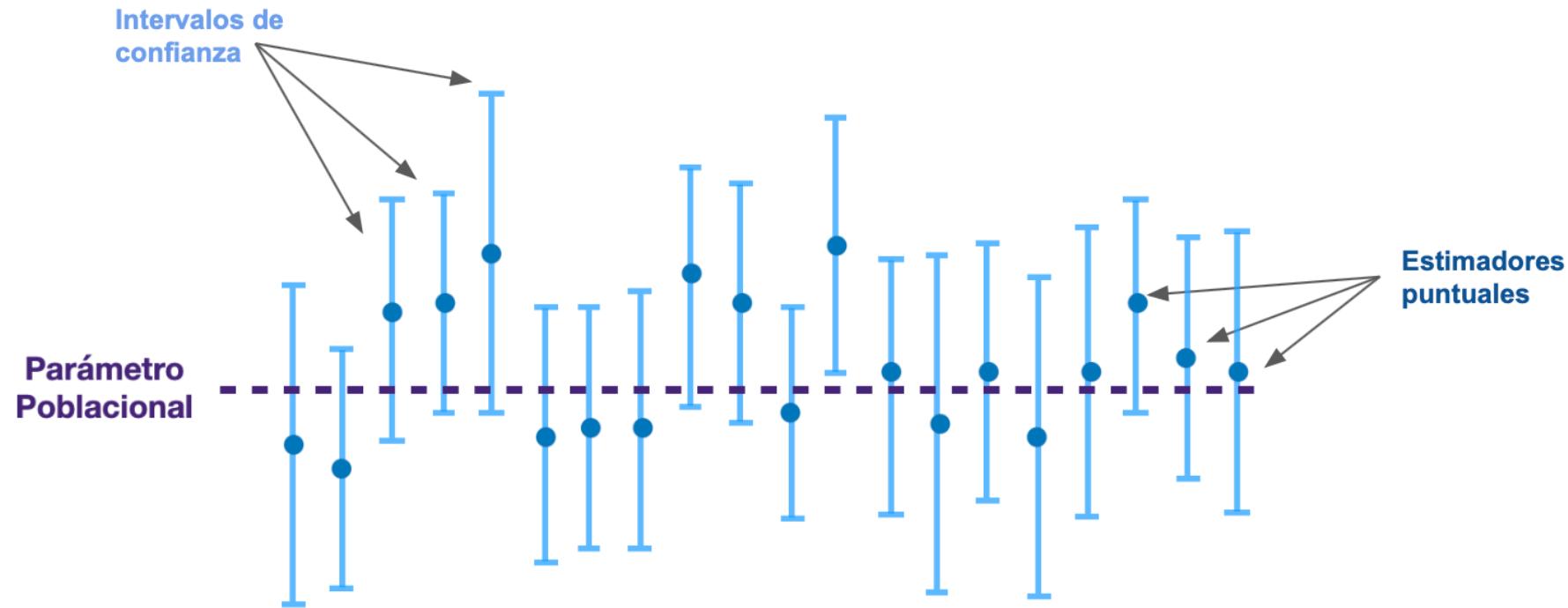


Promedio del
peso de los
pinguinos se
Vuelve normal



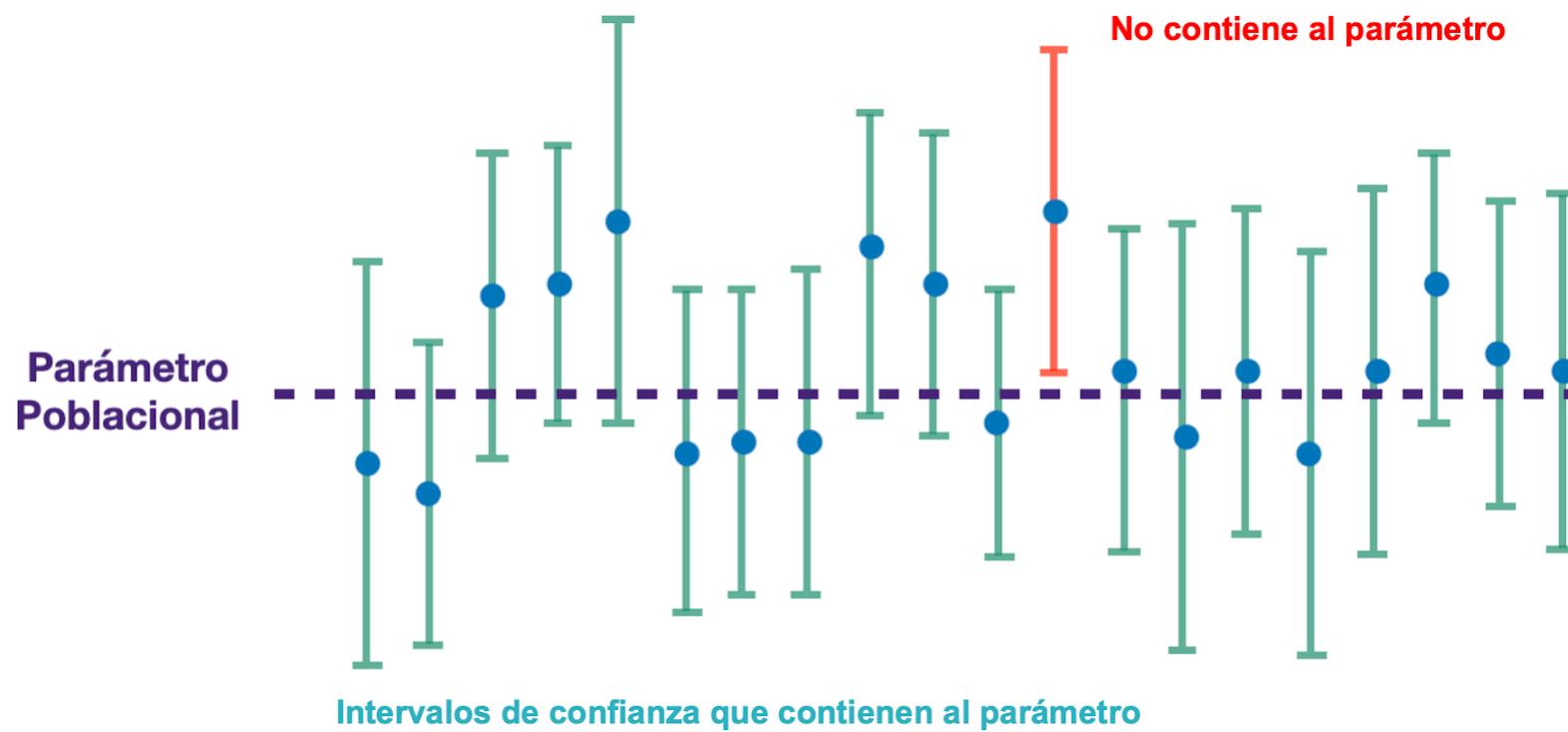
Intervalos de confianza

- Así como el estimador es una variable aleatoria, esto también es cierto para los intervalos de confianza. Por eso también se les llama **intervalos aleatorios**, ya que con diferentes muestras obtendremos un diferente estimador e intervalo.

