Universidad Rey Juan Carlos

ESCUELA TÉCNICA SUPERIOR
DE INGENIERÍA INFORMÁTICA

Asignatura: INFERENCIA ESTADÍSTICA
Grado en Ciencia e Ingeniería de Datos

Diapositivas de la asignatura

(Fecha del material: Noviembre 2024)

Curso académico 2024-2025

Material docente en abierto de la Universidad Rey Juan Carlos

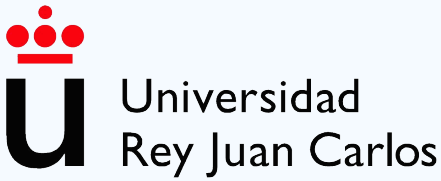
Autores: Carmen Lancho, Víctor Aceña, Isaac Martín de Diego

Índice de las diapositivas

Introducción a la Inferencia Estadística

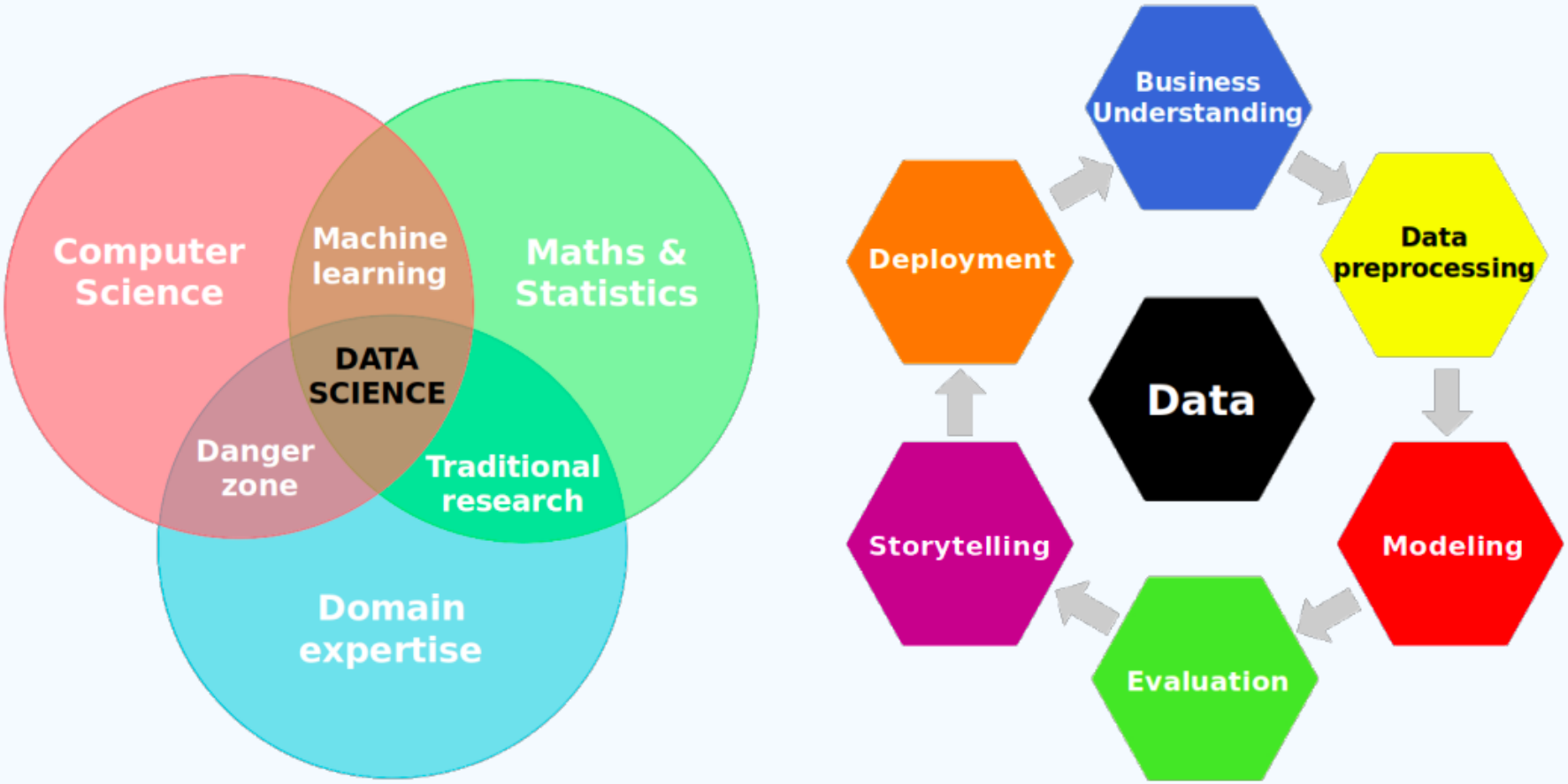
Inferencia Estadística - Grado en Ciencia e Ingeniería de Datos

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)

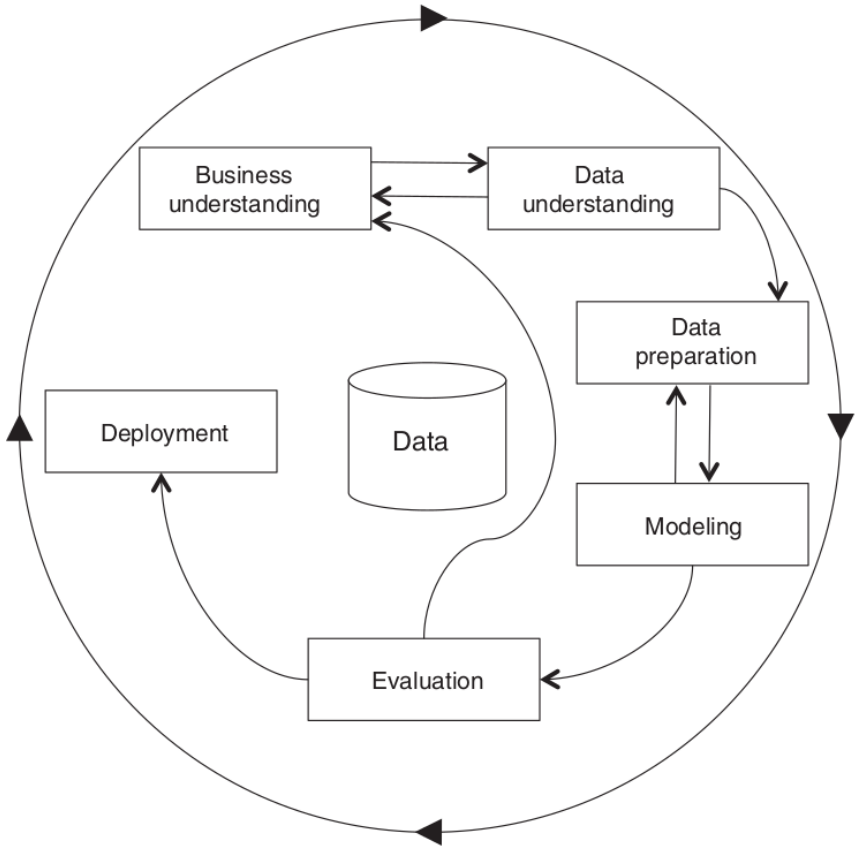


Figure 4 The CRISP-DM life cycle (based on figure 2 in Chapman, Clinton, Kerber, et al. 1999).

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!)

