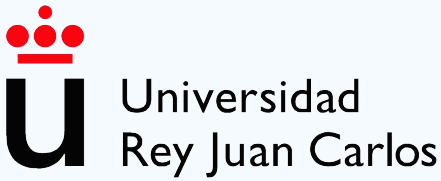
Introducción a la Inferencia Estadística

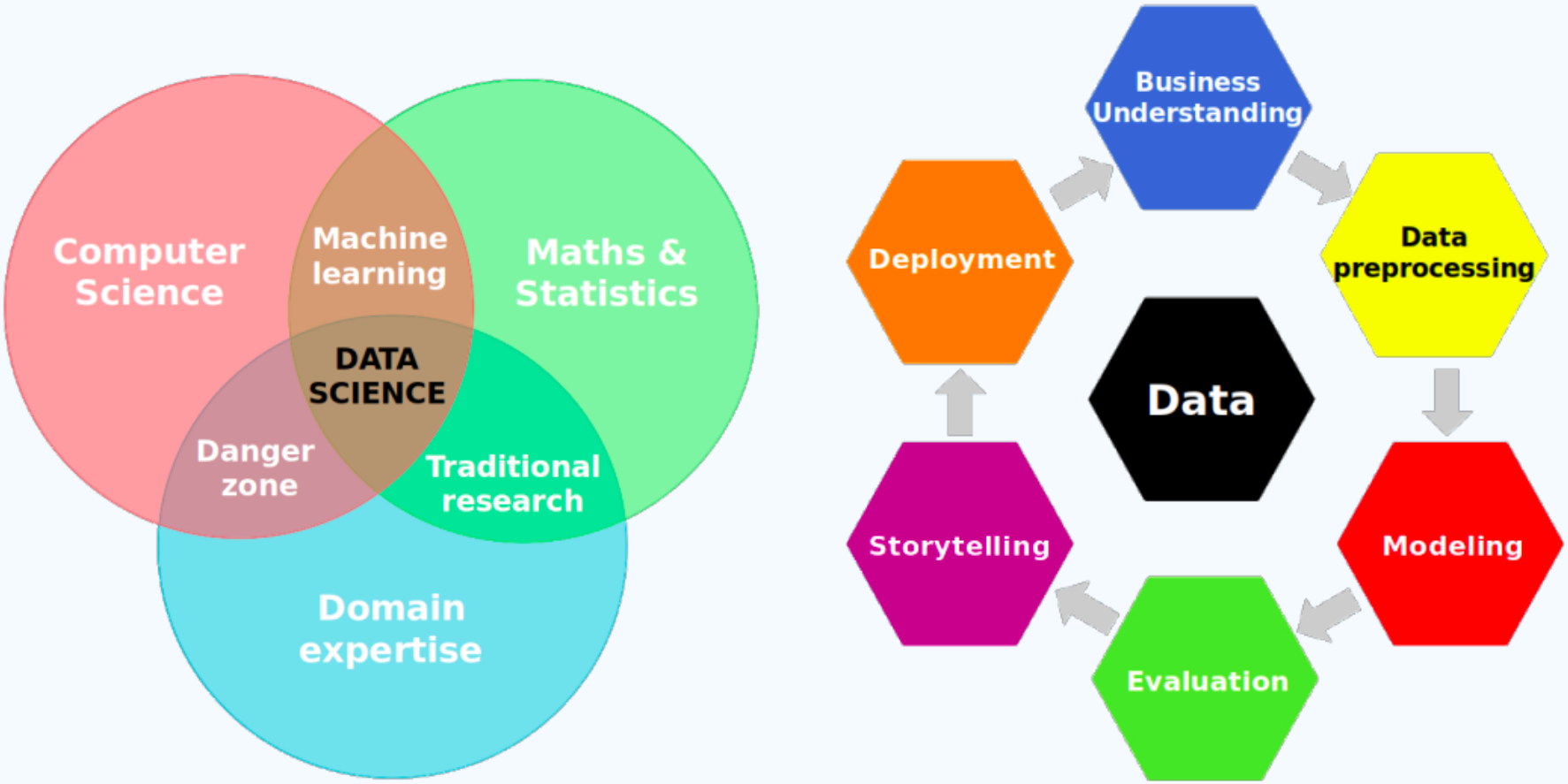
Inferencia Estadística - Grado en Ciencia e Ingeniería de Datos