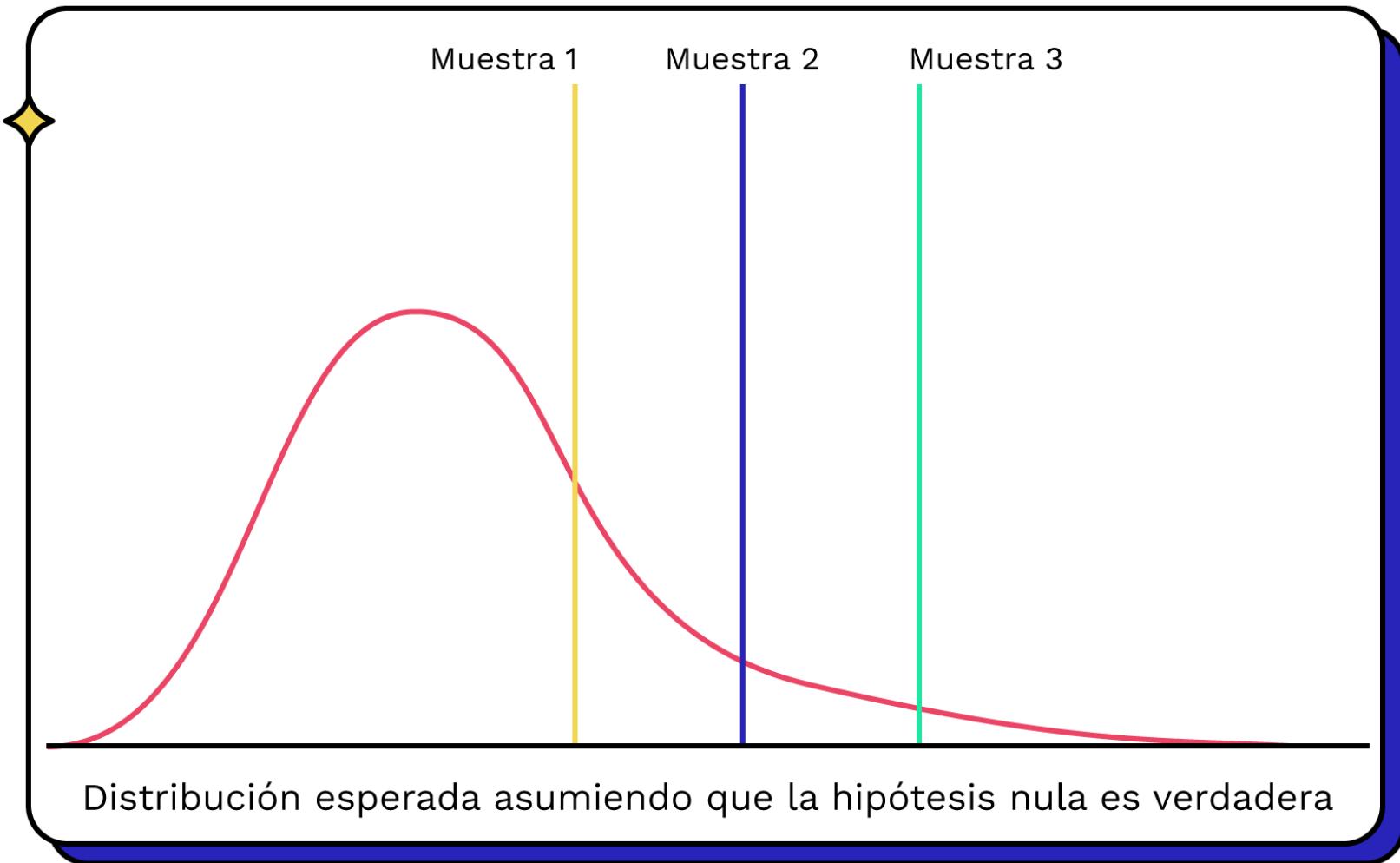


Intervalos de confianza