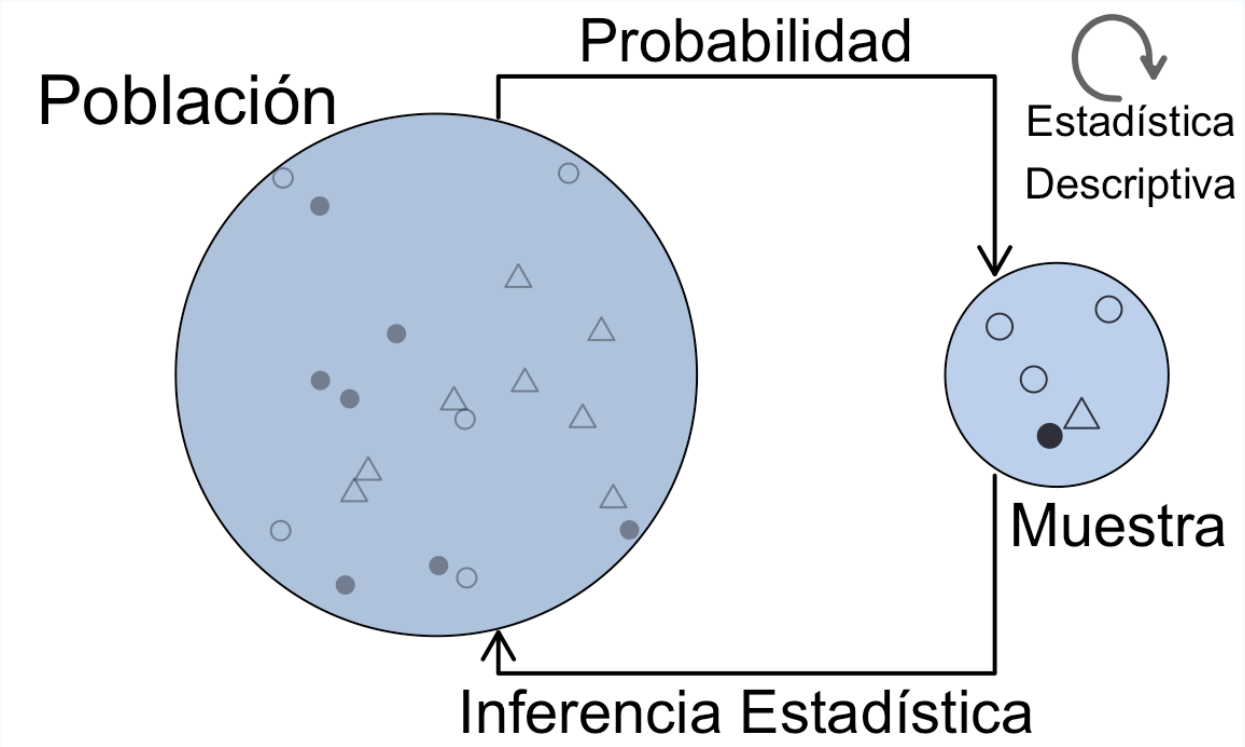
Población y muestra

- 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

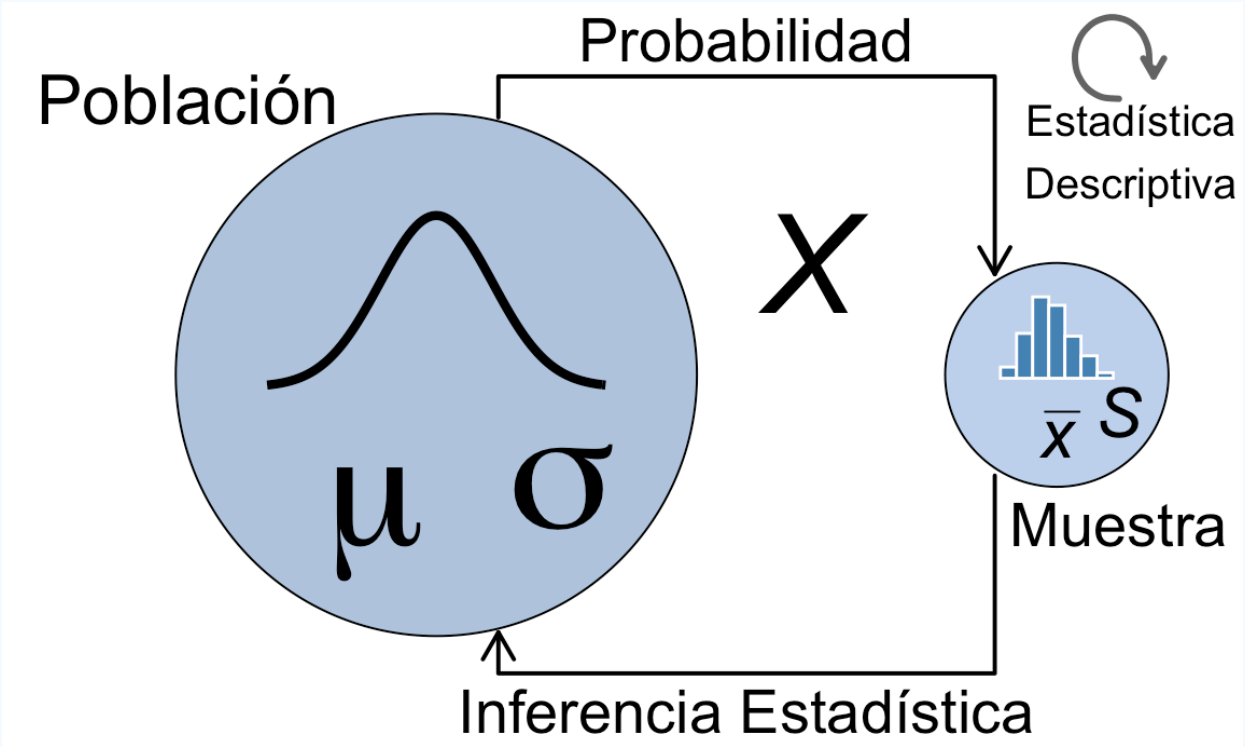
El muestreo estadístico es una técnica fundamental en la estadística que permite extraer conclusiones sobre una población basándose en el análisis de una parte más pequeña de dicha población, conocida como muestra

- **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.

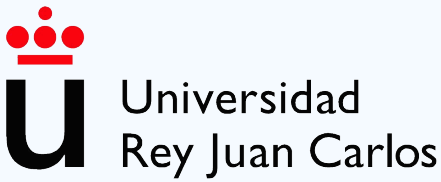
Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

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

Estadística descriptiva. EDA

Inferencia Estadística - Grado en Ciencia e Ingeniería de Datos

Curso académico 2024-2025



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
- 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

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

Tipos de variables

- **Cualitativas o categóricas**: Describen cualidades. Se dividen en:
  - **Nominales**. Sin orden específico. Ej: color de ojos
  - **Ordinales**. Tienen un orden. Ej: niveles de satisfacción
- **Cuantitativas**: Valores numéricos que se pueden medir. Pueden ser:
  - **Discretas**. Valores contables, como el número de hijos
  - **Continuas**. Pueden tomar cualquier valor dentro de un rango, como la altura o el peso.

Tipos de variables

- Marcas de tiempo o identificadores: Como por ejemplo la fecha y hora de una transacción o el código de un producto o el número de identidad.

Escalas de medición

Una escala de medición define cómo se cuantifican o categorizan las variables recogidas sobre un conjunto de datos, influyendo en el análisis estadístico aplicable:

- **Nominal**: categorización sin orden inherente. Por ejemplo, el género, la nacionalidad o el tipo de sangre
- **Ordinal**: categorización con un orden lógico. Por ejemplo, el nivel educativo, o una clasificación de hoteles
- **Métrica**:
  - **Intervalo**: sin cero verdadero, por ejemplo la temperatura en Celsius.
  - **Razón**: con cero verdadero, por ejemplo los ingresos o la distancia.

Conversión entre escalas de medición

- Fechas a categóricas: convertir fechas exactas en mes, día de la semana, etc.
- Cuantitativas a cualitativas: crear clases o rangos a partir de datos numéricos. Por ejemplo convertir el nivel de ingresos en “bajo”, “medio” y “alto”.
- Variables calculadas: creación de nuevas variables a partir de las existentes. Por ejemplo, se crea el Índice de Masa Corporal (IMC) a partir de peso y altura.

Estadística descriptiva

- Aplicación de técnicas matemáticas para resumir un conjunto de datos
- Objetivo: Presentar los datos de manera clara mediante medidas de tendencia central (media, mediana, moda), medidas de dispersión (desviación estándar, varianza), y visualizaciones (tablas de frecuencia, gráficos de barras, histogramas, etc.).
- Sigue un enfoque más formal y estructurado, centrado en describir las características principales de un conjunto de datos
- Limitaciones: Se centra en los datos de manera resumida, sin necesariamente buscar patrones complejos, relaciones o anomalías.

Análisis exploratorio de datos (EDA)

"Exploratory data analysis can never be the whole story, but nothing else can serve as the foundation stone as the first step"

"Exploratory Data Analysis is detective work"

John Tukey

Análisis exploratorio de datos

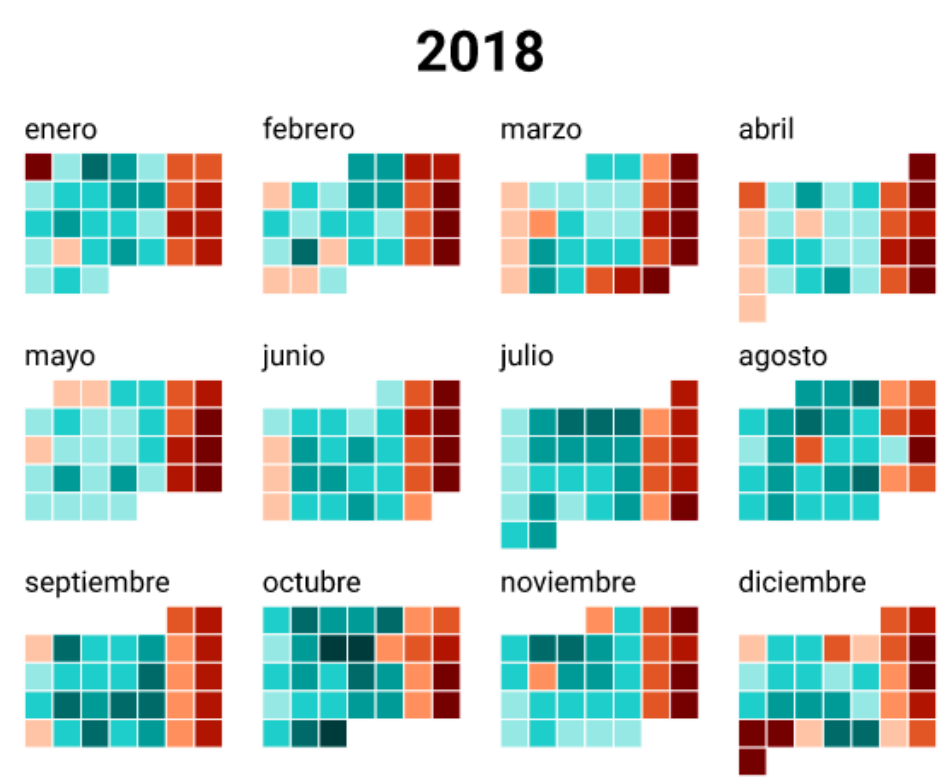
- Enfoque más amplio y flexible que combina herramientas matemáticas y estadísticas son técnicas gráficas para descubrir patrones, tendencias y relaciones en los datos
- Objetivo: Busca explora y entender los datos de manera más profunda e interactiva, generando hipótesis y obteniendo información antes de aplicar técnicas de modelado más formales
- Sigue un enfoque más experimental y visual, permitiendo descubrir patrones inesperados o relaciones ocultas. No hay un guión estricto para realizar un EDA (¡somos detectives!)
- Limitaciones: Para sacar conclusiones generales (no limitadas a la muestra) debe ser seguida por análisis más formales o inferenciales

Importancia del EDA

Supongamos que tenemos los datos de natalidad desde 1977 hasta 2018. ¿Qué haríais con ellos?

Importancia del EDA

Estudio natalidad: https://www.eldiario.es/nidos/no-ninos-nacen-toca-dar-luz-semana-21-probable-hacerlo-lunes-viernes_1_6400307.html