Curso académico 2024-2025



Asignatura

- Inferencia Estadística
- Grado en Ciencia e Ingeniería de Datos



(a) Foundations (b) Applications

Figure 1.1: Data Science

Cross Industry Standard Process for Data Mining (CRISP-DM)

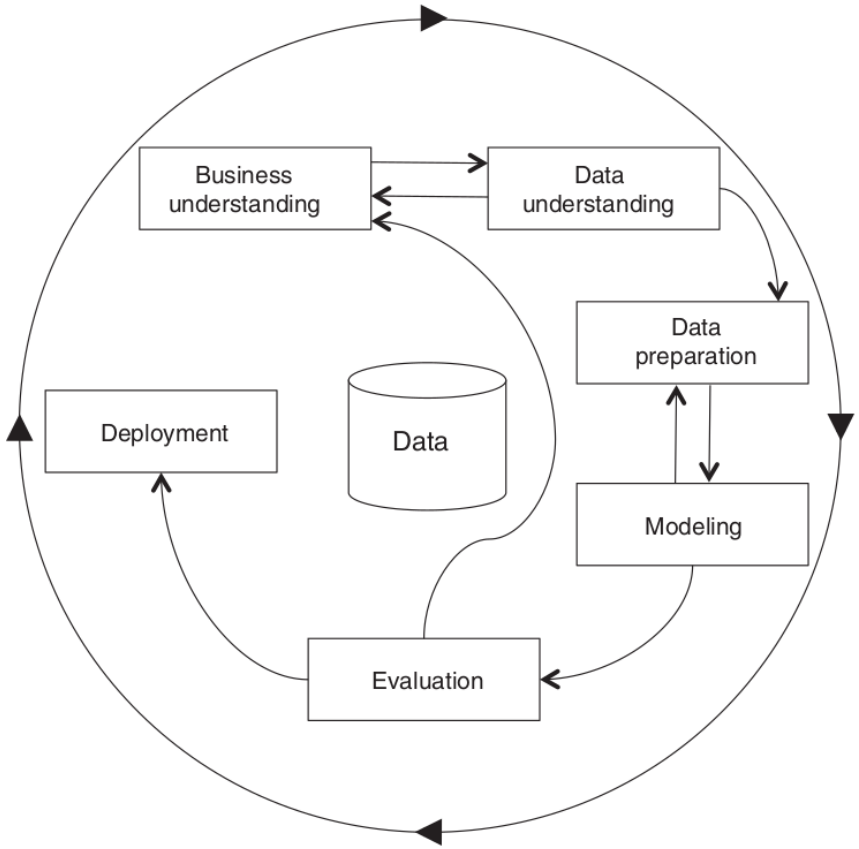


Figure 4 The CRISP-DM life cycle (based on figure 2 in Chapman, Clinton, Kerber, et al. 1999).

Figure 1: Kelleher, J. D., & Tierney, B. (2018). Data science. MIT Press.

CRISP-DM

- **Business Understanding**: Comprender plenamente el problema empresarial que se aborda y diseñar una solución de análisis de datos para el mismo
- **Data Understanding**: Comprender las diferentes fuentes de datos disponibles en la entidad y los diferentes tipos de datos que contienen dichas fuentes
- **Data Preparation**: Poner las distintas fuentes de datos disponibles en un formato adecuado a partir del cual puedan inducirse modelos de aprendizaje automático
- **Modeling**: Crear distintos modelos de aprendizaje automático y selecccionar el mejor para su implantación
- **Evaluation**: Estudiar y validar el rendimiento del modelo para confirmar que es capaz de hacer predicciones precisas antes de ser desplegado
- **Deployment**: Integrar con éxito el modelo de aprendizaje automático en el proceso de la empresa/organización

¿Qué es la Estadística?

- "*La **Estadística** es el arte de aprender a partir de los datos. Está relacionada con la recopilación de datos, su descripción subsiguiente y su análisis, lo que nos lleva a extraer conclusiones.*" Introducción a la Estadística. Sheldon M. Ross.
- "***Statistics** is the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting empirical data. Statistics is a highly interdisciplinary field; research in statistics finds applicability in virtually all scientific fields and research questions in the various scientific fields motivate the development of new statistical methods and theory. In developing methods and studying the theory that underlies the methods statisticians draw on a variety of mathematical and computational tools.*"
`https://www.stat.uci.edu/what-is-statistics/`

¿Qué es la Estadística?