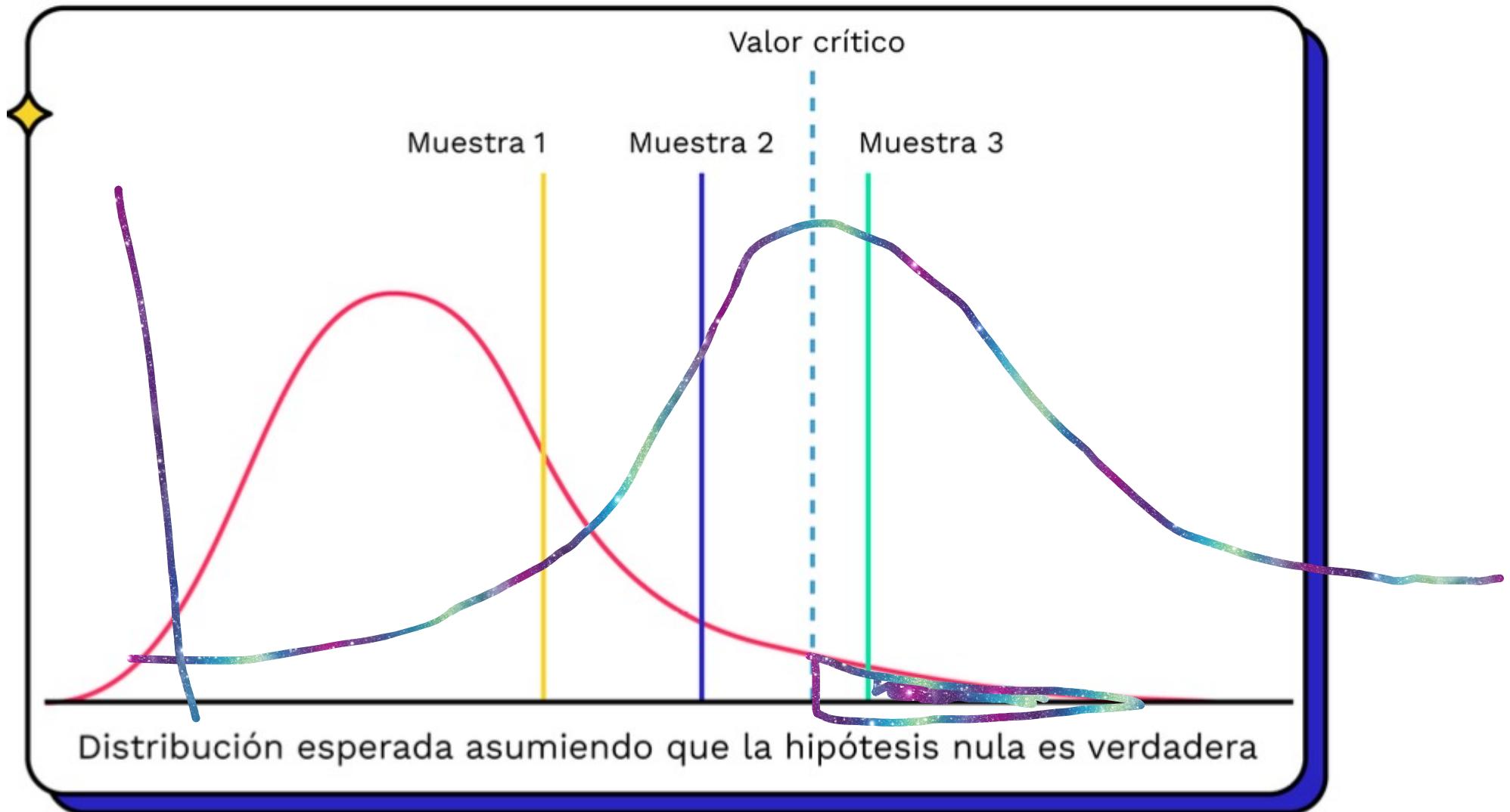
- 95% de confianza:



Prueba de hipótesis



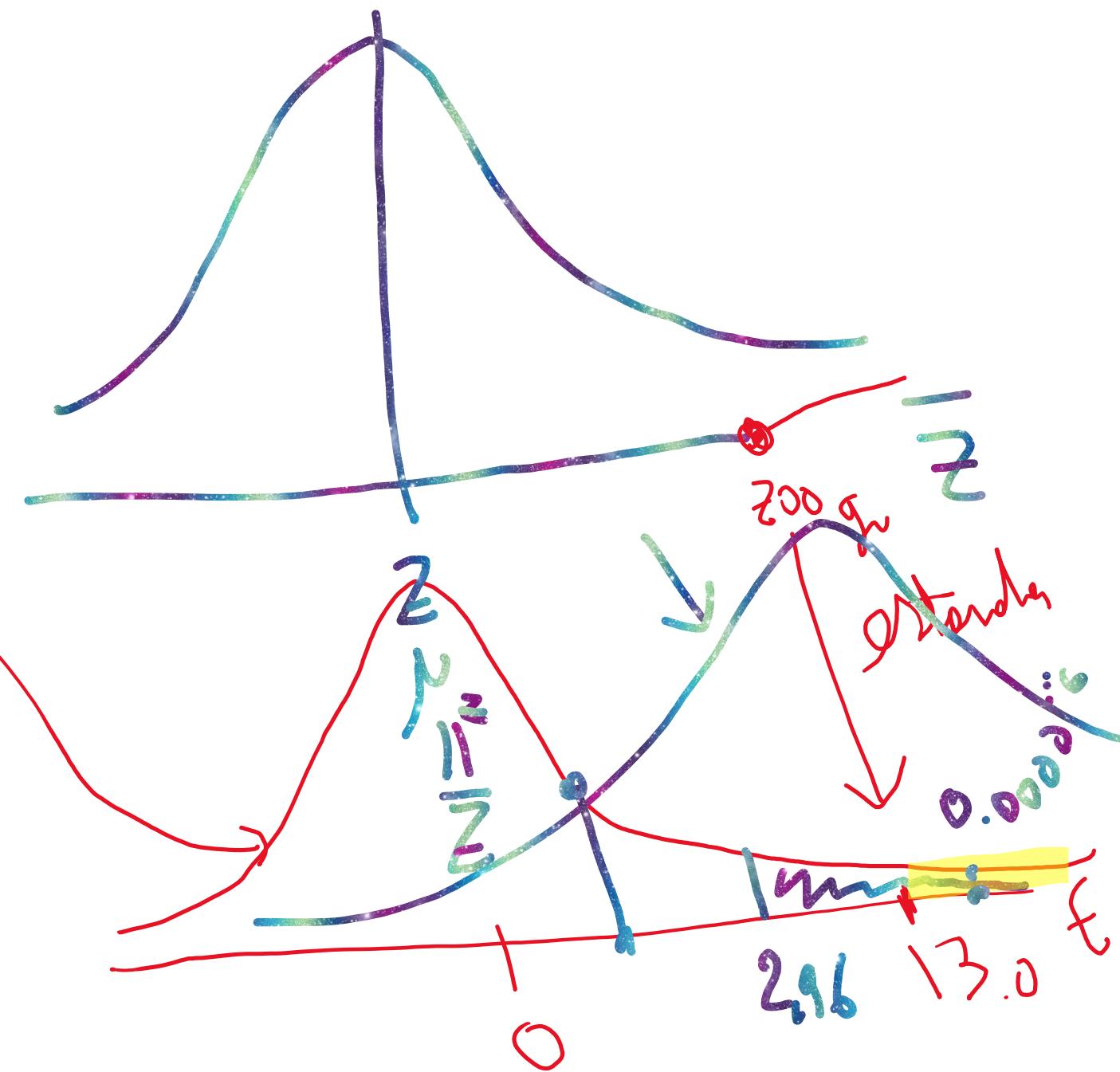
Prueba de hipótesis



$$z = T_{PM} - \varphi_H$$

$$H_0 : z = 0$$

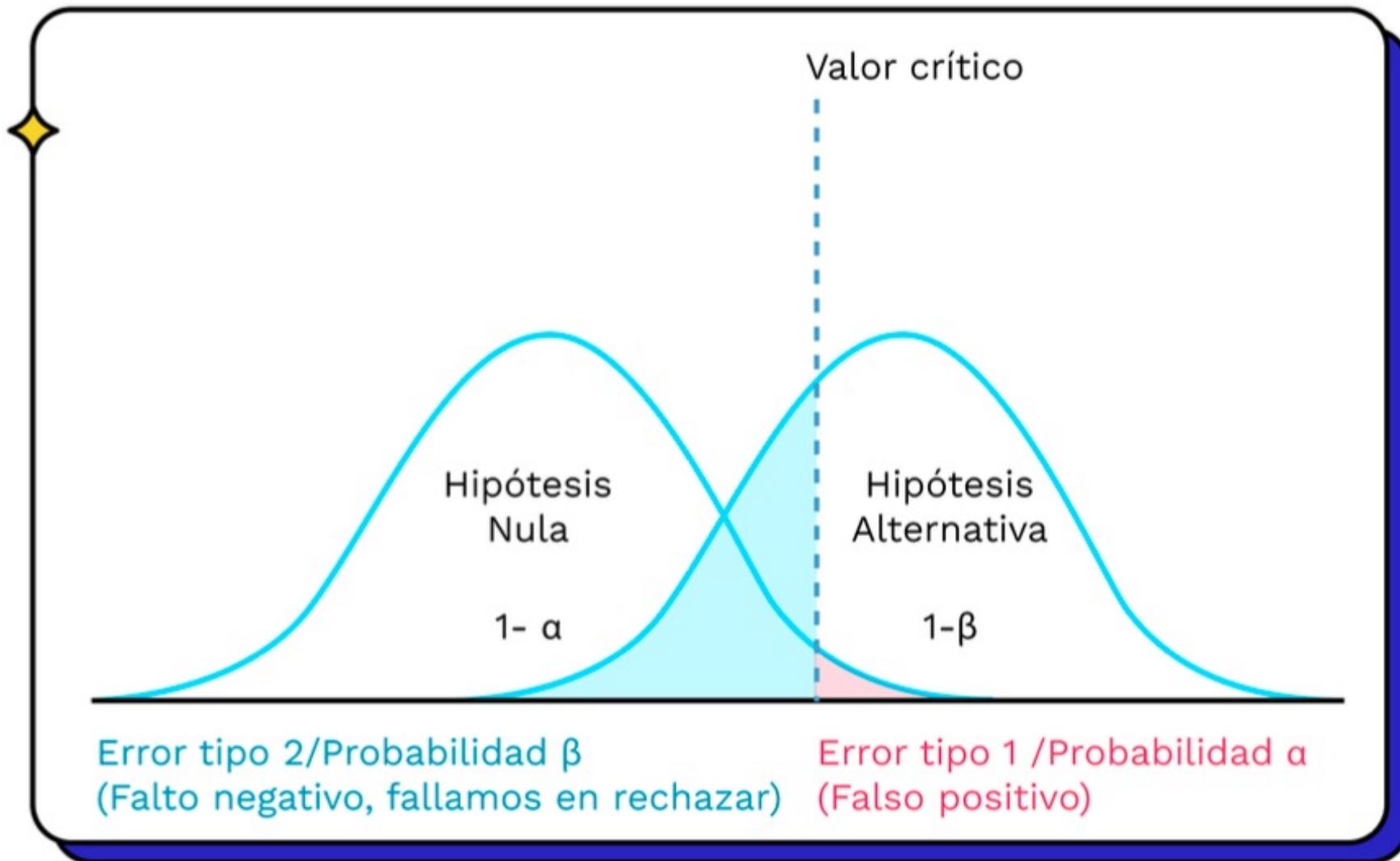
$$H_A : z \neq 0$$



Prueba de hipótesis

Error Tipo 1 (α) o falso positivo	Error tipo 2 (β) o falso negativo
Rechazar la hipótesis nula cuando es cierta	No rechazar una hipótesis nula cuando es falsa.
 Estás embarazado	 No estás embarazada

Prueba de hipótesis



Ejemplo:



Ejemplo:

Por ejemplo, queremos evaluar un proceso industrial en el cual agregamos leche a un contenedor de tetrapack de 1 litro. Tomaremos una muestra aleatoria de los contenedores producidos y evaluaremos su contenido. Por ende, nuestra hipótesis nula es que tiene 1 litro de contenido y la alternativa de que tiene menos de 1 litro.

El valor observado de la desviación estándar es de 0.17 litros y nuestra muestra es de 100 elementos, con un promedio de 0.97

$$H_0: \mu=1 \text{ o } H_0: \mu<1$$

Consideraremos que nuestra prueba es al 95% de confianza, eso quiere decir que aceptaremos un error tipo 1 de 5%.

El valor crítico al cual rechazaríamos esta nula es si $t < -1.64$, ya que ese es el valor al cual la probabilidad de que t sea mayor es 5% o menos.



Test A/B

Una prueba A/B es un experimento estadístico con dos grupos, en los cuales se establecen dos tratamientos, productos o procesos para identificar cual de los dos es superior.

