Estadística para el análisis de expresión génica a gran escala

Lic Ignacio Alcántara (*MSc*)

Ud. Bioestadística- Departamento de Salud Pública

Facultad de Veterinaria - UdelaR. 📧 ignacio.alcantara@fvet.edu.uy

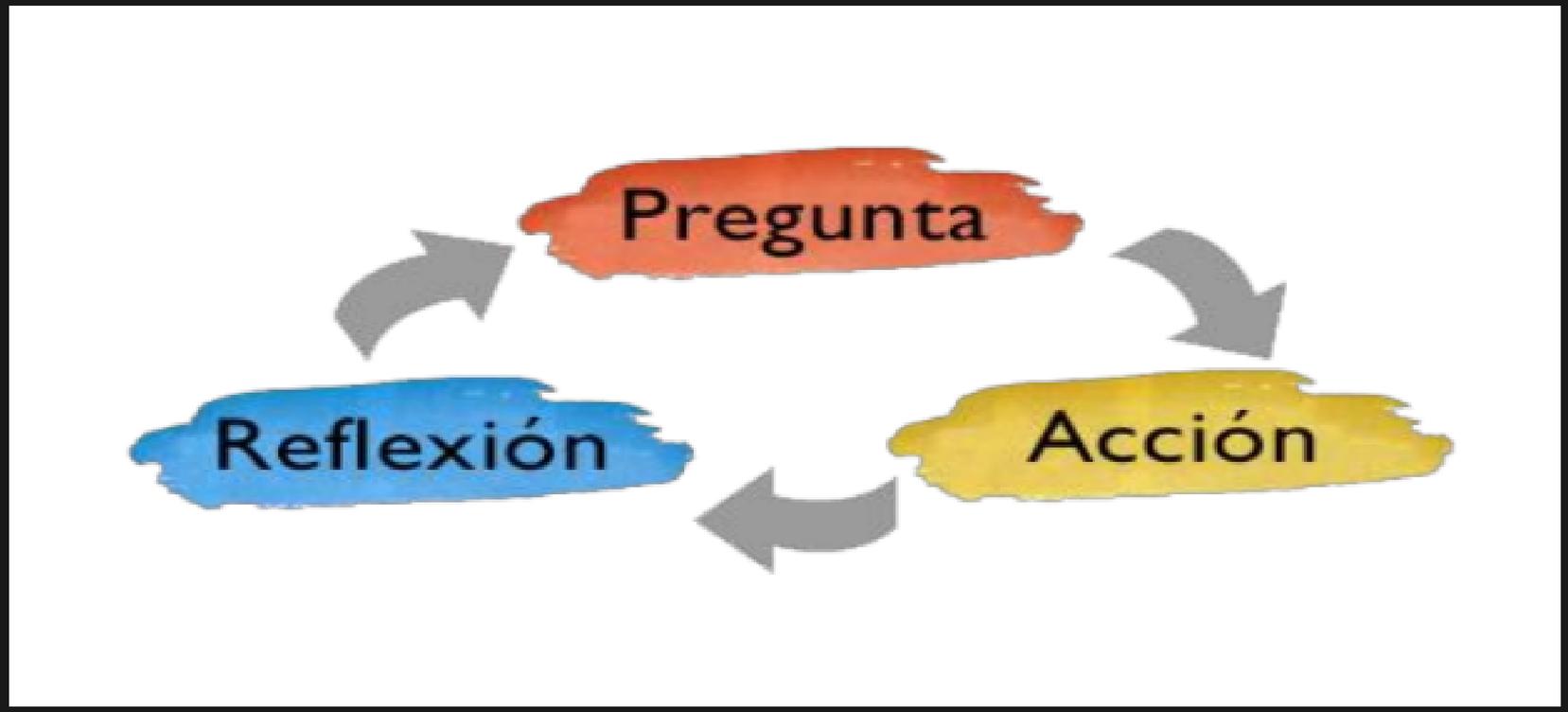
Curso Genómica - Facultad de Ciencias

2024-10-08

Contenido de la clase 🤘

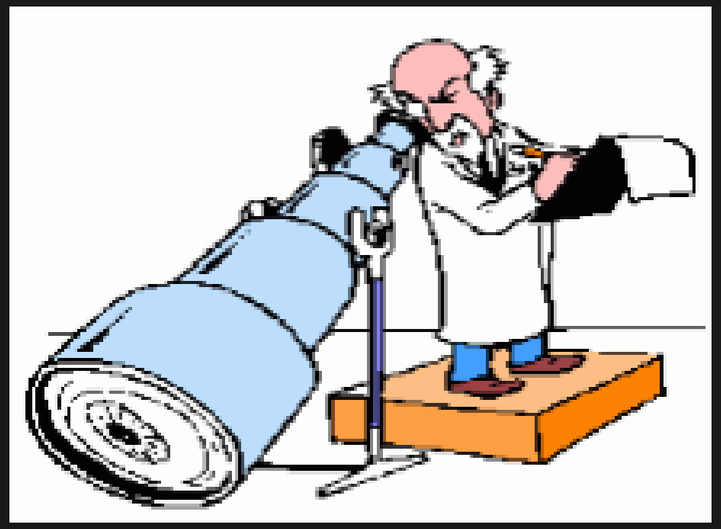
- Repaso de Conceptos Estadísticos
- Tipo de datos Transcriptómicos
- Estadística y sus ramas
- Análisis exploratorio de datos
- Modelos, Prueba de Hipótesis y P-Valor
- Significancia: estadística vs biológica
- Buenas Prácticas Y Reproducibilidad

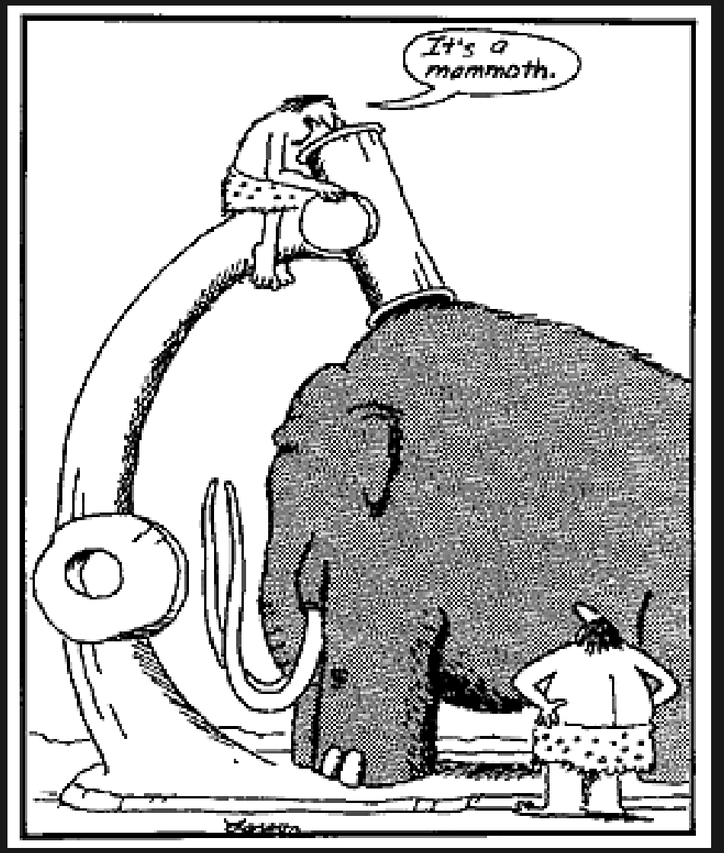
¿Cómo sabemos lo que sabemos?



¿Cómo sabemos lo que sabemos?

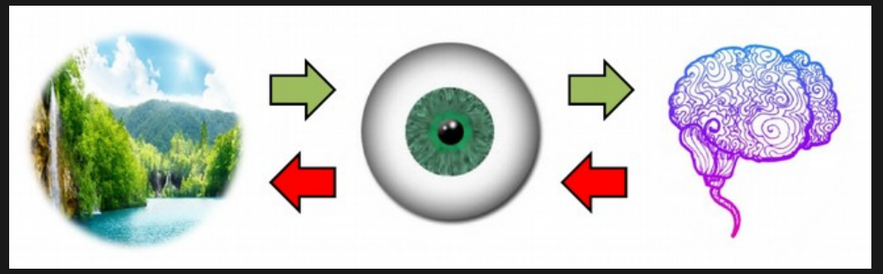
Las ciencias naturales surgen de Observar





¿Cómo sabemos lo que sabemos?