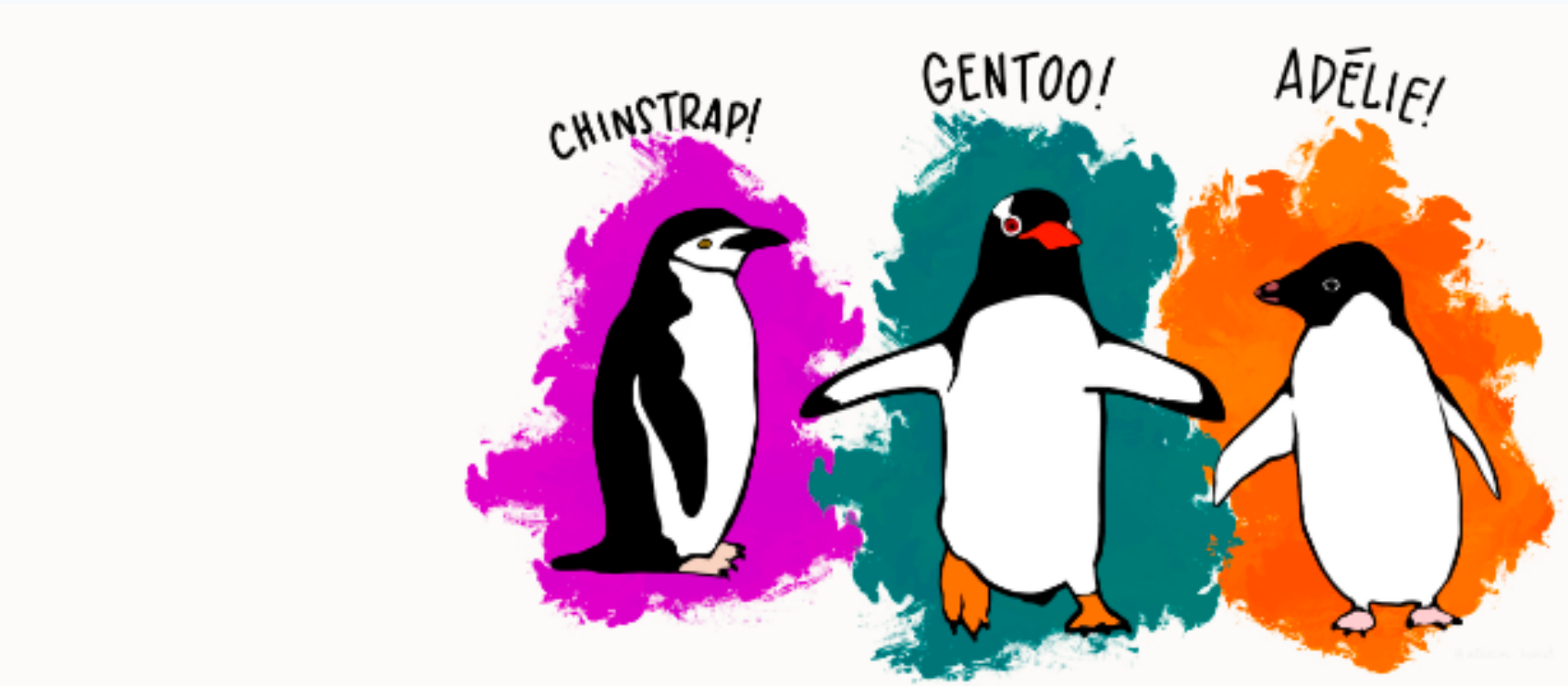


Estadística descriptiva

- Estudiemos algunas de las herramientas de la estadística descriptiva y el EDA
- Herramientas
  - Resúmenes numéricos: media, moda, mediana, cuantiles, tablas de frecuencia, etc
  - Métodos gráficos. diagramas de barras, histograma, boxplot, etc.
- En función de los datos (categóricos o continuos), se usarán unos métodos u otros

Dataset Palmer penguin

https://allisonhorst.github.io/palmerpenguins/articles/intro.html



The Palmer Archipelago penguins. Artwork by @allison_horst.

Dataset Palmer penguin

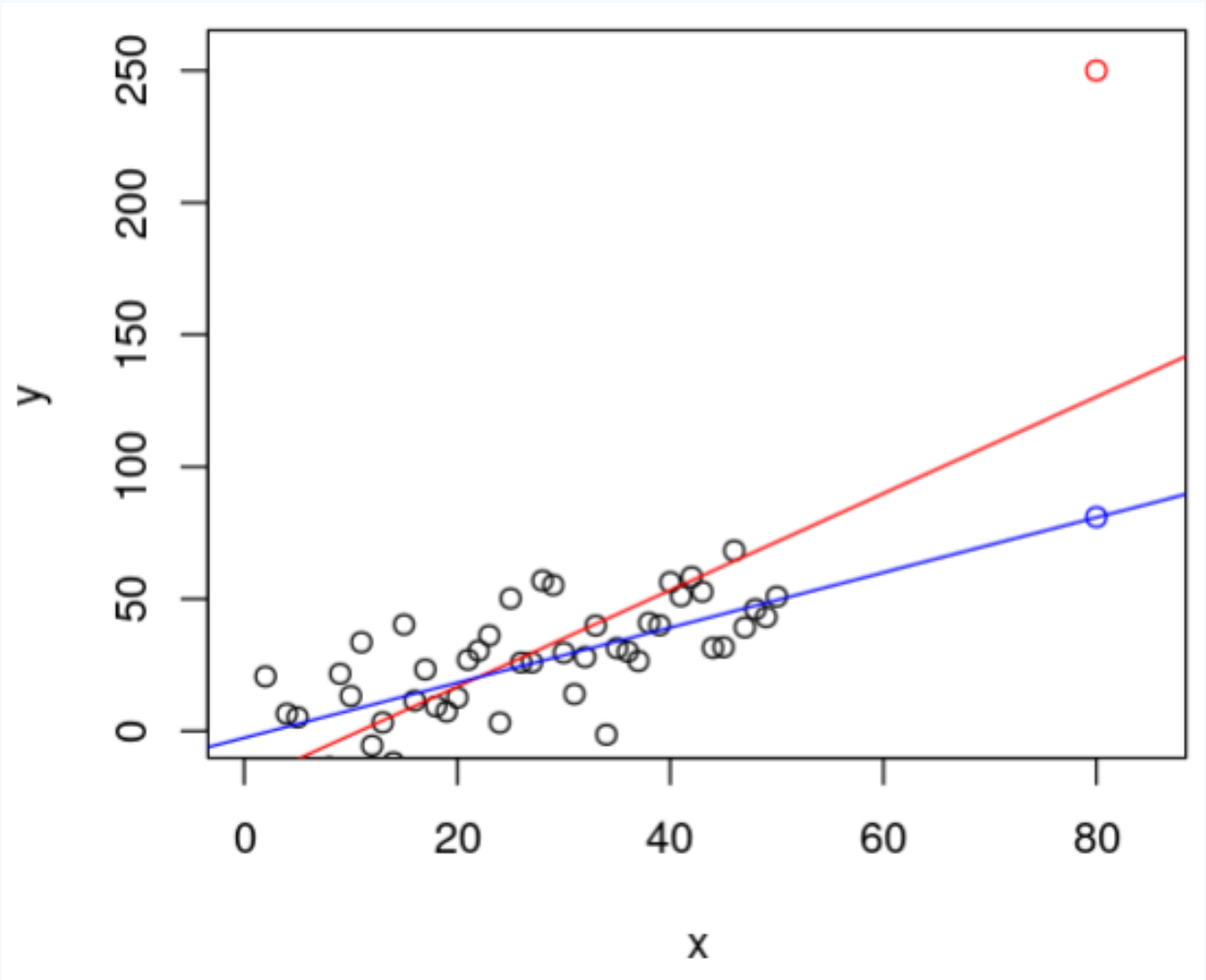
```
Rows: 344
Columns: 8
$ species           <fct> Adelie, Adelie, Adelie, Adelie, Adel
$ island            <fct> Torgersen, Torgersen, Torgersen, Tor
$ bill_length_mm    <dbl> 39.1, 39.5, 40.3, NA, 36.7, 39.3, 38
$ bill_depth_mm     <dbl> 18.7, 17.4, 18.0, NA, 19.3, 20.6, 17
$ flipper_length_mm <int> 181, 186, 195, NA, 193, 190, 181, 19
$ body_mass_g       <int> 3750, 3800, 3250, NA, 3450, 3650, 36
$ sex               <fct> male, female, female, NA, female, ma
$ year              <int> 2007, 2007, 2007, 2007, 2007, 2007,
```

Datos atípicos (outliers)

- Datos que no están en consonancia con el resto, que destaca por ser distinto del resto
- Causas:
  - Errores de medición (humano o del sistema). Ej: Peso de un paciente: 800 kg o un medidor manipulado
  - Contaminación: la muestra contiene datos de una población distinta a la de interés
  - Desviaciones naturales
- Solo se modifica el dato si es un error. Se busca el valor real y, si no es posible, se pone como missing

Datos atípicos

¿Diferencia entre el valor atípico rojo y el azul?



Datos faltantes (missing)

- Dato vacío, dato perdido, NA
- Causas:
  - Error en la medición, la transcripción
  - No se puede lograr el dato
- Acciones:
  - Trabajar únicamente con los datos sin valores faltantes (representan un % bajo del total de los datos)
  - Imputación de missing (media o mediana de la variable, el valor de los puntos más similares, predicción de un modelo de ML)
  - Agrupar los missings en una nueva categoría ¡fácilmente distinguible! Ej: 9999

Tabla de frecuencias

- **Tabla de frecuencias** o **tabla de contingencia**: Muestra, para cada valor que tome una variable categórica (o cada intervalo en una categorizada), o para cada combinación de valores de dos o más variables categóricas, el número de casos que aparecen con dicho valor o combinación de valores

```
table(penguins$species)
```

```
   Adelie Chinstrap    Gentoo
      152        68       124
```

```
prop.table(table(penguins$species))
```

```
   Adelie Chinstrap    Gentoo
0.4418605 0.1976744 0.3604651
```

Tabla de frecuencias

Dada una variable X con dominio $\{X_1, \ldots, X_k\}$ que ha sido medida en una muestra de tamaño n y denotemos por n_i el número de elementos en la muestra que toman el valor X_i. La tabla de frecuencias correspondiente es:

Valores de X	**Frec. absoluta**	**Frec. relativa**	**Frec. acumulada**	**Frec. relativa acumulada**
X_1	n_1	$fr_1 = \frac{n_1}{n}$	$F_1 = n_1$	$Fr_1 = \frac{F_1}{n}$
X_2	n_2	$fr_2 = \frac{n_2}{n}$	$F_1 = F_1 + n_2$	$Fr_2 = \frac{F_2}{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
X_k	n_k	$fr_k = \frac{n_k}{n}$	$F_k = F_{k-1} + n_k$	$Fr_k = \frac{F_k}{n}$

Tabla de frecuencias

¿Para qué tipos de variable sirve?