Ejemplos:

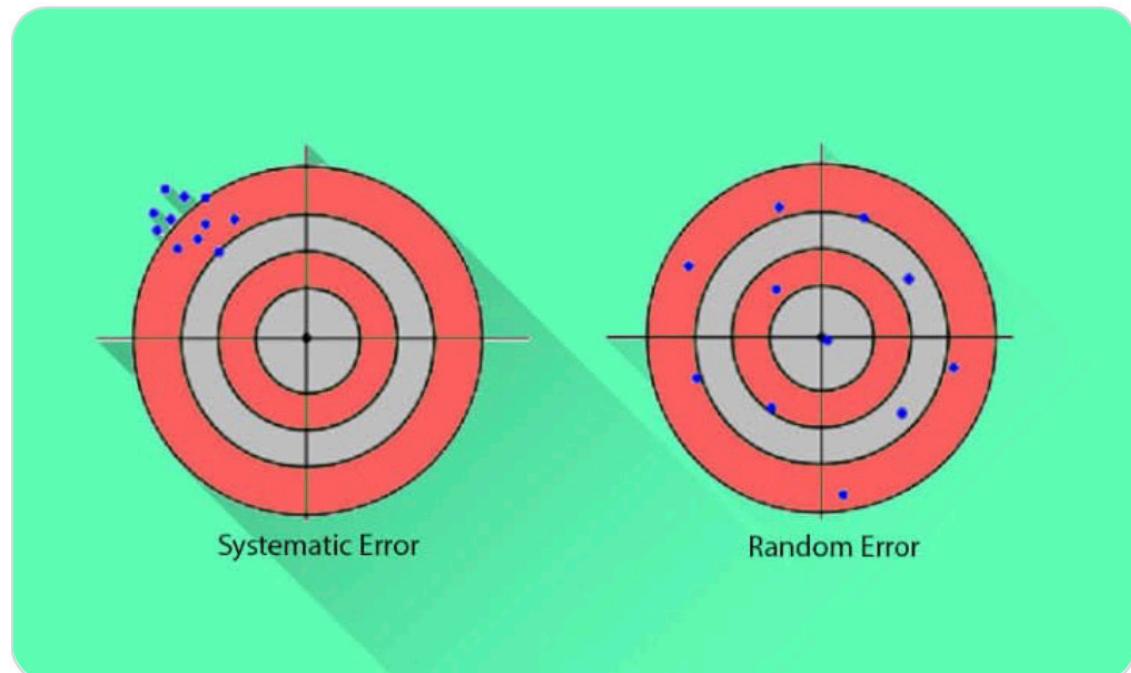
- Queremos testear si un sitio web con cierta estructura produce más clics que otro.
- Un empaque de producto tiene más éxito que otro.
- Un medicamento produce un efecto deseado o no.

Lo que hacemos para comprarlos, es mediante la construcción **de una prueba estadística en la que aplicamos una hipótesis de si ambos grupos son o no diferentes** (dos colas) o si el tratamiento genera mejoras con respecto al estándar anterior (1 cola).

Importancia de la asignación aleatoria

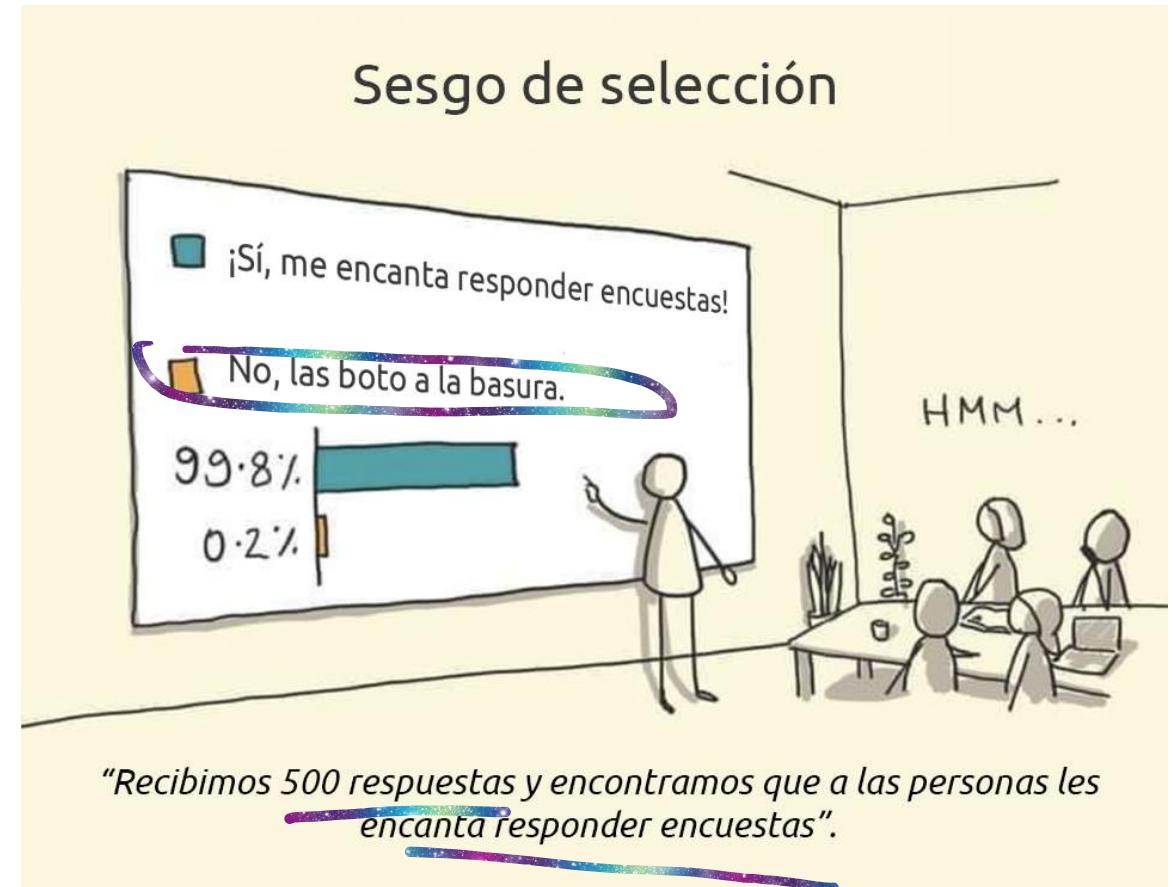
Para que la prueba pueda capturar el impacto o efecto del **tratamiento es clave que ambos grupos sean asignados aleatoriamente**