- El mundo es uno solo, pero las **visiones** son diversas
- La correspondencia entre lo que **decimos** y lo que **ocurre** a nuestro al rededor es lo que llamamos **Realidad**
- La realidad que percibimos esta limitada por el **marco filosófico**

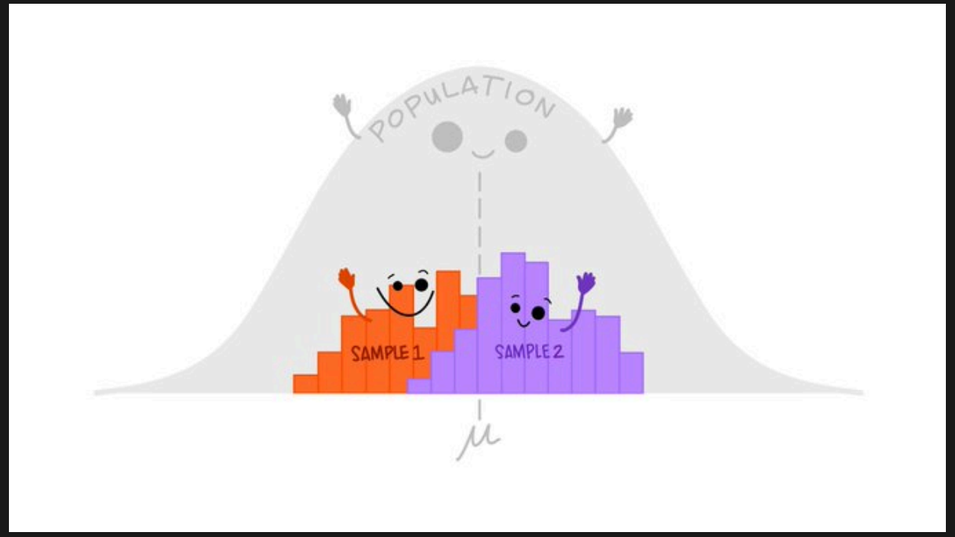


Población y Muestra

- **Población (tamaño N)**: **Colección completa** de elementos con alguna característica de interés observable, *acotada en un tiempo y en un espacio* determinado.
- Es el conjunto entero al que se desea describir, y en el que se pretende tengan validez nuestras conclusiones.
- **Censo**: Estudio de **todos los elementos** de una población
- **Muestra (tamaño n)**: **Subconjunto de elementos** de una población seleccionados de acuerdo a un plan de acción previamente establecido (**muestreo**), que permita obtener conclusiones extensivas a toda la población.

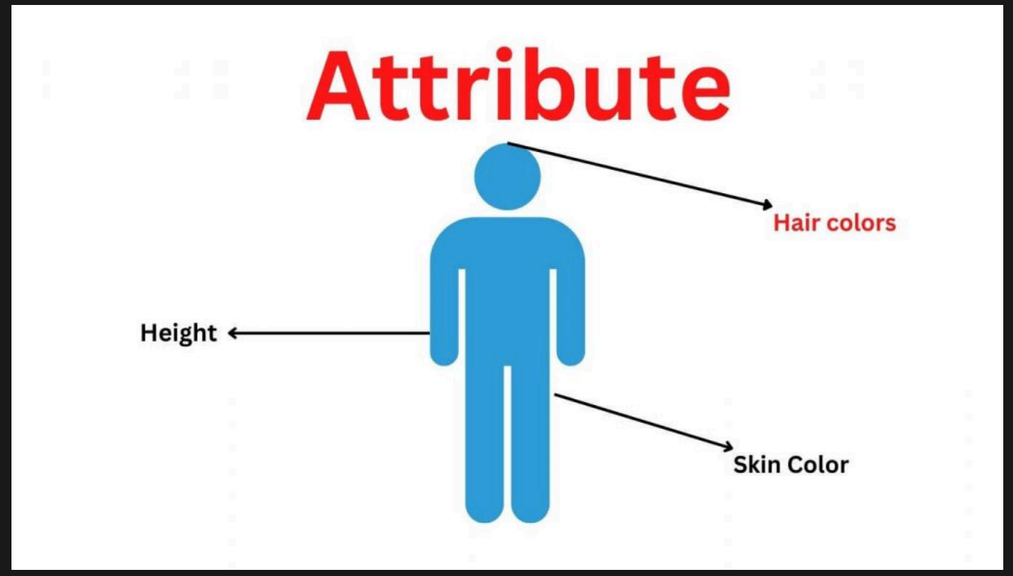
Muestreo

- Técnica que conduce a la obtención de una muestra **representativa** de la población.
- El objetivo es asegurar que cada observación en la población tiene la misma probabilidad de ser incluida en la muestra, y su elección es por azar.
- El alcance y validez de nuestras conclusiones estará determinado por la adecuada elección de la muestra.



Variables

- Una **variable** es una **característica o atributo** de interés medida/identificada sobre cada **elemento individual** de una población o muestra.
- El **valor** que asume la variable en cada **elemento** de la población es un **Dato**