"Statistics, data mining, and machine learning are all concerned with collecting and analyzing data. For some time, statistics research was conducted in statistics departments while data mining and machine learning research was conducted in computer science departments. Statisticians thought that computer scientists were reinventing the wheel. Computer scientists thought that statistical theory didn't apply to their problems.

Things are changing. Statisticians now recognize that computer scientists are making novel contributions while computer scientists now recognize the generality of statistical theory and methodology. Clever data mining algorithms are more scalable than statisticians ever thought possible. Formal statistical theory is more pervasive than computer scientists had realized.

Students who analyze data, or who aspire to develop new methods for analyzing data, should be well grounded in basic probability and mathematical statistics. Using fancy tools like neural nets, boosting, and support vector machines without understanding basic statistics is like doing brain surgery before knowing how to use a band-aid."

Wasserman, L. (2006). *All of nonparametric statistics*. Springer Science & Business Media.

¿Qué es la Estadística?

La **Estadística** es la ciencia que se encarga de recolectar, organizar, analizar e interpretar datos para tomar decisiones informadas. Su objetivo principal es comprender y describir la *variabilidad* inherente en los datos y utilizar esta comprensión para hacer predicciones y tomar decisiones bajo condiciones de incertidumbre.

¡Al abordaje!

1. ¿Cuál es el número medio de horas que pasan las personas presentes en este aula en Instagram? ¿Y si diferenciamos por edad? ¿Y por género?
2. ¿Cuál es el número medio de horas que pasan los alumnos de la URJC en Instagram?

Clasificación de la Estadística

- Estadística **descriptiva**
- Estadística **inferencial** o inferencia estadística

Estadística descriptiva

- Se ocupa de resumir y describir las características de un conjunto de datos mediante herramientas gráficas y numéricas, como tablas, gráficos, medias, medianas, varianzas, etc.
- Su objetivo es proporcionar una visión clara y comprensible de la estructura y características de los datos y de la información contenida en ellos.
- Ejemplo: Tiempo medio que tardan los alumnos de la asignatura de Inferencia Estadística del Grado en Ciencia e Ingeniería de Datos en llegar a la universidad

Inferencia Estadística

- Extracción de conclusiones acerca de una población a partir de una muestra de datos
- Utiliza muestras de datos para hacer generalizaciones o inferencias sobre una población más amplia. Involucra el uso de métodos como la estimación de parámetros, pruebas de hipótesis y la construcción de intervalos de confianza
- La inferencia estadística permite tomar decisiones y hacer predicciones basadas en datos muestreados
- Ejemplo: El tiempo medio que tardan los alumnos de IE de GCID en llegar a la universidad, ¿es representativo del tiempo que tardan todos los estudiantes?

¿Para qué sirve la Inferencia Estadística?

- A diferencia de la mera descripción de datos, la inferencia permite ir más allá de lo observado y hacer generalizaciones, estimaciones y decisiones en presencia de incertidumbre
- Fundamental para cualquier análisis de datos que aspire a ser predictivo o que busque comprender fenómenos más amplios que los capturados por los datos disponibles
- Requiere una recopilación cuidadosa de datos y un análisis riguroso que tenga en cuenta la variabilidad inherente y las posibles fuentes de error
- Clave en la Ciencia de Datos:
  - Modelado predictivo
  - Análisis experimental
  - Decisiones basadas en datos
  - Gestión de la incertidumbre

¿Para qué sirve la Inferencia Estadística?

- Responder a las preguntas que hagamos sobre los datos:
  - ¿Cuál es la estatura media de los estudiantes de la URJC?
  - ¿Quién ganará la Eurocopa?
    `https://twitter.com/kikollan/status/1801667477132202152`
  - ¿Este medicamento es efectivo para reducir la ansiedad?
  - Nueva versión de una app, ¿están los usuarios satisfechos?