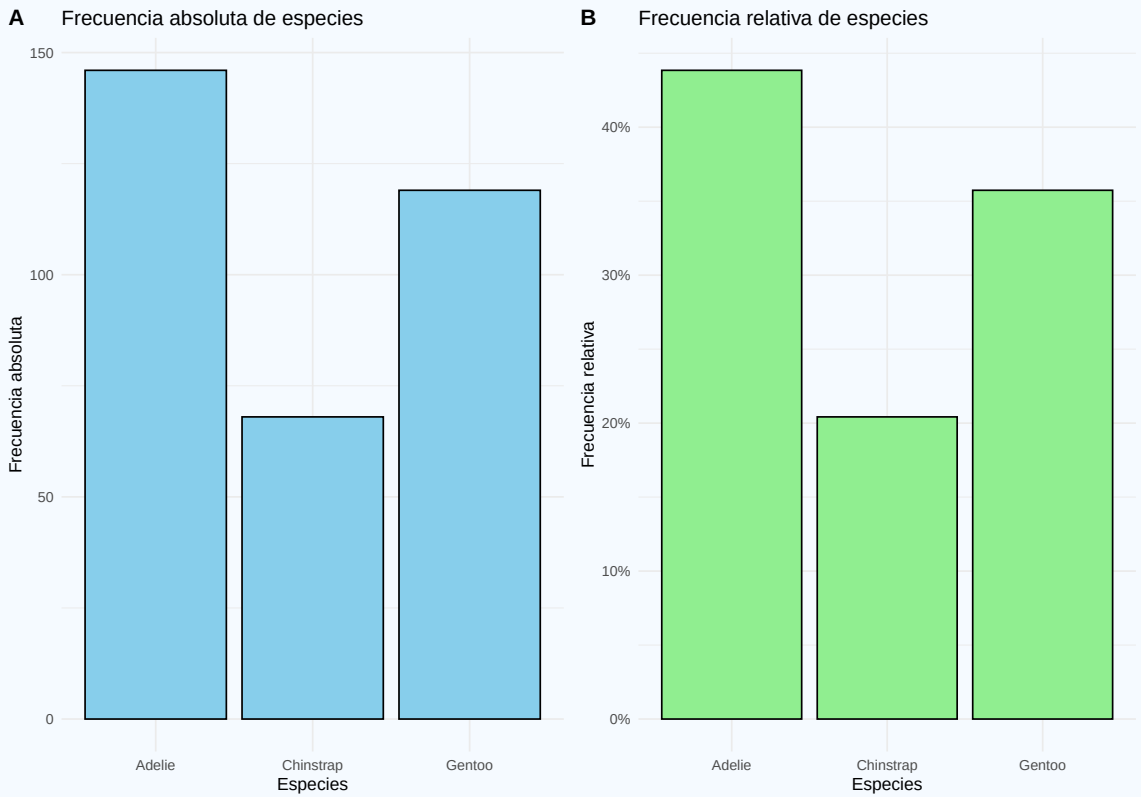
Tabla de frecuencias

¿Para qué tipos de variable sirve?

- Categóricas nominales. Color de ojos.
- Categóricas ordinales. Grado de satisfacción
- Continuas. Altura.
  - Dividimos en intervalos: (0,140],(140,155],...,(210,225]

Diagrama de barras

- Para variables cualitativas



No hay diferencia entre hacer la comparación con las frecuencias relativas o absolutas en este caso

Diagrama de barras

¿Y si quisiéramos comparar a nuestros pingüinos con un grupo de otra isla?

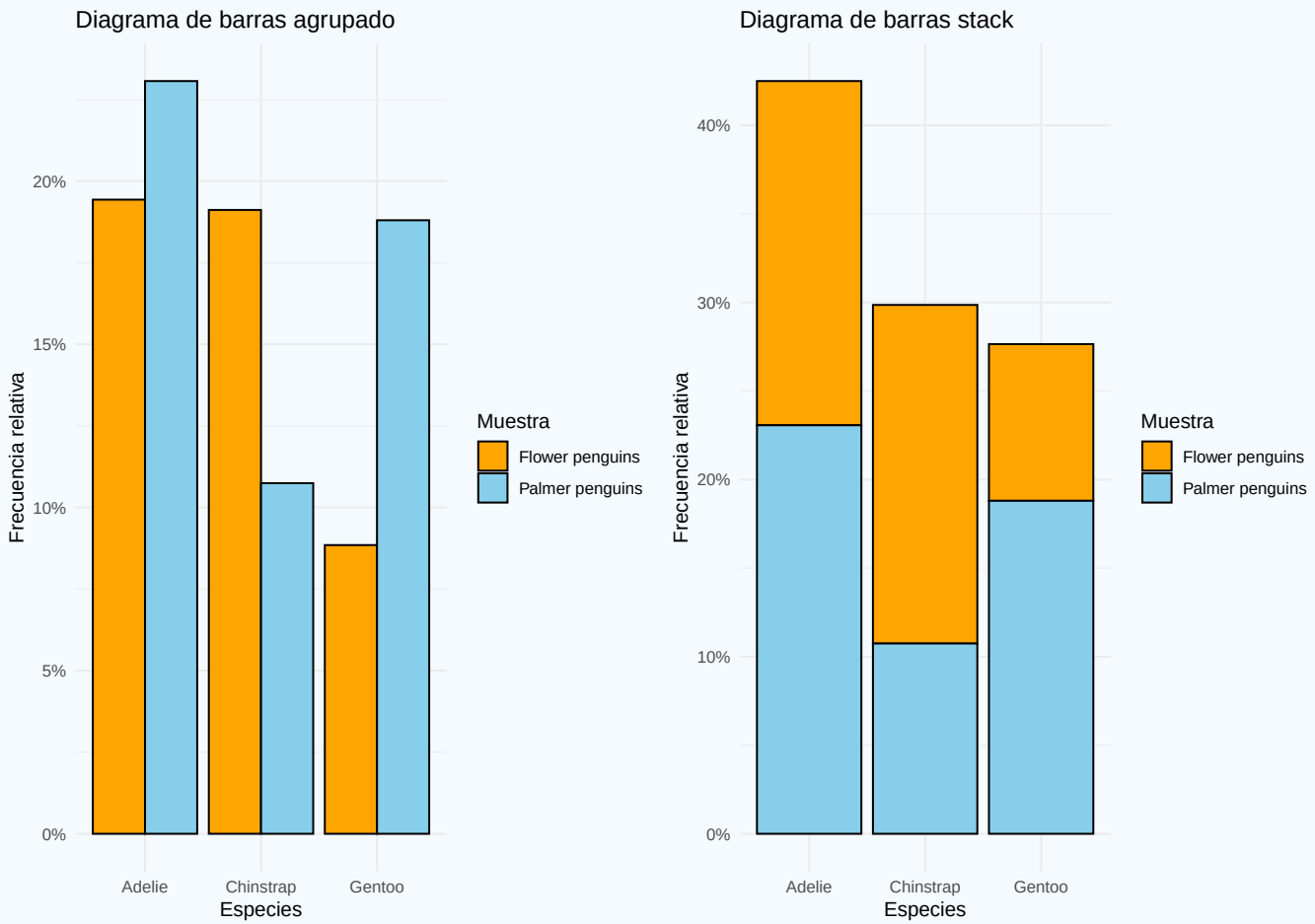


Tabla de frecuencias (contingencia)

	female	male
Adelie	73	73
Chinstrap	34	34
Gentoo	58	61

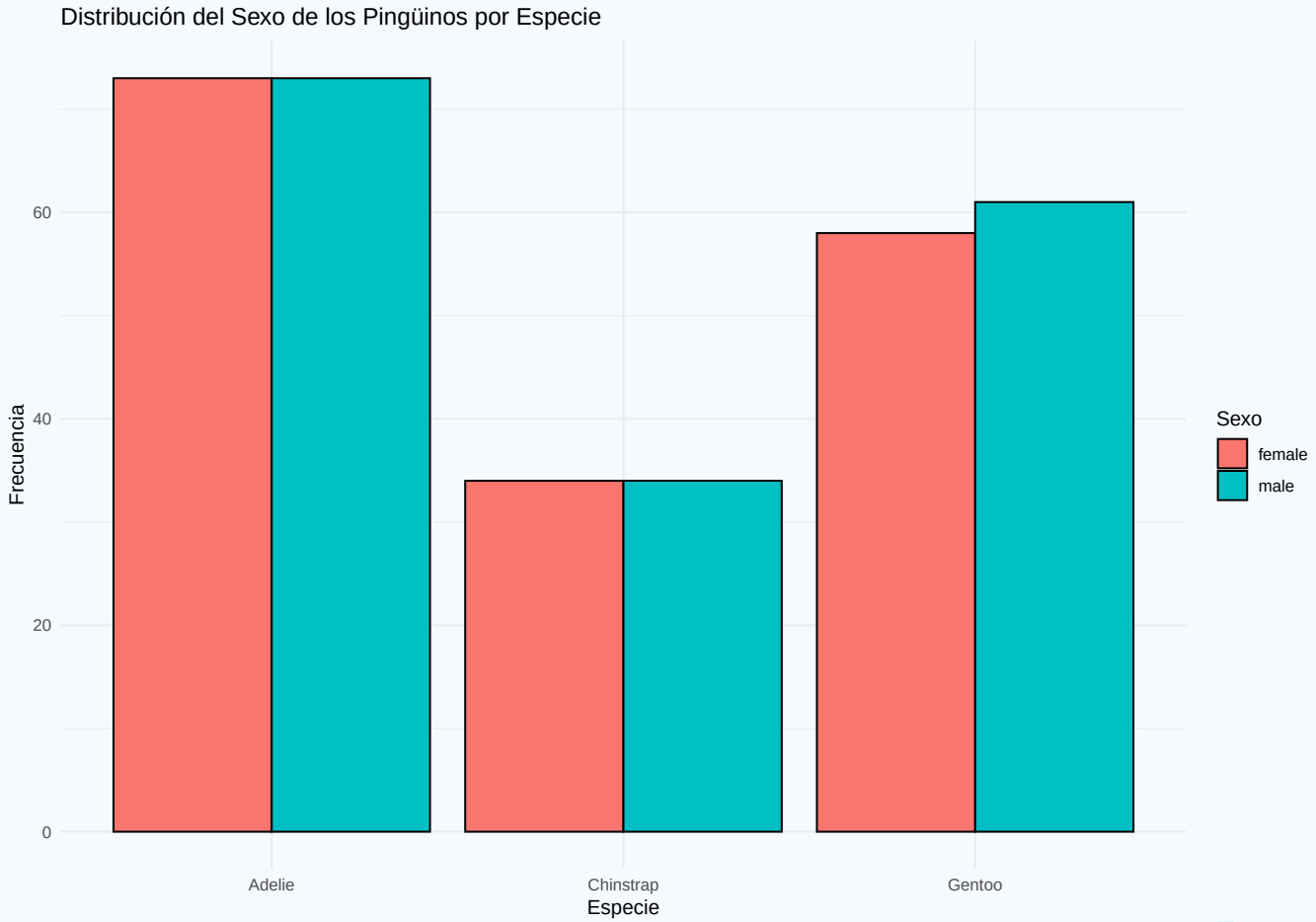
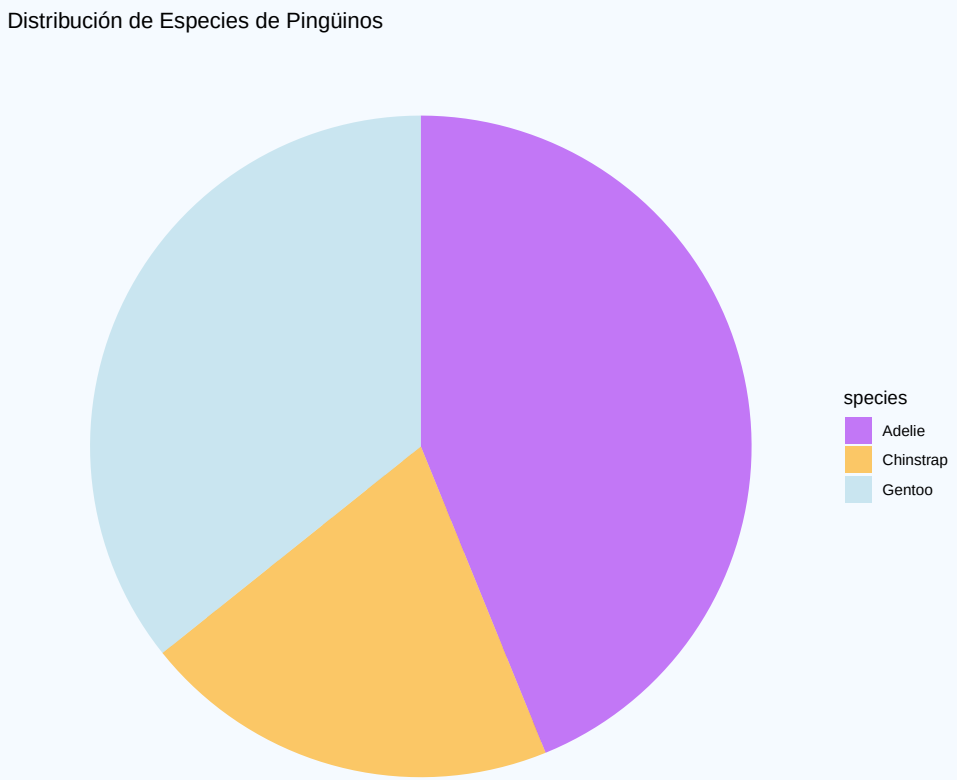