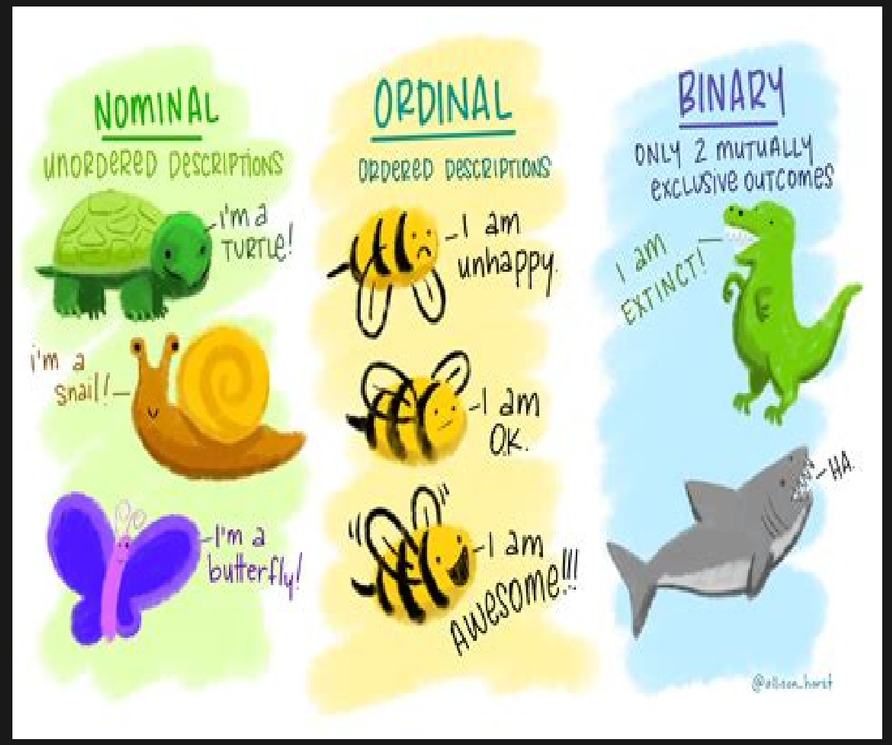
Variables: tipos y escalas

- Cualitativas o categóricas

El dato se asigna a una de las categorías disponibles

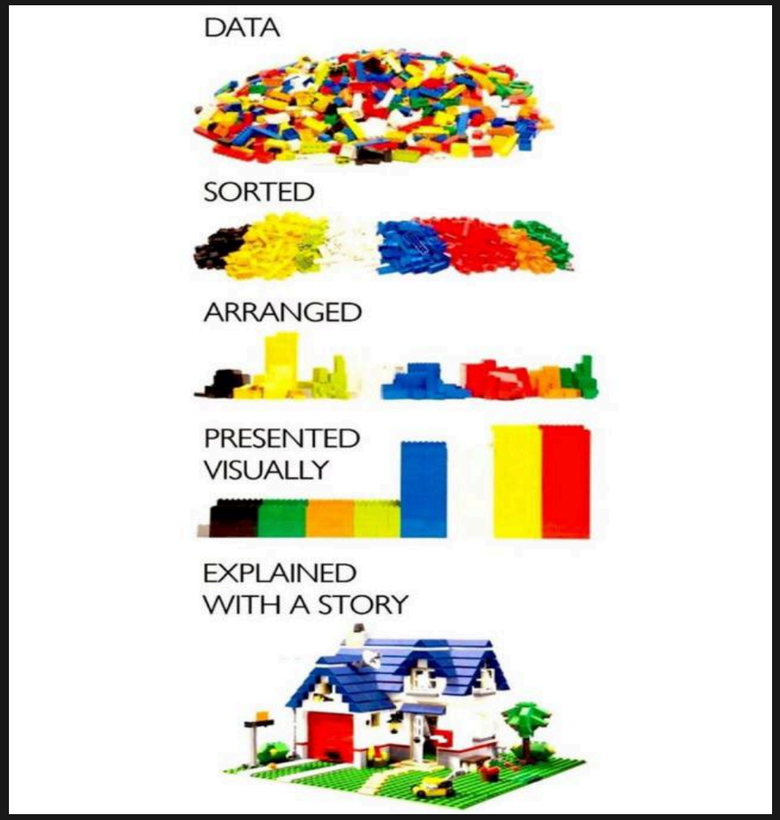
- Cuantitativas o numéricas

Son cuantificables: surgen de contar o medir



Ordenar y resumir los datos

- Las mediciones que no han sido organizadas, procesadas o manejadas de alguna manera se les llama *datos crudos*



Tipos de datos Transcriptómicos

Microarreglos y RNA-Seq

¿Qué tipos de datos son?

- Microarreglos (Microarrays): miden la cantidad de expresión de genes predefinidos mediante la hibridación de secuencias conocidas a sondas en una matriz.
  - Datos relativos, predefinidos (solo para genes con sondas específicas), basados en intensidad de fluorescencia, con una resolución limitada.

- RNA-seq (Secuenciación de RNA): utiliza la secuenciación de próxima generación (NGS) para cuantificar la cantidad de transcritos en una muestra de manera más precisa y sin un set predefinido de genes.
  - Datos de recuento (conteo) absoluto de lecturas, no predefinidos (permite detectar nuevos transcritos), mayor resolución y sensibilidad.

Proceso de obtención del *Dato*

1. Calidad de Lectura

1. Alineamiento

1. Cuantificación de expresión

1. **Normalización:** Asegurar datos comparables y consistentes para el análisis exploratorio o de expresión diferencial, mientras se limitan los falsos positivos o negativos

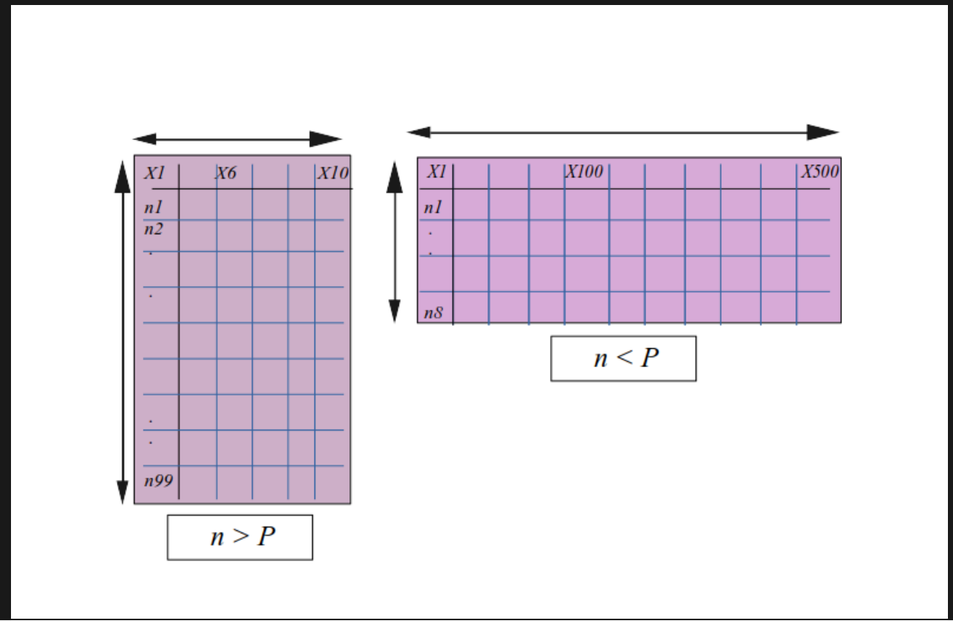
Normalización

- **Intra-Muestra:** Los genes más largos suelen tener más lecturas mapeadas que los genes más cortos a un mismo nivel de expresión.
- **Entre muestras (mismo set de datos):** RNA-seq es una medida relativa de la abundancia de transcriptos. El tamaño de la población de transcriptos en su conjunto afecta los niveles relativos de transcripción.
- **Entre set de datos:** Cuando se integran datos de RNA-seq de múltiples estudios independientes. Estos conjuntos de datos generalmente se secuencian en diferentes momentos, métodos variables, instalaciones y contienen otros factores experimentales. Esto da lugar a un efecto de lote (**batch effect**).
- **El efecto de lote suele ser la mayor fuente de expresión diferencial cuando se combinan los datos. Puede enmascarar cualquier diferencia biológica real y llevar a conclusiones incorrectas.**

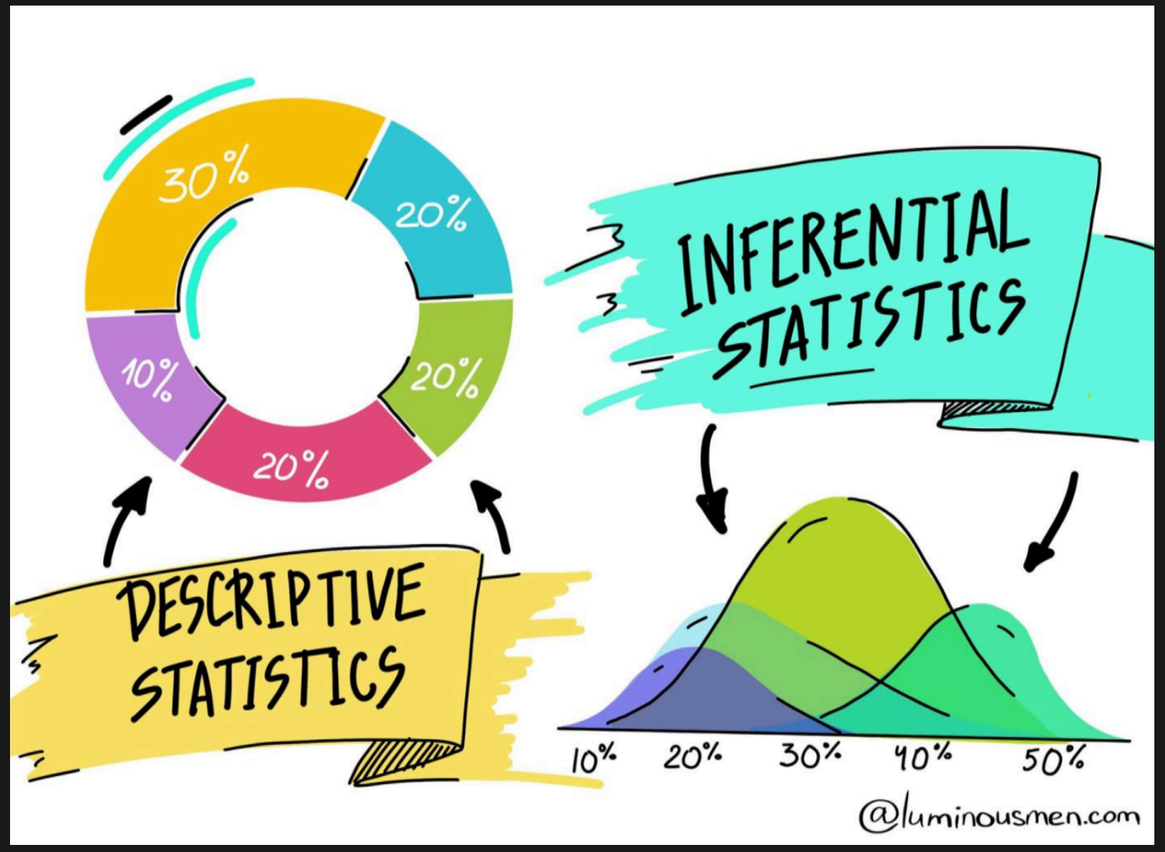
Dimensión de los datos

Cuando p supera a n…

…since the standard fitting algorithms all require p < n; in fact the usual rule of thumb is that there be five or ten times as many samples as variables. But here we consider situations with n around 50 or 100, while p typically varies between 1,000 and 20,000 (Hastie and Tibshirani 2003).