Ejemplo en Machine Learning

Evaluación del rendimiento de un modelo de Aprendizaje Automático para predecir transacciones bancarias:

- Trabajas como **científico de datos** en un banco. Tu tarea es desarrollar un **modelo de predicción de fraude** para identificar transacciones fraudulentas
- Problema:
  - El banco tiene millones de transacciones diarias (y solo una pequeña fracción son fraudulentas). Es imposible revisar manualmente todas las transacciones o entrenar y testear tu modelo con todas ellas
  - Necesitas un método que te permita **inferir** cómo se comportará tu modelo con la población completa de transacciones a partir de una muestra

Ejemplo en Machine Learning

¿Qué hacemos como científicos de datos?:

- Obtener una muestra representativa del total de transacciones
- Entrenar un modelo de aprendizaje automático usando dichos datos que permita discernir entre transacciones fraudulentas o no. Evaluar el rendimiento del modelo –> Resultados restringidos a la muestra
- Gracias a la Inferencia Estadística podemos generalizar las conclusiones de la muestra a la población (total de transacciones).
  - Ej: Lograr un Intervalo de confianza para el rendimiento del modelo. Con nivel de confianza del 95%, el rendimiento del modelo en el total de transacciones está en el intervalo [72%,81%].
- Si probamos varios modelos de aprendizaje automático, podemos comparar su rendimiento mediante contrastes de hipótesis y así determinar si la diferencia entre ellos es estadísticamente significativa o no

Datos

Para todo esto, ¿qué necesitamos en primer lugar?

¡DATOS!

	species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex
0	Adelie	Torgersen	39.1	18.7	181	3750	male
1	Adelie	Torgersen	39.5	17.4	186	3800	female
2	Adelie	Torgersen	40.3	18.0	195	3250	female
3	Adelie	Torgersen	36.7	19.3	193	3450	female
4	Adelie	Torgersen	39.3	20.6	190	3650	male
...	...	...	...	...	...	...	...
328	Chinstrap	Dream	55.8	19.8	207	4000	male
329	Chinstrap	Dream	43.5	18.1	202	3400	female
330	Chinstrap	Dream	49.6	18.2	193	3775	male
331	Chinstrap	Dream	50.8	19.0	210	4100	male
332	Chinstrap	Dream	50.2	18.7	198	3775	female

333 rows × 7 columns

Datos y variables

- Los datos son las observaciones o medidas que recopilamos del mundo que nos rodea
- Estos pueden ser números, categorías o cualquier tipo de información cuantificable. Ejemplo: glucosa en sangre, nivel de un terremoto, color de ojos, resultados de una encuesta, etc.
- Llamaremos elementos a los individuos u observaciones sobre las que se recojen un conjunto de atributos o características
- ¡Muy importante el diseño y la recogida de datos!

Datos y variables

- **Variables estadísticas**: los atributos o características que se miden sobre los individuos
- El conjunto de valores que puede tomar una variable se denomina **dominio**.
- Notación: $X, Y, ..$
- Ejemplo: $X =$ Resultado de lanzar un dado de 6 caras. Dominio de X es $\{1, 2, 3, 4, 5, 6\}$
- Cada valor del dominio tiene asignada una probabilidad $\rightarrow$ concepto de **variable aleatoria**

Población y muestra

- La **población** es el conjunto completo de todos los elementos o individuos que se desean estudiar
- Por ejemplo,todos los estudiantes de una universidad, todos los árboles en un bosque, o todos los productos fabricados en una planta
- ¿Es posible estudiar siempre toda la población?
  - Destrucción de las observaciones (vida útil del componente)
  - Coste elevado (experimentos biológicos)
  - Poblaciones muy muy grandes

Población y muestra

- ¿Solución? Tomar una muestra: subconjunto de la población
- La muestra tiene que ser representativa de la población. En caso contrario, las conclusiones que se extraigan no serán válidas para la población (¡y este era el objetivo!)