Gráfico de tarta

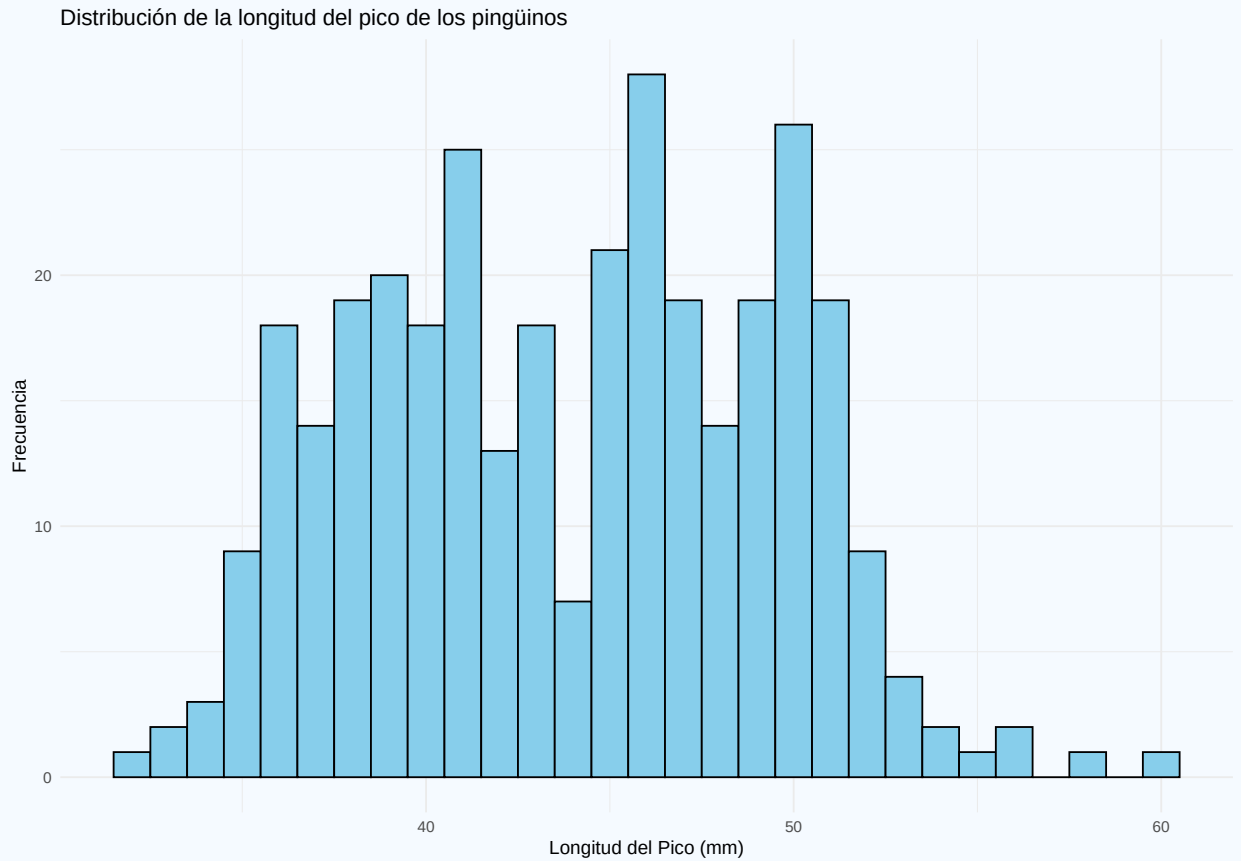
- Para variables cualitativas
- También llamado diagrama de sectores



Problema: ojo humano tiene problemas para percibir correctamente diferencias en sectores angulares

Histograma

- Para variables cuantitativas
- Refleja la distribución de los datos



¿Cómo se construye el histograma?

Resumen numérico de variables

- Medidas de centralidad
- Medidas de posición
- Medidas de dispersión

Medidas de centralidad

- **Moda**: valor más frecuente de la distribución
- **Media**. Dada una muestra de n observaciones $\mathbf{x} = (x_1, \ldots, x_n)$ de la variable X su media es

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

Muy afectada por valores atípicos/extremos

Medidas de centralidad

- **Mediana**: valor que ocupa la posición central de los datos, i.e., deja el 50% de los puntos a su izquierda (por debajo de él) y el otro 50% a la derecha (por encima de él). Sea $\mathbf{x}$ una muestra con n observaciones, ordenados de menor a mayor, entonces:
  - Si n es impar, la mediana es justamente el valor que ocupa justamente la posición central $\lfloor n/2 \rfloor + 1$, $Med(\mathbf{x}) = x_{(\lfloor n/2 \rfloor + 1)}$
  - Si n es par, la mediana será la media de los dos valores centrales, esto es, $Med(\mathbf{x}) = \frac{x_{(n/2)} + x_{(n/2+1)}}{2}$

  Medida robusta

Media, moda, mediana

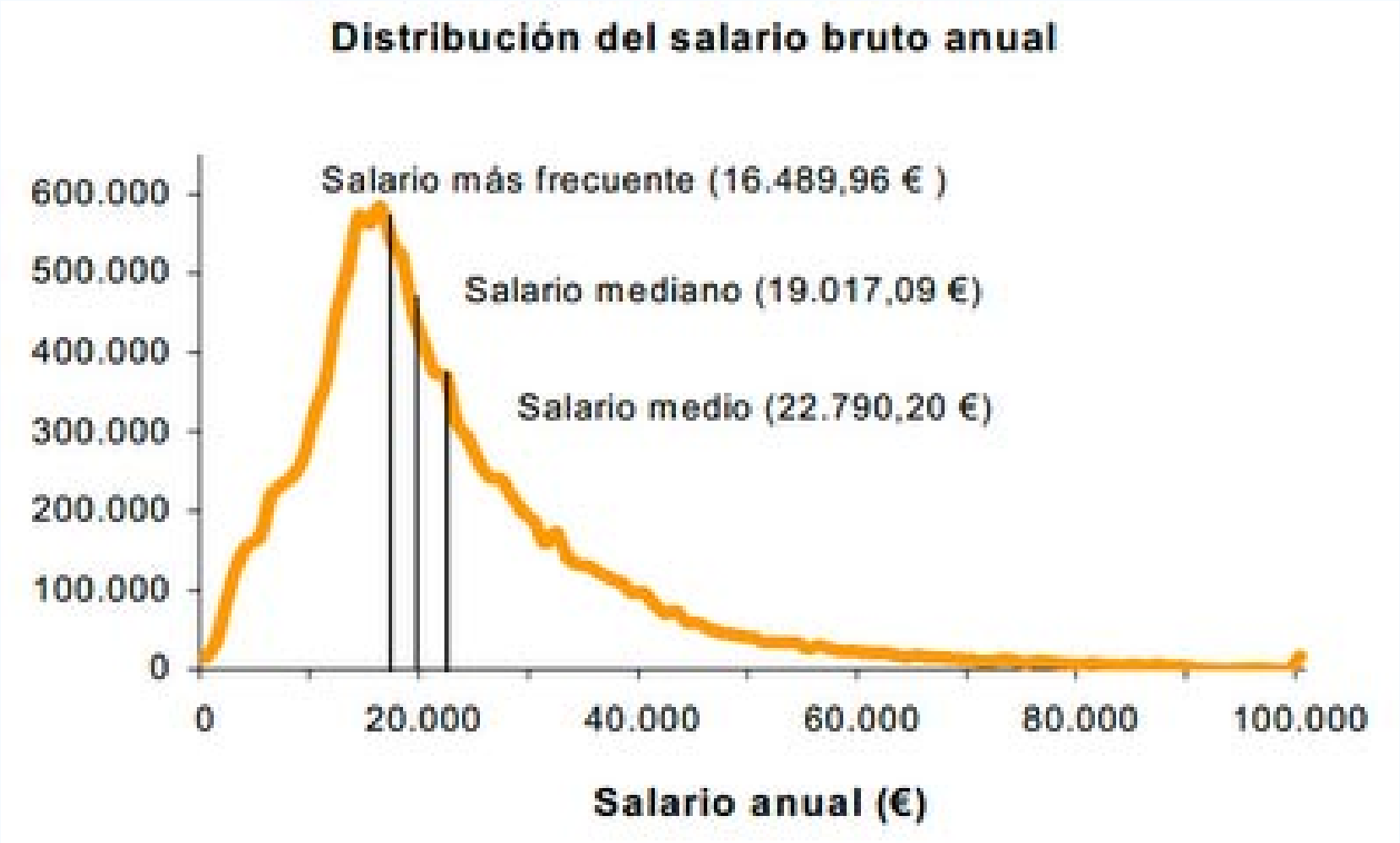


Figure 1: Microsiervos

Resumen numérico de variables

Medidas de posición

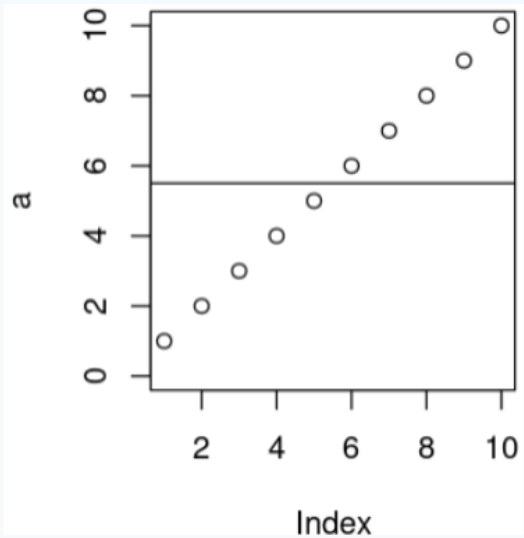
- Valores **mínimo** y **máximo** de la variable: x_{min}, x_{max}
- Primer y tercer **cuartil**: Los valores que dejan por debajo un $p\%$ de los datos, siendo $p = 25\%$ en el caso del primer cuartil (Q_1) y $p = 75\%$ en el caso del tercer cuartil (Q_3). El segundo cuartil es la mediana.
- **Deciles**: Mismo concepto que los cuartiles pero de 10 en 10