Estadística



Descriptivo vs Inferencial

Estadística Descriptiva:

- **Ordenar** y procesar un **conjunto de observaciones** de tal manera que permita extraer información para su **comunicación**.

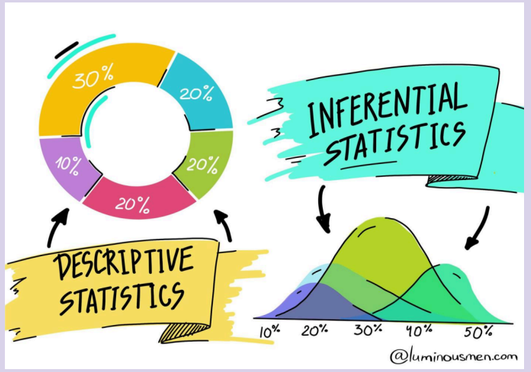
Estadística Inferencial:

- Toma **decisiones**, hace **estimaciones y/o predicciones** acerca de una población de la que se extrajo una muestra.

Estadística Descriptiva

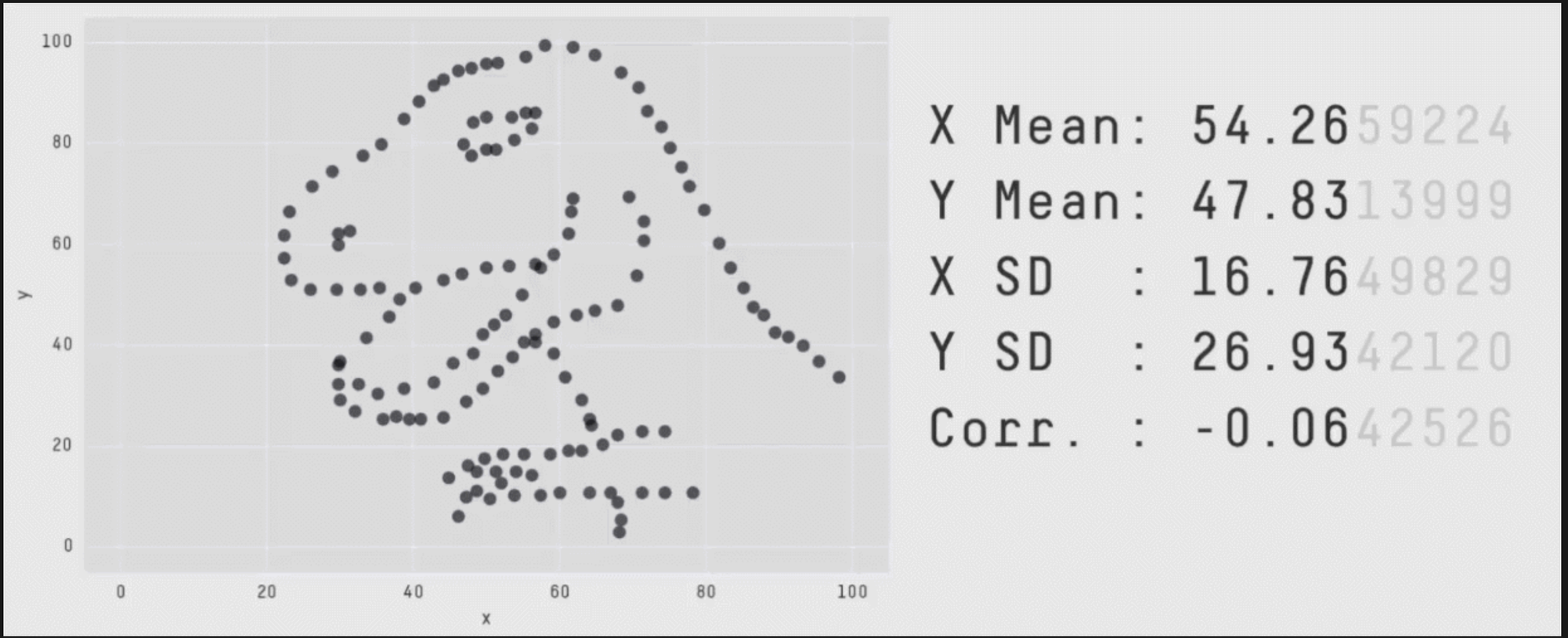
Ayuda a responder

- ¿Qué tipo de números obtuve y qué tan dispersos son?
- ¿Cómo están distribuidos los valores de las variables explicativas y la de respuesta?
- ¿Hay valores extremos, sospechosos y/o inexactos?
- De las variables disponibles, ¿cuáles tienen sentido para el análisis?



Visualización vs Cálculos

- Los cálculos numéricos son exactos mientras que los gráficos son aproximados



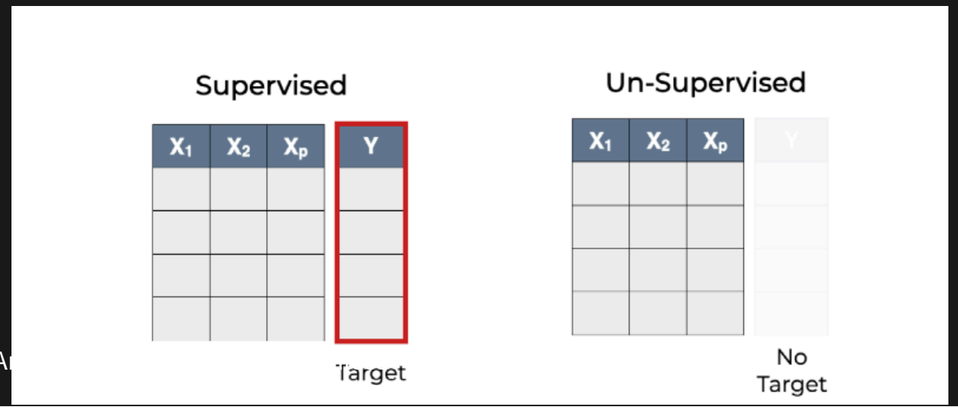
Análisis Supervisado y No Supervisado

- Análisis Supervisado:

Se utiliza cuando se tiene un conjunto de datos etiquetado, es decir, se conoce la variable de respuesta. El objetivo es aprender una relación entre las variables predictoras (x) y la respuesta (y) para predecir o clasificar nuevos datos. Ejemplos: Regresión lineal, clasificación (árboles de decisión, SVM, redes neuronales).

- Análisis No Supervisado:

No se tiene una variable de respuesta conocida; el objetivo es identificar grupos, patrones o estructuras en los datos. Ejemplos: Clustering, análisis de componentes principales (PCA), reducción de dimensionalidad.



Resumiendo las Correspondencias:

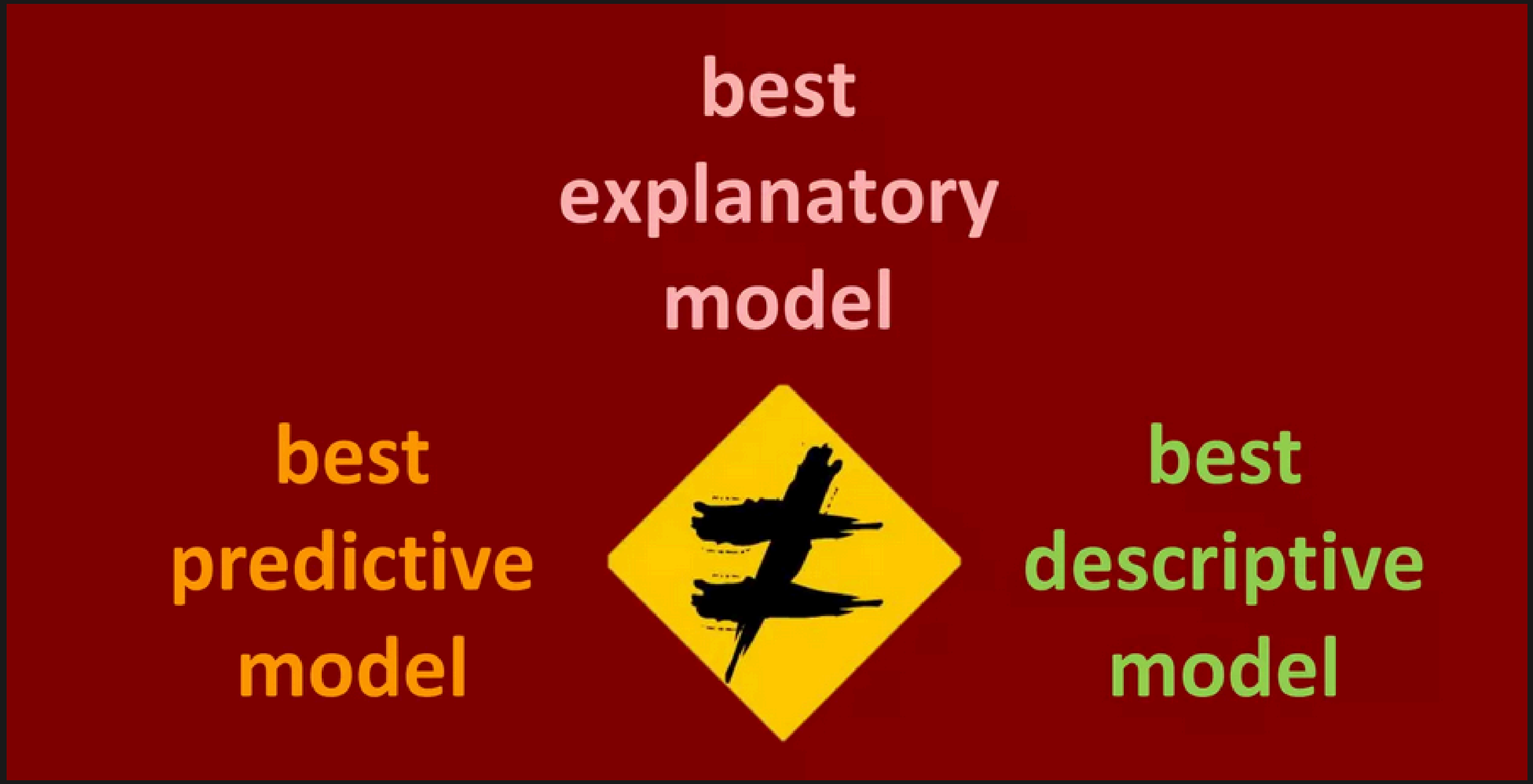
- **No Supervisado ~ Estadística Descriptiva:** Ambos describen o exploran datos sin un objetivo específico de predicción o inferencia.

- **Supervisado ~ Estadística Inferencial:** Ambos buscan hacer predicciones o inferencias basadas en los datos, utilizando modelos que se ajustan a los datos etiquetados o a muestras de la población.

Let’s the data talk!

- La generación de grandes cantidades de información, exige a los investigadores **dar sentido a esa información** entendiendo qué nos **dicen los datos**.
- Analizar la información, mediante un amplio **conjunto de herramientas** que permitan entender los patrones detectados

Cada pregunta tiene su análisis



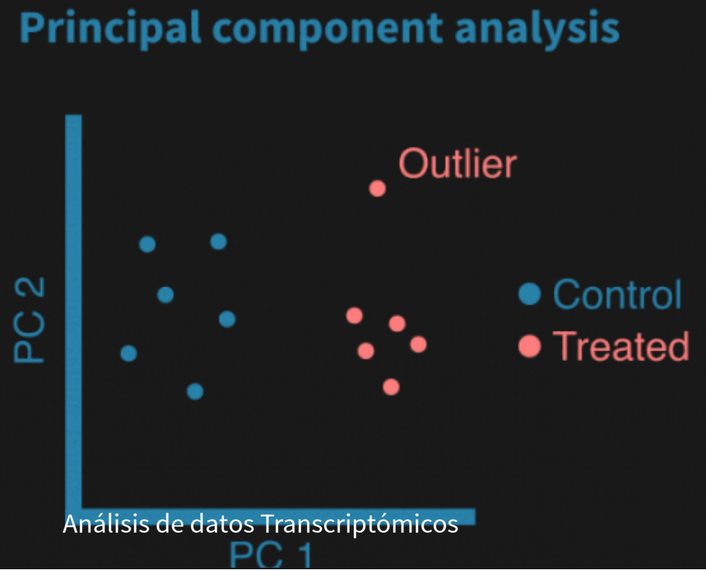
Herramientas para análisis exploratorios

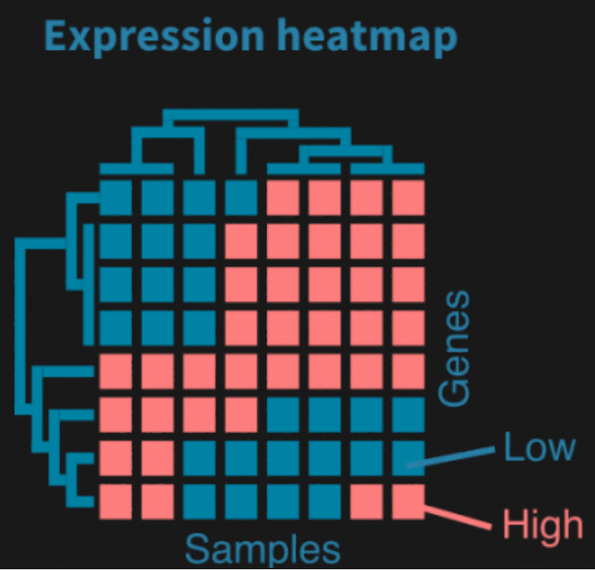
PCA (Análisis de componentes principales)

Heatmaps (Mapas de calor)

Correlación

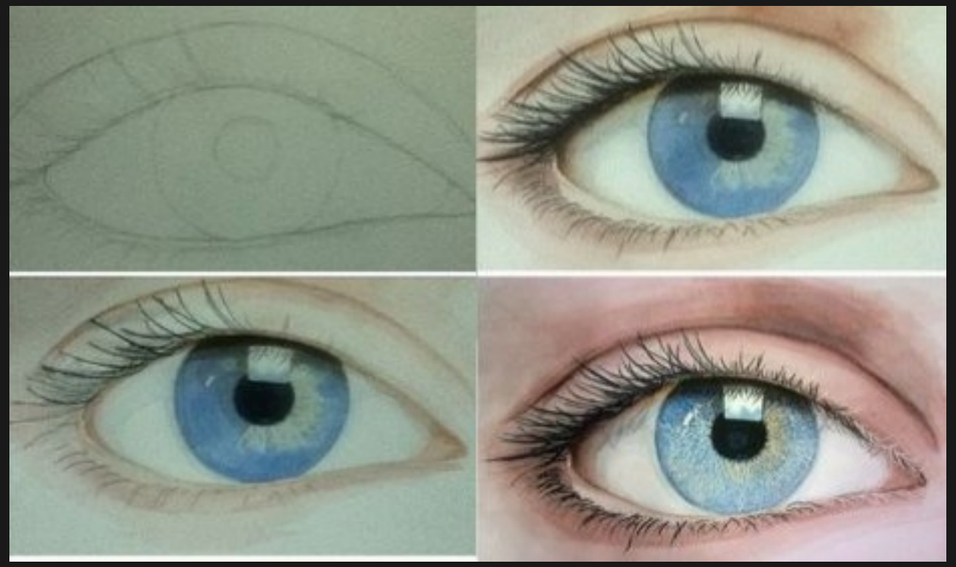
Detección de “outliers”