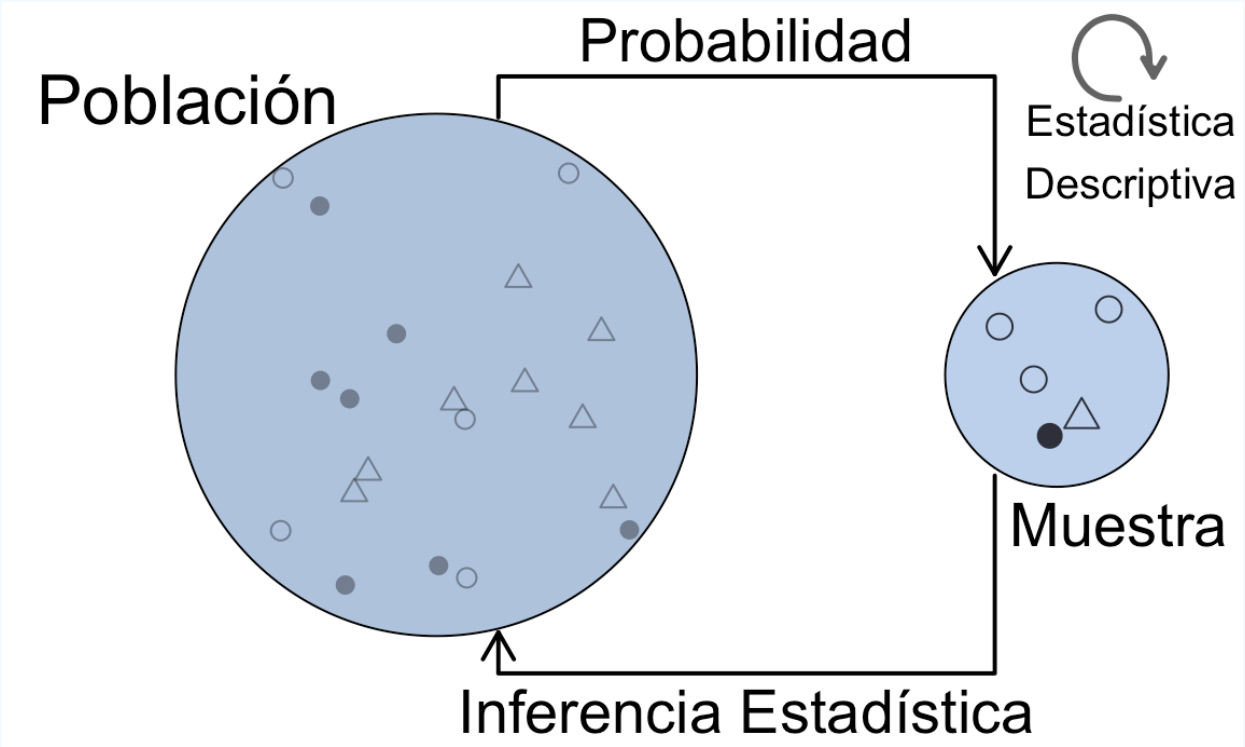
Población y muestra

- Una vez que se ha recolectado una muestra, la **estadística descriptiva** se utiliza para organizar, resumir y presentar los datos de manera comprensible mediante:
  - Medidas de tendencia central: como la media, mediana y moda, que resumen el centro de los datos.
  - Medidas de dispersión: como el rango, la varianza y la desviación estándar, que describen la variabilidad de los datos.
  - Gráficos: como histogramas, gráficos de caja y gráficos de dispersión, que visualizan los datos.
- La estadística descriptiva se centra en describir lo que los datos muestran, sin hacer inferencias o generalizaciones sobre la población.

Población y muestra

- La **inferencia estadística** utiliza los datos de la muestra para hacer estimaciones, predicciones y generalizaciones sobre la población completa:
  - Estimación: Utilizar los datos de la muestra para estimar parámetros de la población, como la media.
    - Puntuales (un solo valor)
    - Por intervalo (un rango de valores con un nivel de confianza asociado).
  - Contraste de hipótesis: Probar afirmaciones sobre la población utilizando los datos de la muestra. Esto implica formular una hipótesis nula y una hipótesis alternativa, y usar pruebas estadísticas para decidir cuál es más consistente con los datos observados.
- La inferencia estadística se basa en la teoría de la probabilidad para evaluar la incertidumbre y la variabilidad en las estimaciones y pruebas.

Población y muestra



$$DATOS = MODELO + ERROR$$

- Los **datos** representan la realidad (procesos de negocios, clientes, productos, actividades, fenómenos físicos, etc.) que se quiere comprender, predecir o mejorar
- El **modelo** es una representación **simplificada** de la realidad que proponemos para describirla e interpretarla más fácilmente.
- El **error** refleja la diferencia entre nuestra representación simplificada de la realidad (el modelo) y los datos que realmente describen esa realidad de forma precisa.