Para que no existan posibles **sesgos sistemáticos**



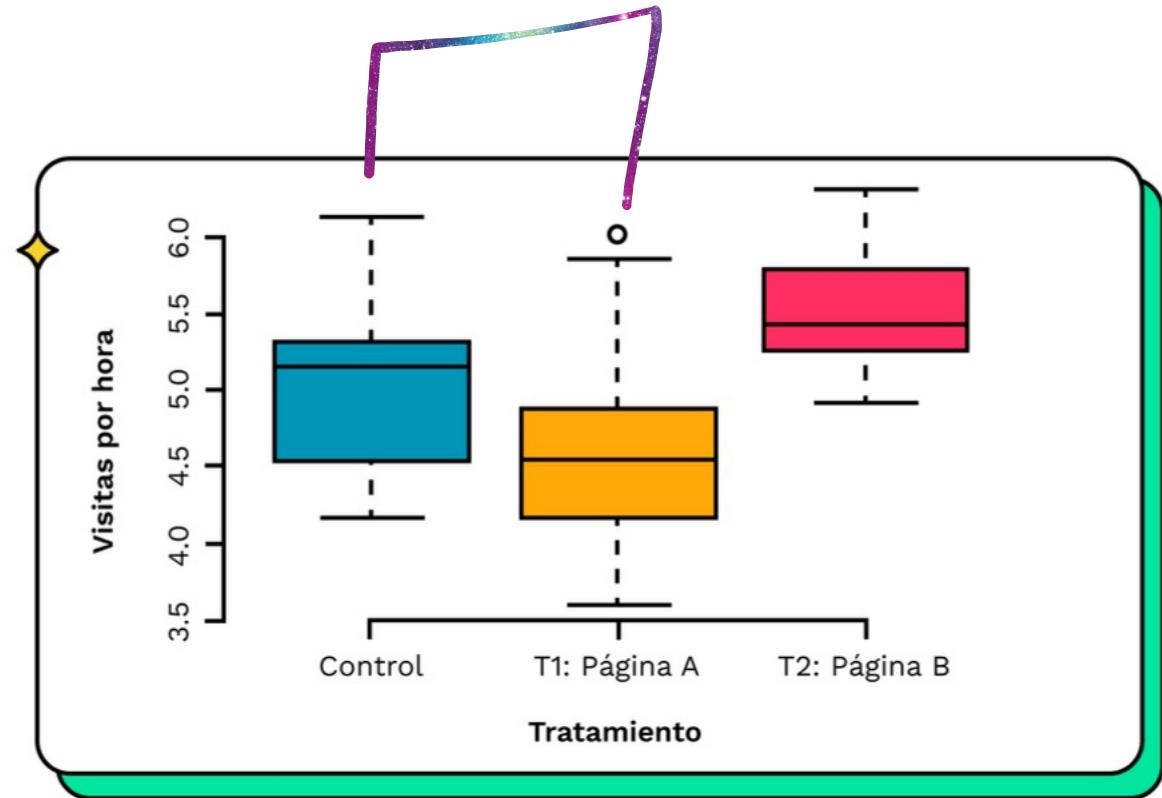
Sesgo de selección

- Este corresponde a un **sesgo sistemático** en los resultados que se desprende de la manera en la cual las **observaciones que participan en el estudio son seleccionadas**.
- Asegurar que no exista **sesgo de selección**, es decir que otras variables expliquen las diferencias entre tratados y controles.