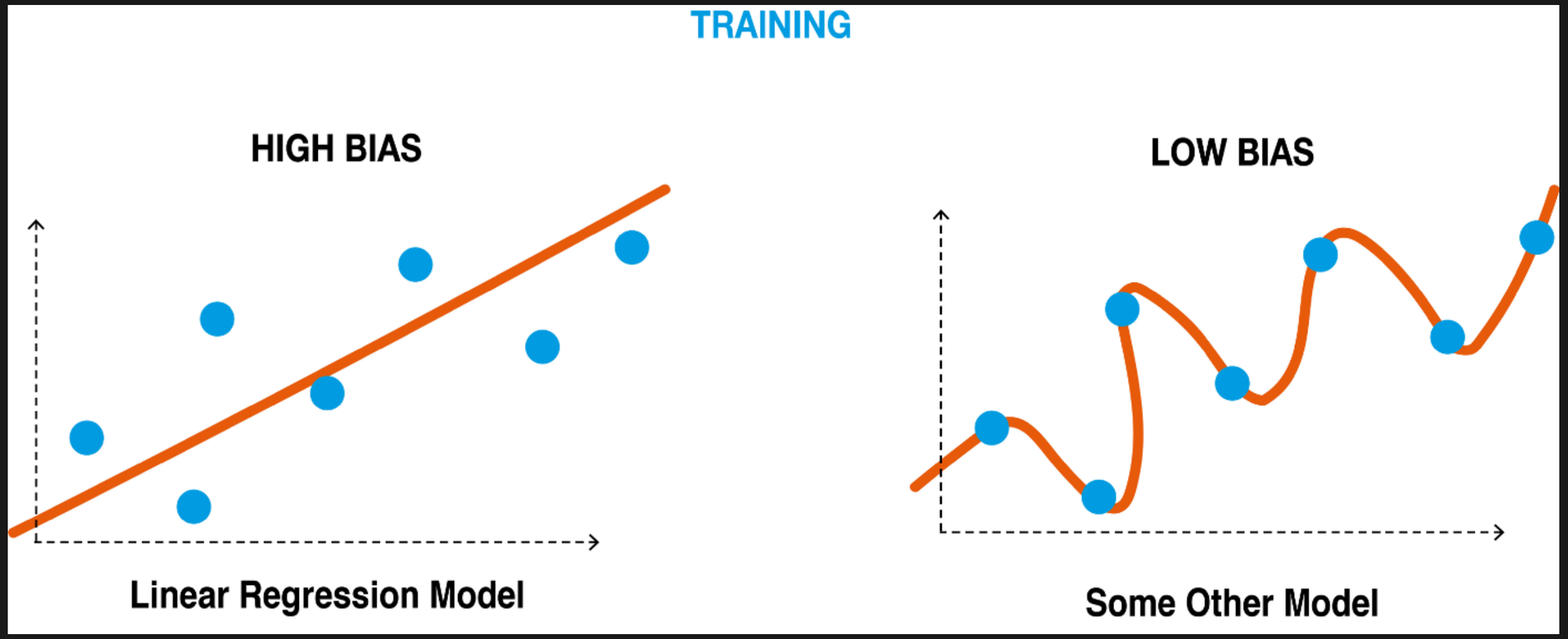


Modelos estadísticos

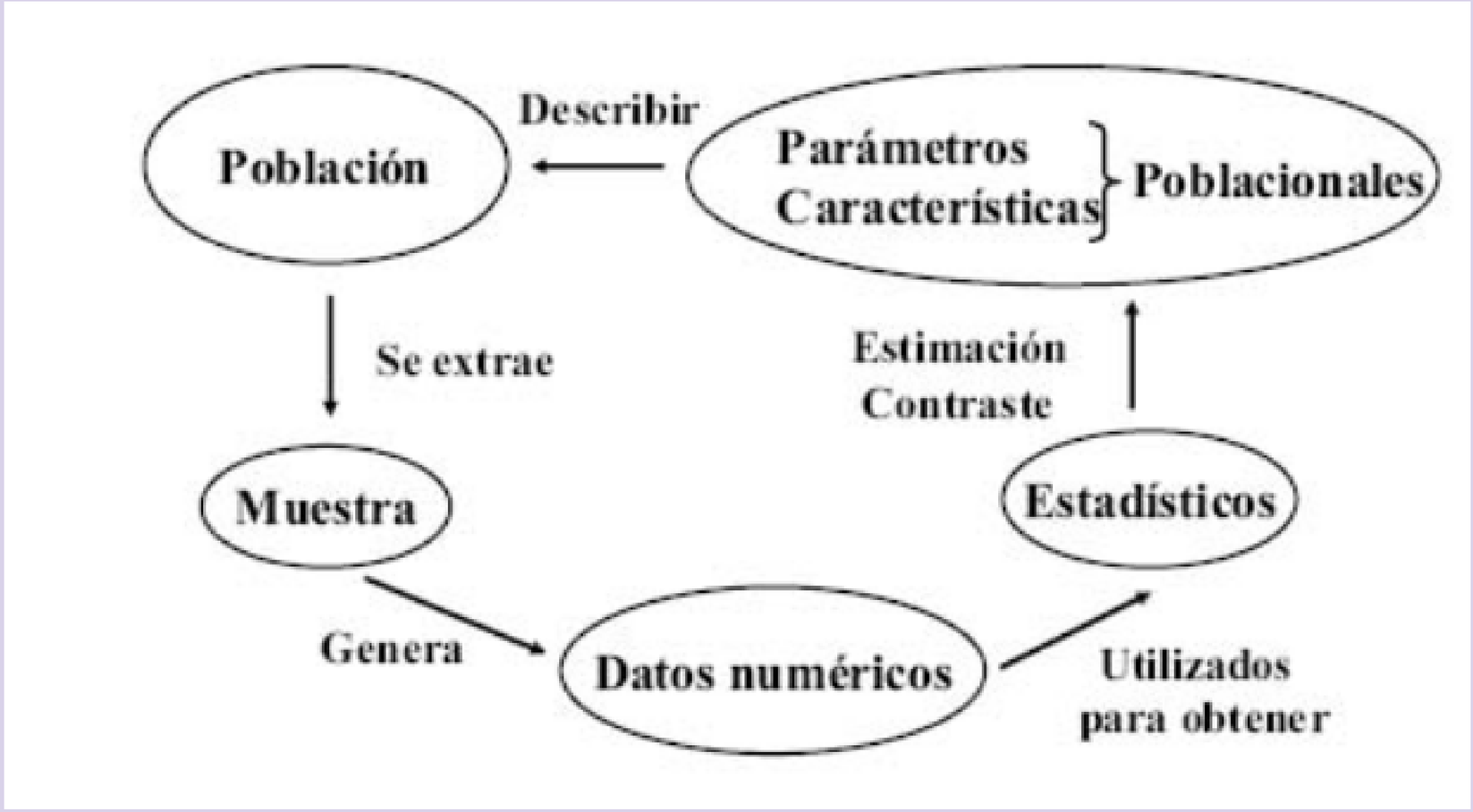
- Abstracción o simplificación que permite representar (matemáticamente) fenómenos de la naturaleza
- ¿Cuál es el Modelo más adecuado de un ojo?



¿Cuál es el modelo que mejor representa a los datos?



El proceso de inferencia



Hipótesis Nula vs Alternativa

- **Hipótesis nula** (H_0) es una hipótesis que está sujeta a pruebas directas. Por ejemplo, si la hipótesis de trabajo es: “Una persona que se levanta temprano en la mañana tendrá una menor satisfacción con la vida”. El objetivo de la hipótesis nula es rechazarla, por lo que se plantea de forma que represente el escenario de “no efecto”; H_0: “No hay diferencias de satisfacción entre las personas que se levantan temprano y las que no”.

El proceso comienza con la suposición de que la hipótesis nula es cierta. Sin embargo, como es poco probable que sea verdadera, se espera que sea rechazada desde el principio.