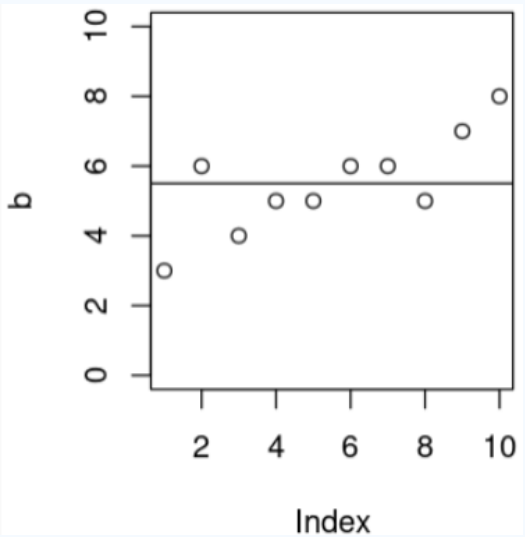
Resumen numérico de variables

Medidas de dispersión: ¿Cómo varían los datos en torno a los valores centrales?

```
a <- c(1, 2, 3, 4, 5, 6, 7, 8 , 9, 10)
b <- c(3, 6, 4, 5, 5, 6, 6, 5, 7, 8)
cat("Media a: ", mean(a), ', Desviación típica a: ', round(sd(a),2), '\n',
             "Media b: ", mean(b), ', Desviación típica b: ', round(sd(b),2))
```

```
Media a:  5.5 , Desviación típica a:  3.03
 Media b:  5.5 , Desviación típica b:  1.43
```





Medidas de dispersión

- **Rango** o **recorrido**: $Rango = x_{max} - x_{min}$
- **Varianza**: Mide la dispersión de los valores de la variable respecto a la media
  - Varianza **muestral**: $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - Varianza **poblacional**: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(x_i - \mu)^2$, siendo N el tamaño de la población y μ su media
- **Desviación típica**: raíz cuadrada de la varianza muestral o poblacional.
  - Desviación típica **muestral**: $s = \sqrt{s^2}$
  - Desviación típica **poblacional**: $\sigma = \sqrt{\sigma^2}$

  Interpretación más sencilla al medir la dispersión en las mismas unidades que la variable

Medidas de dispersión

- **Rango intercuartílico**: diferencia entre el tercer y el primer cuartil $IQR = Q_3 - Q_1$
- **Coeficiente de variación**: representa la desviación típica en unidades de la media $CV = s/\bar{x}$. Se suele expresar en porcentaje. Por ejemplo, CV $= 60\%$ indica que el valor de la desviación típica es 0.6 veces la magnitud de la media.

Diagrama de cajas y bigotes (boxplot)

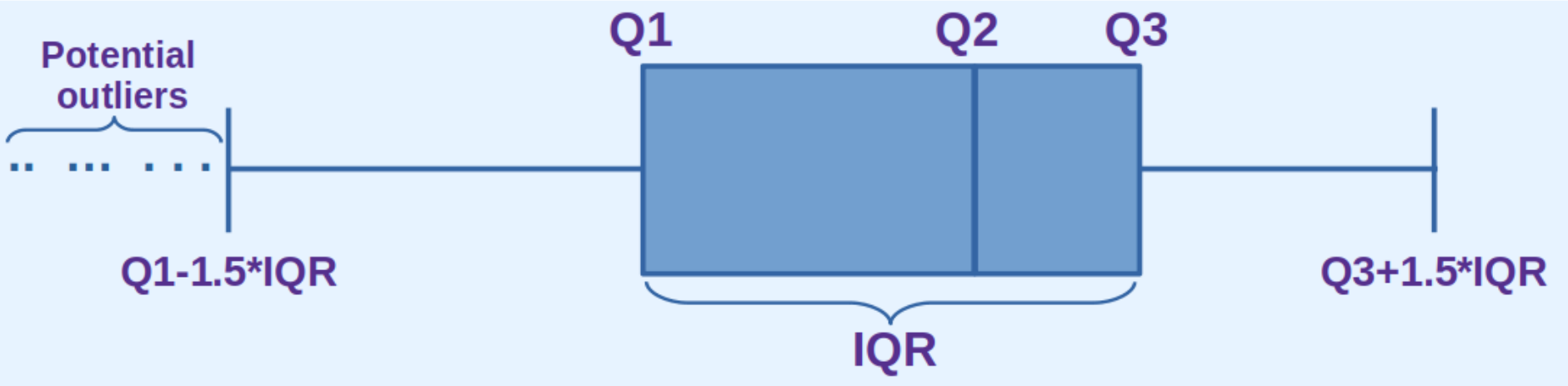
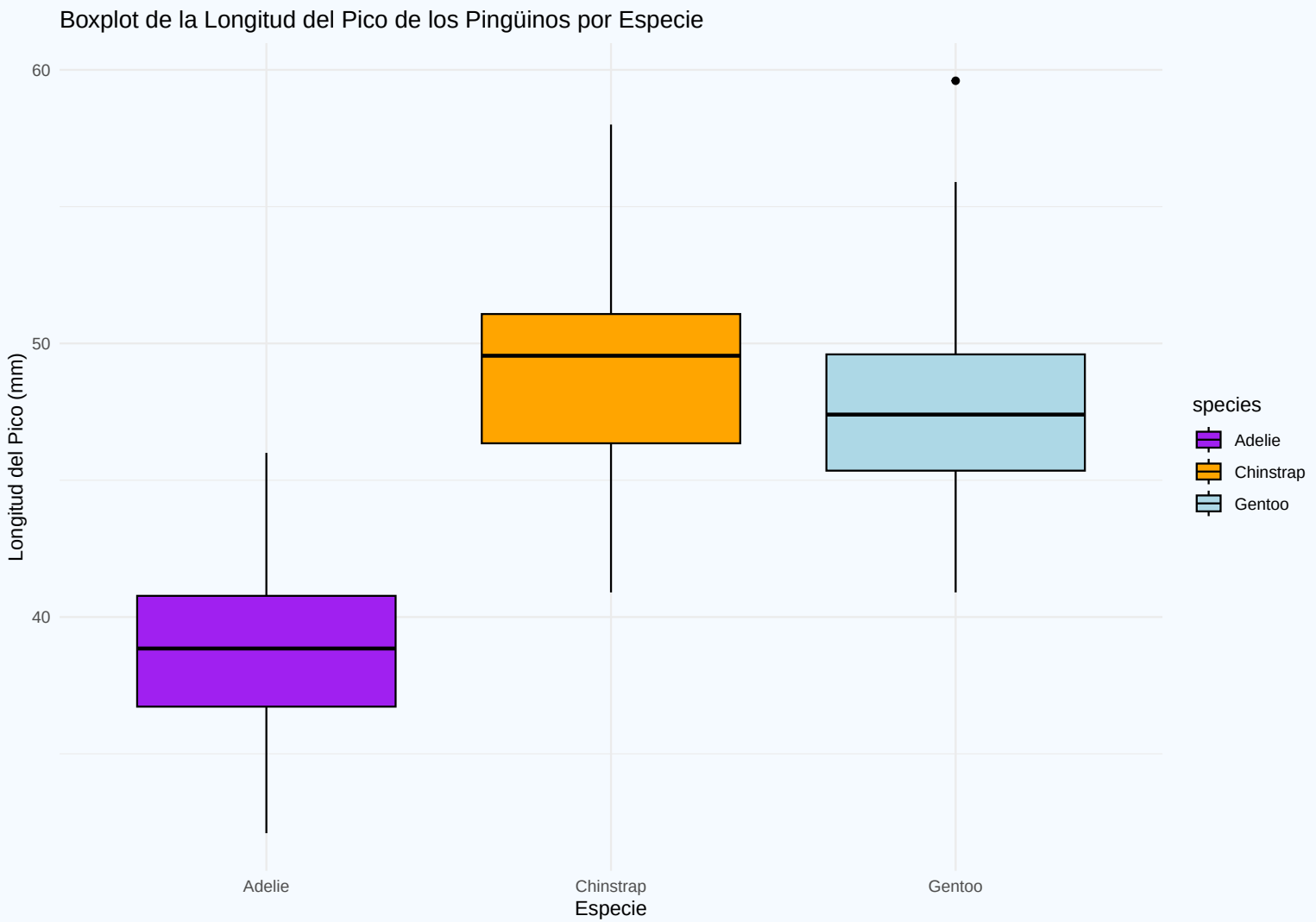


Diagrama de cajas y bigotes (boxplot)