Muestreo

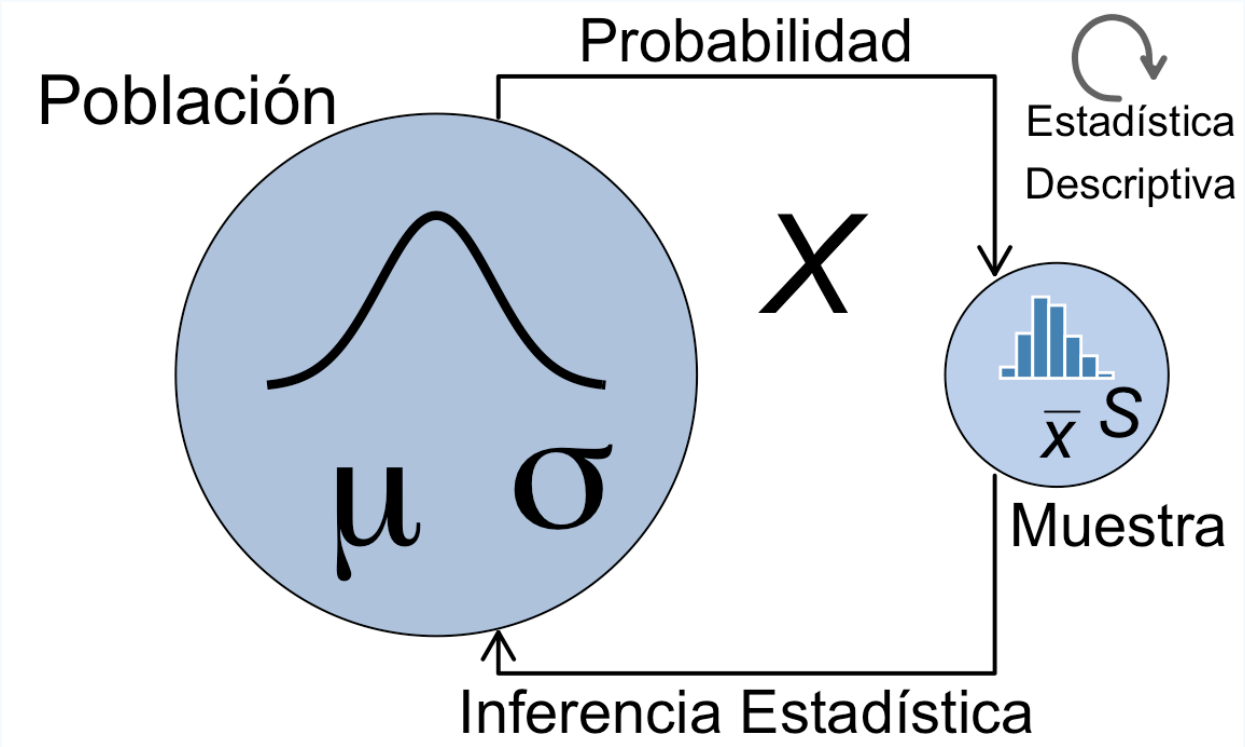
El muestreo estadístico es una técnica fundamental en la estadística que permite extraer conclusiones sobre una población basándose en el análisis de una parte más pequeña de dicha población, conocida como muestra

- **Muestreo aleatorio simple** (con y sin reemplazamiento)
- Muestreo sistemático
- Muestreo estratificado
- Muestreo por conglomerados
- ...

Parámetros y estadísticos

- Estamos interesados en conocer alguna característica de una población (ejemplo la altura media de los portugueses). Dicha característica se denota como **parámetro poblacional**.
- Los **parámetros poblaciones** son valores teóricos desconocidos que se definen sobre la población. Son sobre los que haremos *inferencia*. Normalmente se representan con letras griegas (por ejemplo, media poblacional μ).
- ¿Cómo haremos dicha inferencia?
  - Como no podemos trabajar con toda la población, se obtiene una muestra representativa de la misma
  - En la muestra, se estudian las características de interés (estatura) y se usará para calcular una estimación del parámetro poblacional. Esta estimación se llama **estadístico muestral** y ya es conocido (a diferencia del parámetro poblacional).
  - El estadístico muestral es una función real definida sobre los datos de la muestra que estima el valor de un parámetro poblacional. Se representa con letras latinas (por ejemplo, $\overline{x}$ para la media muestral).