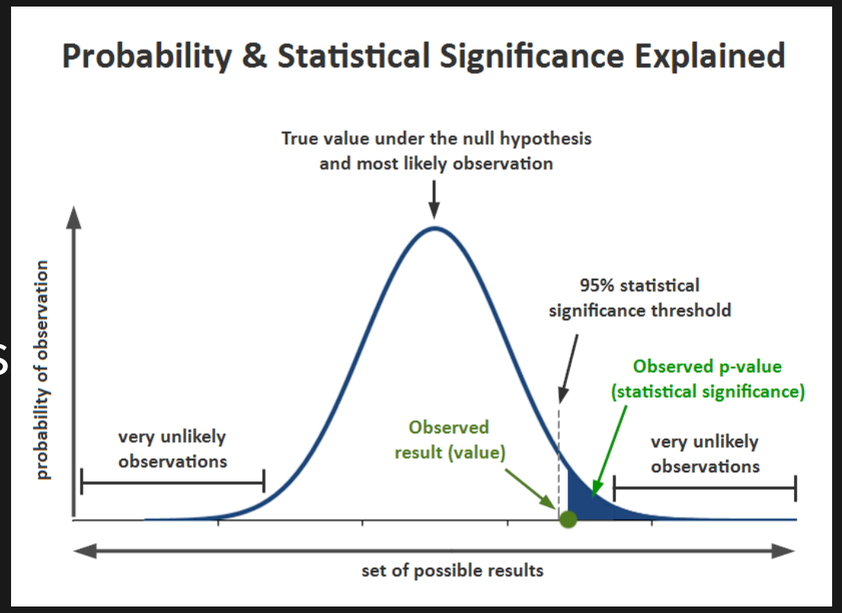
- **Hipótesis alternativa** (H_1) es una alternativa a la hipótesis nula. Esta hipótesis se “acepta” cuando la hipótesis nula es rechazada. La hipótesis alternativa es una nueva afirmación o la hipótesis que realmente se desea probar. En otras palabras, la hipótesis alternativa es la que se busca adoptar. H_1: “Hay diferencias de satisfacción entre las personas que se levantan temprano y las que no”

Significancia estadística

- Ejemplo con t test
  - Comparar el valor de medias entre dos grupos
  - ¿Las medias son tan diferentes que debemos asumir que esas diferencias no pueden ser atribuidas al azar?
  - $H_0 : \mu_1 = \mu_2$
  - Estadistico t

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{SE_{(\overline{x_1} - \overline{x_2})}}$$

  - **p-valor:** Probabilidad de obtener una relación entre las medias mayor a la observada, si H_0 fuera cierta
  - Si p-valor es menor a cierto umbral (ej: 5%), significa que las medias observadas son verdaderamente distintas y por ende rechazo H_0



Modelos estadísticos para la detección de Genes Diferencialmente expresados

La detección de genes diferencialmente expresados (DEGs) es un aspecto clave en los análisis transcriptómicos.La cual se analizan mediante **Modelos Lineales Generalizados**.

- Los datos de RNA-seq son generalmente recuentos de lecturas, y las distribuciones adecuadas para modelarlos son cruciales:
  - Distribución de Poisson: asume que media y varianza son iguales. Sin embargo, en la mayoría de los experimentos de RNA-seq, la varianza tiende a ser mayor que la media (sobredispersión), lo que hace que este modelo sea poco adecuado.
  - Distribución Binomial Negativa: Para abordar la sobredispersión, se utiliza este modelo, que permite que la varianza sea mayor que la media, ajustándose mejor a los datos reales de RNA-seq.

Modelos estadísticos para la detección de Genes Diferencialmente expresados (II)

- EdgeR y DESeq2 son dos de las herramientas más populares para analizar datos de RNA-seq. Ambos paquetes utilizan distribuciones binomiales negativas para modelar los recuentos, permitiendo manejar la sobredispersión
- Ajuste del modelo: Se ajusta un modelo lineal generalizado para cada gen.
- Los genes que muestran diferencias significativas entre condiciones (tratamiento vs control, por ejemplo) se identifican utilizando estadísticas basadas en la razón de verosimilitudes o en tests Wald.
- Dado que los análisis de RNA-seq involucran miles de pruebas simultáneas, es fundamental corregir los p-valores para evitar falsos positivos.
- El método más común es la tasa de descubrimientos falsos (False Discovery Rate, FDR). Esto asegura que el porcentaje de genes identificados como significativamente expresados erróneamente se mantiene bajo control.

P valores y fold change

- **p-valor:** Representa la probabilidad de obtener un resultado al menos tan extremo como el observado si la hipótesis nula es cierta (no hay diferencia en la expresión).

En el contexto de RNA-seq, se utilizan valores ajustados por FDR para seleccionar DEGs.

- **Fold-change:** Mide la magnitud del cambio en la expresión génica entre dos condiciones (tratamiento vs control, por ejemplo). Se usa en combinación con el p-valor ajustado para identificar genes relevantes.

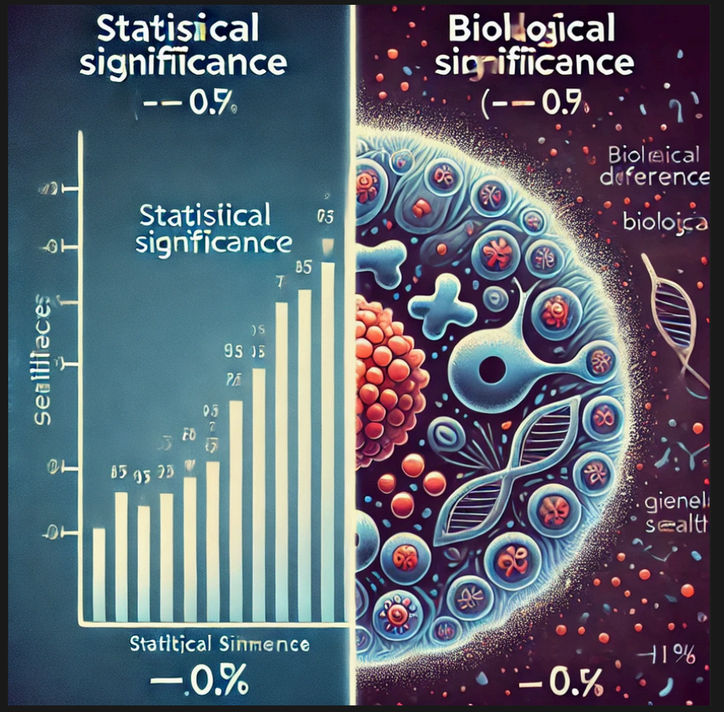
Significancia: estadística vs biológica

- **Significancia biológica**

Refiere a la importancia práctica o relevancia de un hallazgo de investigación en términos de su impacto en los organismos vivos o ecosistemas. Se enfoca en las implicaciones del mundo observado (“real”) de los resultados de un estudio.

- **Significancia estadística**

Medida de la probabilidad de que un hallazgo de investigación no sea producto del azar. Es un concepto matemático que ayuda a los investigadores a determinar si sus resultados son fiables y pueden generalizarse a una población más amplia.



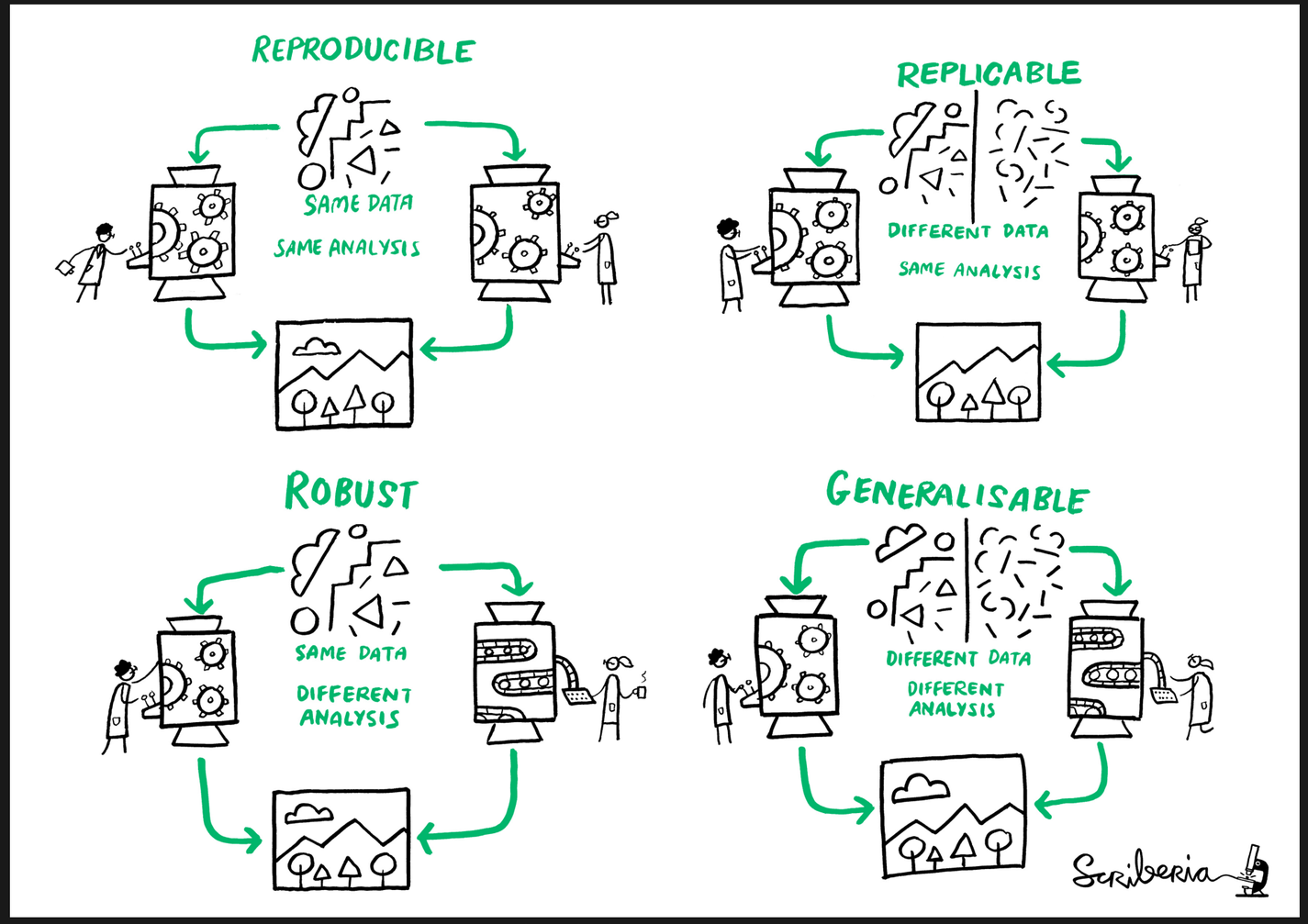
Significancia: estadística vs biológica (II)