Diagramas de dispersión

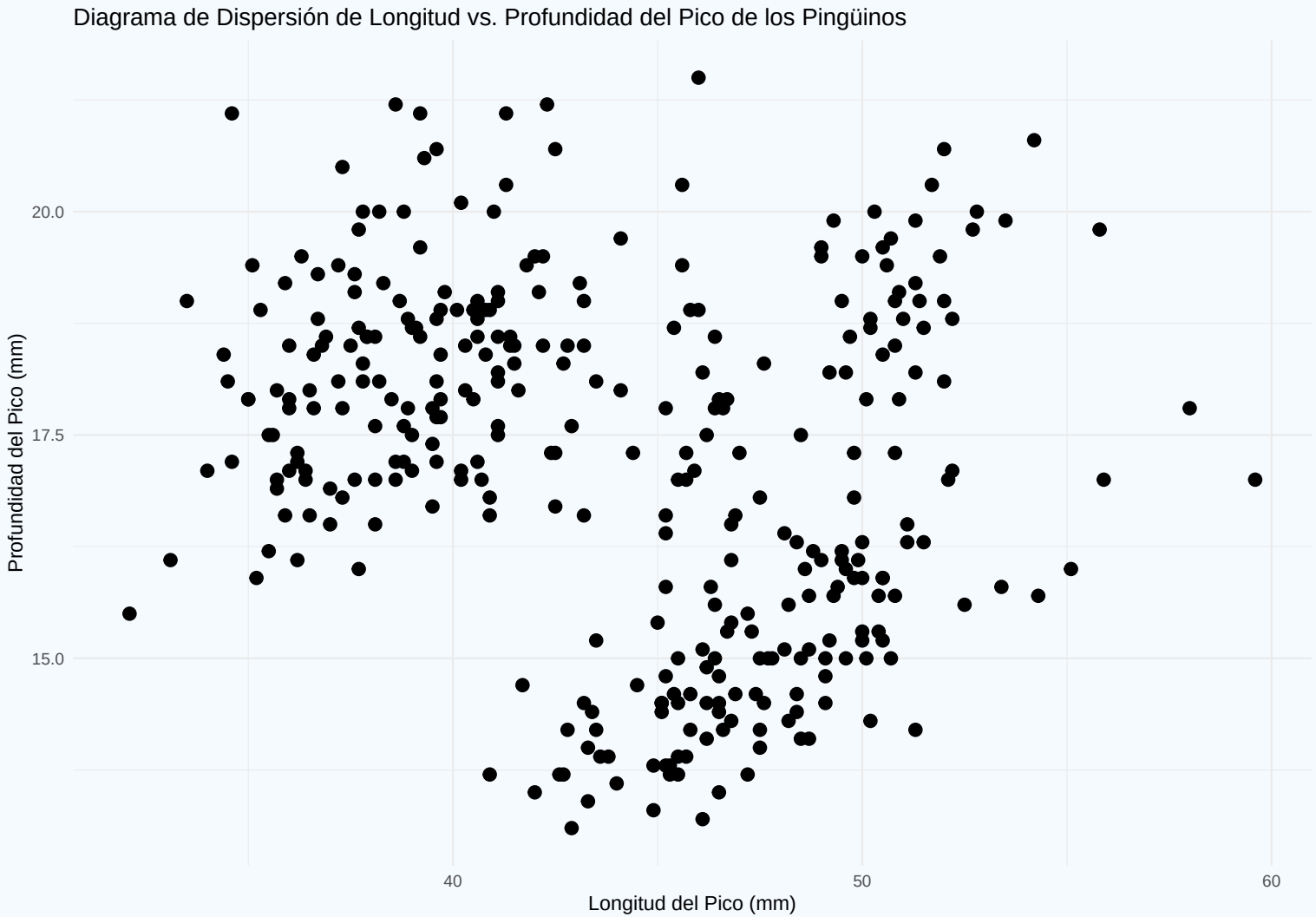
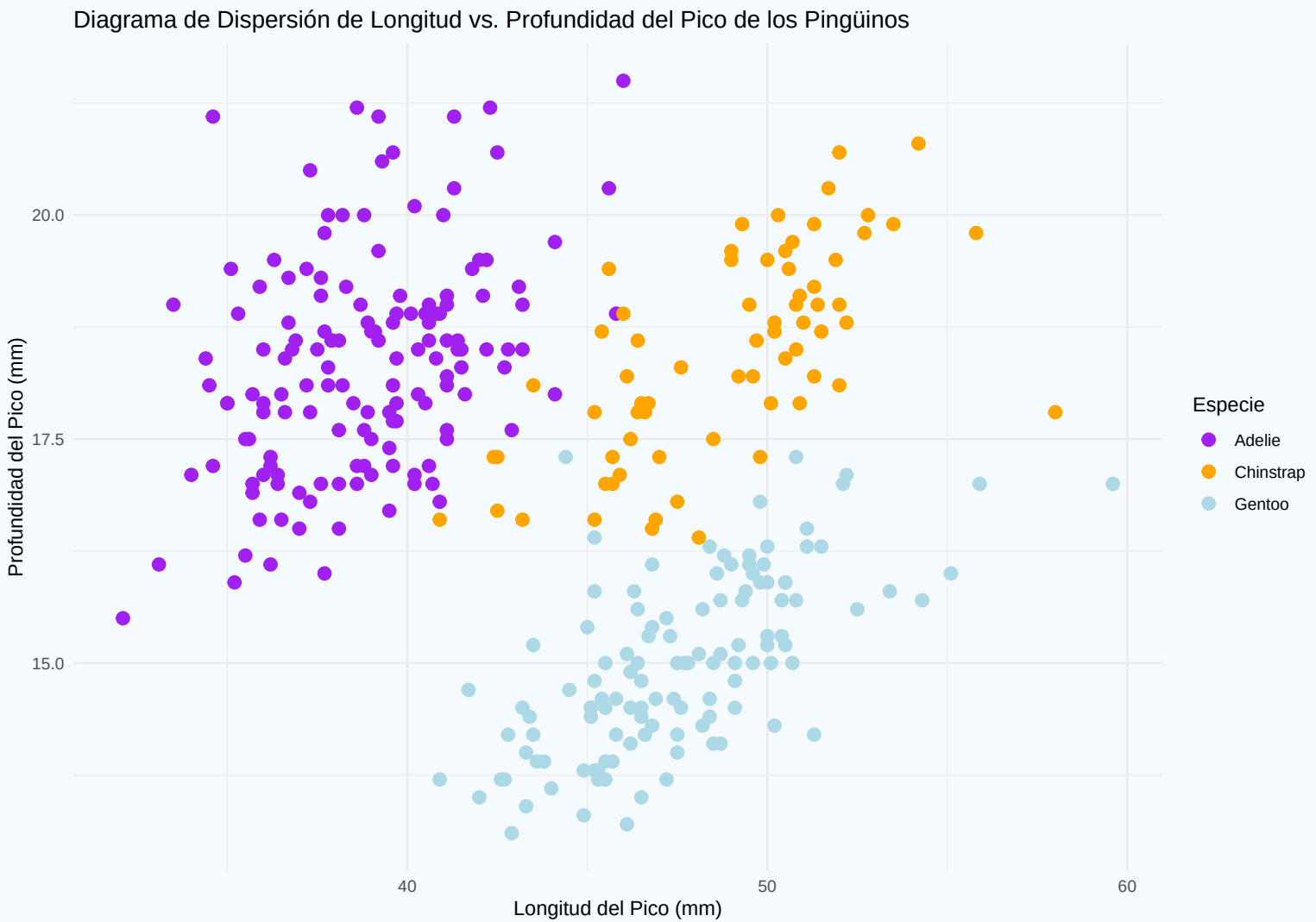