Parámetros y Estadísticos



Estadística paramétrica

- Se basa en la suposición de que los datos siguen una distribución de probabilidad conocida, como la distribución Normal, binomial, Poisson, etc.
- Estas suposiciones deben ser comprobadas para dar validez a este tipo de pruebas.
- Los parámetros de estas distribuciones, como la media y la varianza, se utilizan para resumir la información de los datos y realizar inferencias.
- Objetivo: obtener información sobre el parámetro de interés mediante la obtención de muestras de la variable aleatoria
- Más potentes que los no paramétricos (i.e., tienen una mayor probabilidad de detectar un efecto verdadero) si las suposiciones son correctas.

Estadística paramétrica. Familias paramétricas.

Sea una variable aleatoria X cuya distribución pertenece a una cierta familia paramétrica $\{f_\theta\}$ donde $\theta \in \Theta$.

La distribución de X es conocida excepto por el valor del parámetro θ , del cual lo único que se conoce es su rango de posibles valores Θ, denominado espacio paramétrico.

Algunos ejemplos de familias paramétricas:

- $X \sim N(\mu, \sigma^2) \rightarrow \theta = (\mu, \sigma^2)$
- $X \sim Bernoulli(p) \rightarrow \theta = p$
- $X \sim Exp(\lambda) \rightarrow \theta = \lambda$

Estadítica no paramétrica

- No hace suposiciones fuertes sobre la distribución de los datos.
- Más flexibles y robustos a las violaciones de las suposiciones, pero pueden ser menos potentes si las suposiciones de los métodos paramétricos son verdaderas.
- Los métodos no paramétricos a menudo se basan en el orden de los datos, en lugar de sus valores exactos.

Estadística paramétrica y no paramétrica

La elección entre métodos paramétricos y no paramétricos depende de la naturaleza de los datos y de las suposiciones que estemos dispuestos a hacer. Si los datos cumplen con las suposiciones de una prueba paramétrica, esa prueba puede ser la opción más potente. Si no, una prueba no paramétrica puede ser más apropiada.

Ejemplos

- Paramétrica. Prueba *t de Student*, el *análisis de varianza (ANOVA)* y la *regresión lineal* que veréis en el segundo cuatrimestre
- No paramétrica. Prueba de *Mann-Whitney U*, la prueba de *Kruskal-Wallis* y la prueba de *Chi-cuadrado*

Frecuentistas vs Bayesianos

Existen dos enfoques en Estadística:

- Enfoque frecuentista
- Enfoque Bayesiano

Enfoque frecuentista

- Interpretan la probabilidad como la frecuencia relativa de un evento en un número infinito de repeticiones del experimento.
- Se obtienen datos a través de una muestra y con técnicas estadísticas se extrae información de los mismos mediante estimadores. En base a esas estimaciones se toman decisiones en el dominio de aplicación.
- Ampliamente utilizados y son la base de muchas técnicas estadísticas clásicas.
- Los parámetros son considerados como valores fijos y desconocidos que se estiman a partir de los datos.

Enfoque bayesiano

- Tiene su fundamento en el teorema de Bayes, formulado por el matemático británico Thomas Bayes en el siglo XVIII
- Es un principio fundamental en la teoría de la probabilidad que describe la forma de actualizar las probabilidades de una hipótesis basándose en nueva evidencia o información

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

donde:

  - $P(A|B)$ probabilidad a posteriori
  - $P(B|A)$ es la verosimilitud
  - $P(A)$ es la probabilidad a priori
  - $P(B)$ probabilidad de B

Enfoque bayesiano

- El teorema de Bayes permite actualizar la probabilidad de una hipótesis A a la luz de nueva evidencia B. Básicamente, proporciona una forma de ajustar nuestras creencias iniciales (probabilidad a priori) en base a la nueva información disponible (evidencia)
- Interpretan la probabilidad como una medida de la creencia o confianza en un evento. Esta creencia puede ser actualizada a medida que se obtiene más información.
- Los parámetros son considerados como variables aleatorias y se describe su incertidumbre a través de distribuciones de probabilidad.
- Los métodos bayesianos permiten la incorporación directa de conocimientos previos en el análisis a través de la distribución a priori.

Referencias

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical statistics for data scientists: 50+ essential concepts using R and Python*. O'Reilly Media

Kelleher, J. D., & Tierney, B. (2018). *Data science*. MIT Press.

Ross, S. M. (2018). *Introducción a la estadística*. Reverté.

Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

Canavos, G. C., & Medal, E. G. U. (1987). *Probabilidad y estadística* (p. 651). México: McGraw Hill.