- En un ejemplo hipotético de dos nuevos agentes quimioterapéuticos para tratar el cáncer, el Fármaco A aumentó la supervivencia en al menos 10 años con un valor de $P < 0,01$. Por lo tanto, este estudio tiene significancia estadística y significancia clínica (aumento de la supervivencia en 10 años). Un segundo agente quimioterapéutico, el Fármaco B, aumenta la supervivencia en al menos 10 minutos con un valor $P < 0,01$. El estudio sobre el Fármaco B también encontró significancia estadística, pero no significancia clínica (un aumento de 10 minutos en la expectativa de vida no es clínicamente significativo).
- En un estudio separado, aquellos que tomaron el Fármaco C vivieron un promedio de 8 años después de comenzar el tratamiento, en comparación con vivir solo 2 años más para aquellos que no tomaron el Fármaco C ($p = 0.08$). En este segundo estudio, no hay significancia estadística. ¿Y Cliníca?

Buenas Prácticas Y Reproducibilidad

Pregúntese…

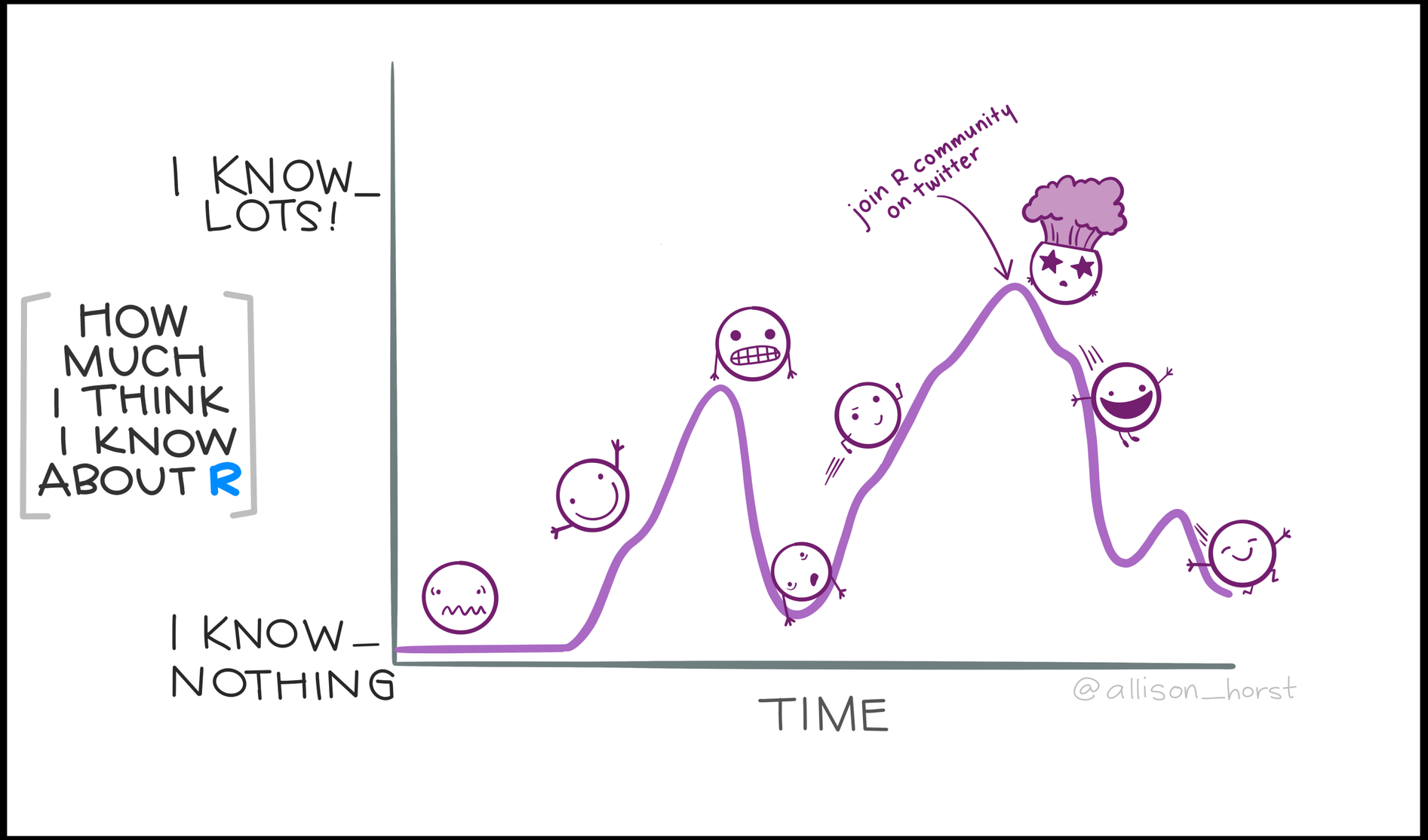
¿Podría un investigador reproducir el estudio con la información que proporciono en este documento?



Conclusiones

- La estadística se basa en el análisis de evidencia observada de forma directa o indirecta
- Dejemos que los datos hablen sin torturarlos
- Entender el contexto del problema y realizar un adecuado análisis exploratorio de los datos
- No todos los cambios estadísticamente significativos son biológicamente significativos, y viceversa
- La Ciencia es una práctica que necesita ser abierta y reproducible

Muchas gracias



Bibliografia citada y recomendada

Arango, N., Chaves, M. E., & Feinsinger, P. (2009). Principios y práctica de la enseñanza de ecología en el patio de la escuela. Fundación Senda Darwin.

Hastie, T., & Tibshirani, R. (2003). Expression Arrays and the p≫ n Problem. See< http://Www-Stat. Stanford. Edu/~ Hastie/Papers/Pgtn. Pdf, 1-14.

Krzanowski, W.J. (2014). Multivariate Analysis: Overview . In Wiley StatsRef: Statistics Reference Online (eds N. Balakrishnan, T. Colton, B. Everitt, W. Piegorsch, F. Ruggeri and J.L. Teugels). https://doi.org/10.1002/9781118445112.stat06467

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome biology, 15, 1-21.

Sharma H. Statistical significance or clinical significance? A researcher's dilemma for appropriate interpretation of research results. Saudi J Anaesth. 2021 Oct-Dec;15(4):431-434. doi: 10.4103/sja.sja_158_21. Epub 2021 Sep 2. PMID: 34658732; PMCID: PMC8477766.

Vandever C. Introduction to Research Statistical Analysis: An Overview of the Basics. HCA Healthc J Med. 2020 Apr 28;1(2):71-75. doi: 10.36518/2689-0216.1062. PMID: 37425244; PMCID: PMC10324782.

Yuwen Liu, Jie Zhou, Kevin P. White, RNA-seq differential expression studies: more sequence or more replication?, Bioinformatics, Volume 30, Issue 3, February 2014, Pages 301–304, https://doi.org/10.1093/bioinformatics/btt688

Sesgo (bias) y Varianza