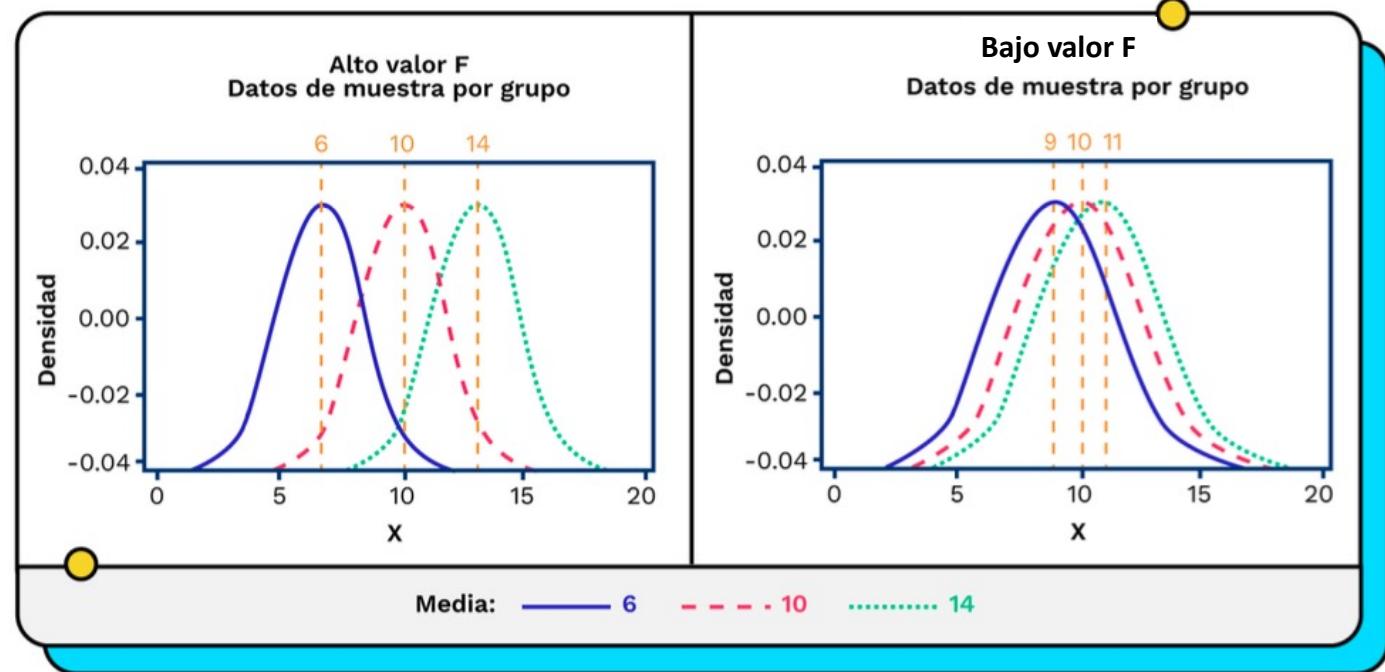
Diferencia de media

- Compara la media de dos grupos
- Evalúa si la diferencia entre ambas medias es estadísticamente significativa
- Se utiliza para analizar datos numéricos agrupados



ANOVA (Análisis de Varianza)

- Se utiliza para comparar las medias de dos grupos o más.
- La hipótesis nula es que no hay diferencia significativa entre las medias de los grupos.
- Se calcula la varianza entre los grupos y la varianza dentro de los grupos para determinar si la diferencia observada entre los grupos es significativa.



$H_0: \mu_1 = \mu_2 = \mu_3$ Vs $H_0: \mu_1 \neq \mu_2 \neq \mu_3$

- ANOVA produce un valor F y un valor p.
- Si el valor p es menor que el nivel de significancia elegido, se puede rechazar la hipótesis nula.
- Se concluye que hay diferencias significativas entre las medias de los grupos.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios
Intra grupos	$SS_{within} = \sum_{j=1}^k \sum_{i=1}^n (x_i - \bar{X}_j)^2$	n-k	$MS_{within} = \frac{SS_{within}}{n-k}$
Entre grupos Diferencia entre la media de cada grupo y en total	$SS_{between} = \sum_{j=1}^k (\bar{X}_j - \bar{X})^2$	k-1	$MS_{between} = \frac{SS_{between}}{n-k}$
Covarianza	$SS_{total} = \sum_{i=1}^n (x_i - \bar{X})^2$	n-1	$MS_{total} = \frac{SS_{total}}{n-1}$

Y el F observado es:

$$F^{\text{observado}} = \frac{MS_{between}}{MS_{within}}$$



Evaluación del taller

- A cada bloque del taller le acompaña un notebook de trabajo y la actividad.
- Se puede trabajar en grupos de hasta 3 o individualmente.
- Deben mandarme un mail con su notebook desarrollado.

¿Consultas?