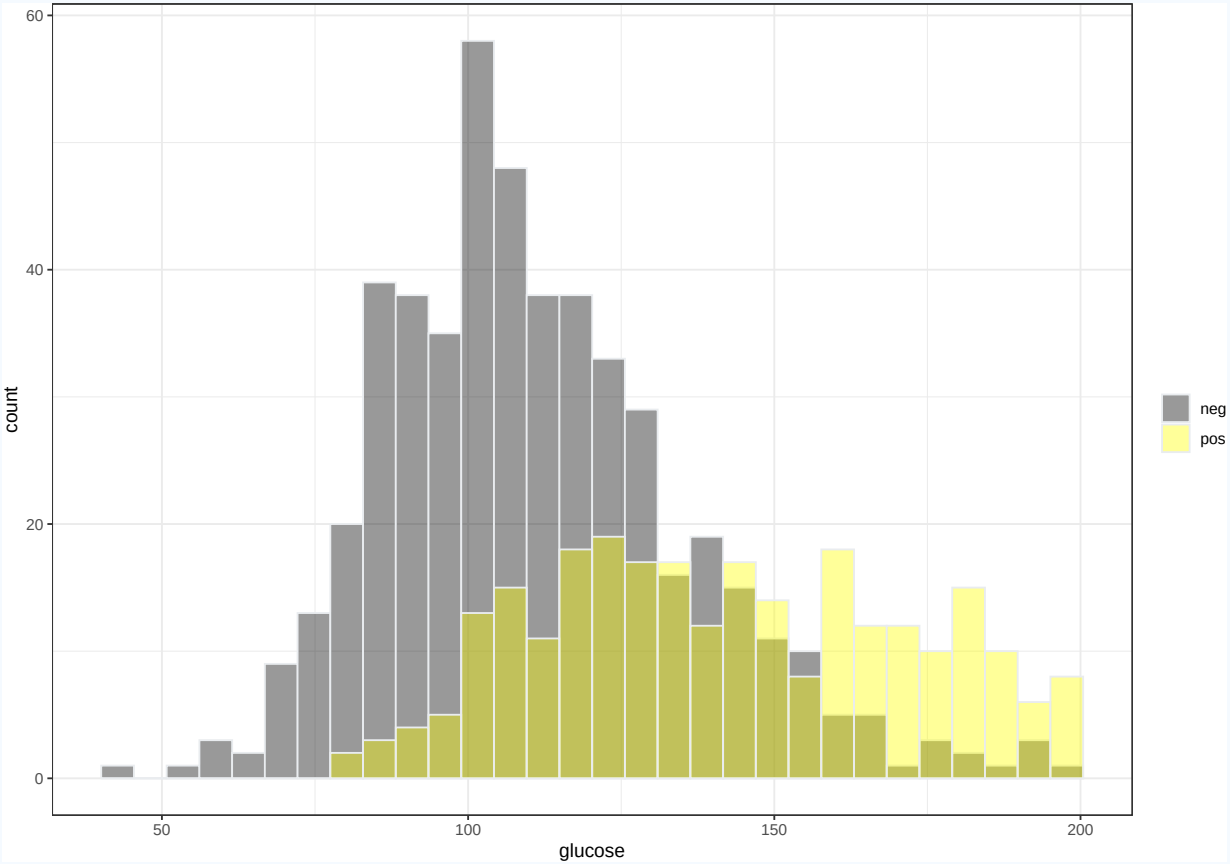
Diagrama de dispersión



Más y más y más

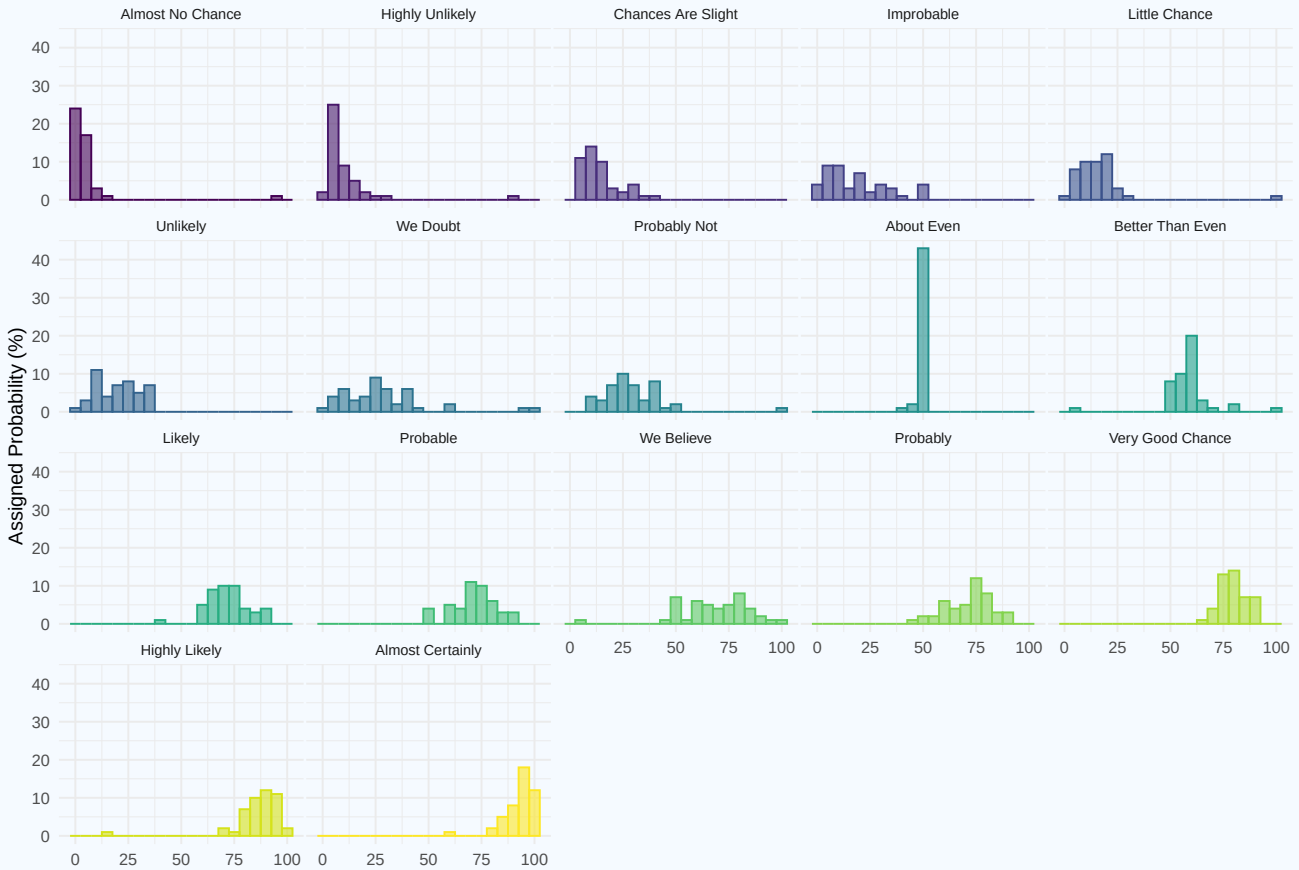
- R graph gallery `https://r-graph-gallery.com/`
- R Gallery book `https://bookdown.org/content/b298e479-b1ab-49fa-b83d-a57c2b034d49/`
- ¿El mejor gráfico hecho? The Minard map `https://bigthink.com/strange-maps/229-vital-statistics-of-a-deadly-campaign-the-minard-map/`

Histogramas conjuntos

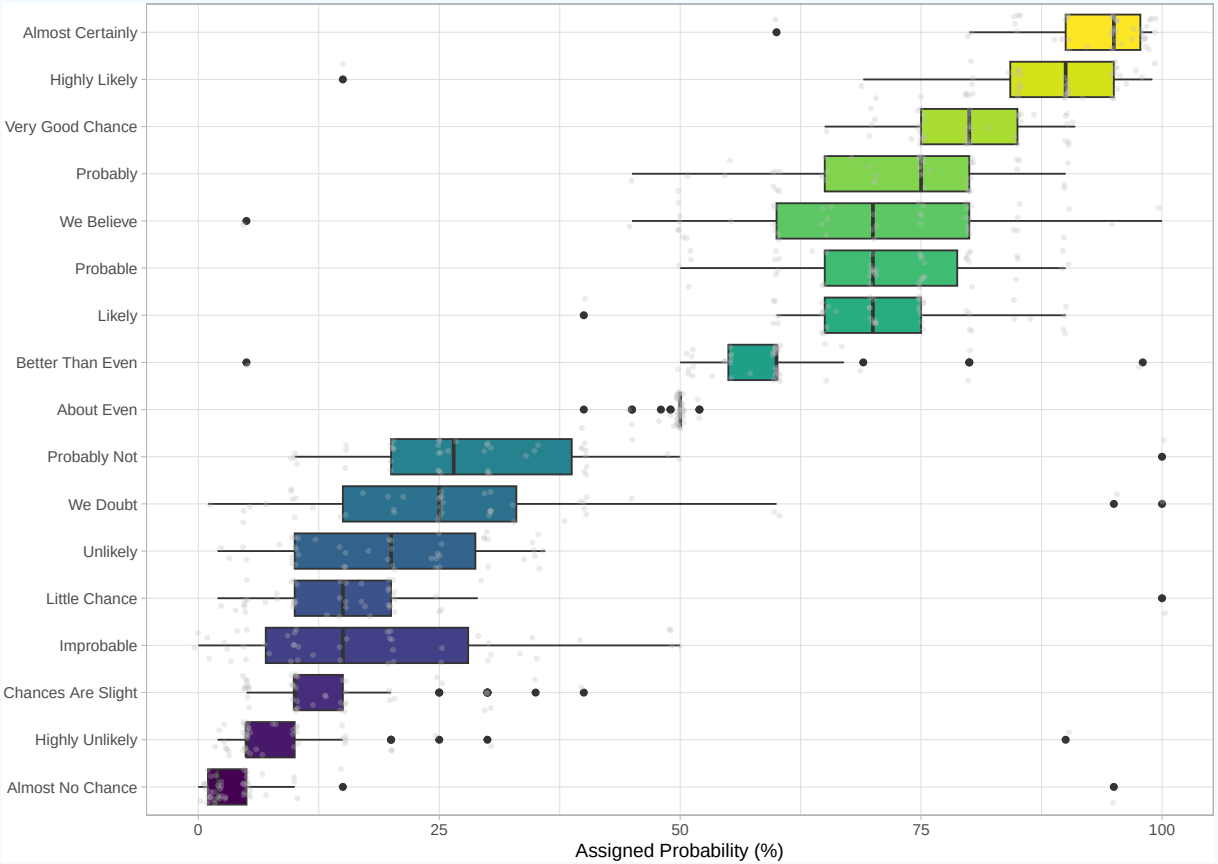


Histogramas conjuntos

Origen del gráfico: From Data to Viz

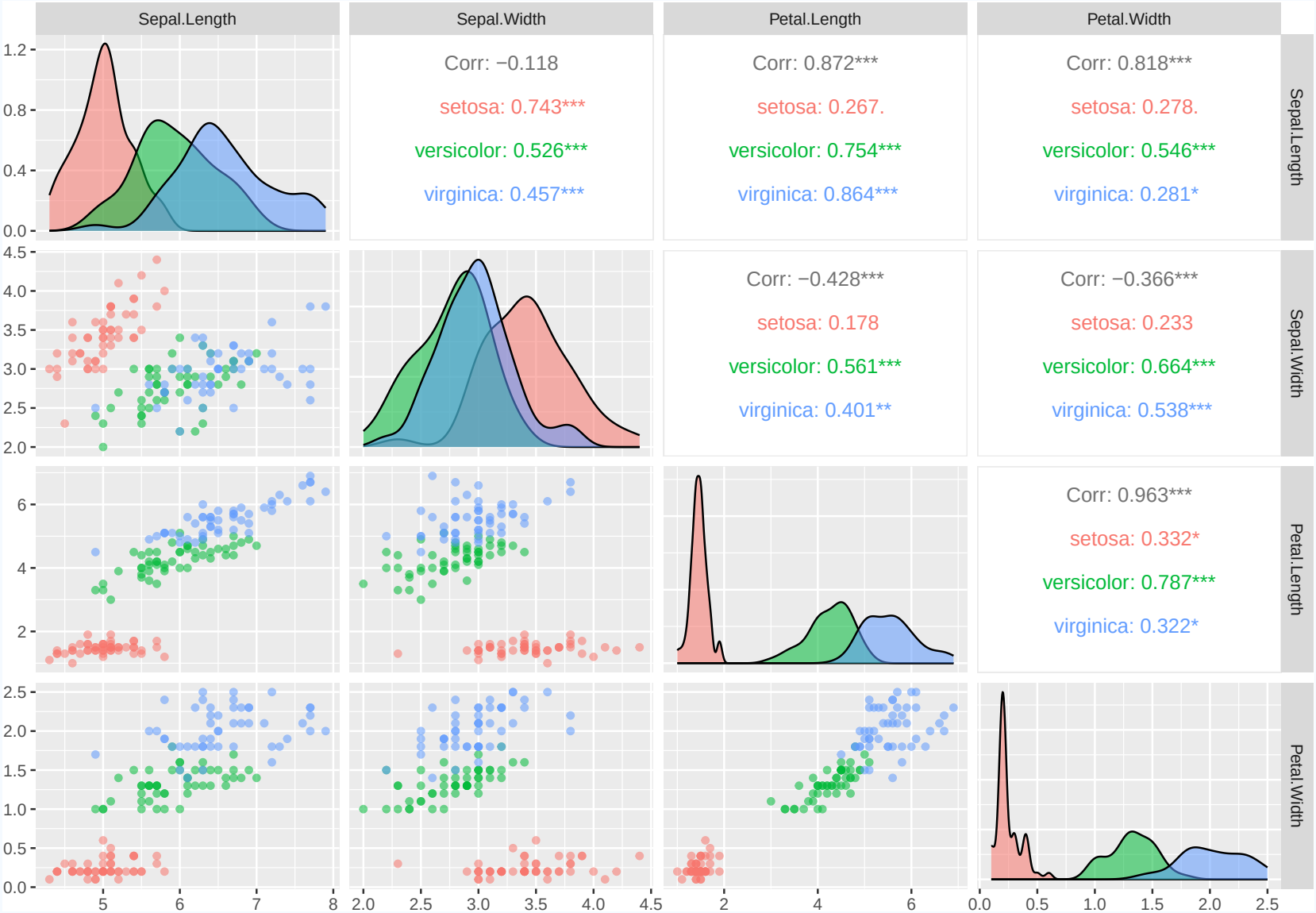


Origen del gráfico: From Data to Viz



ggpairs

https://r-charts.com/es/correlacion/ggpairs/



Más y más

- Gráficos multivariantes
- Gráficos de correlación
- Series temporales
- Mapas
- Pirámides de población
- QQplot
- etc

Más y más

https://elartedeldato.com/

https://rkabacoff.github.io/datavis/ Modern Data Visualization with R

https://r-graph-gallery.com/ggplot2-package.html

https://r-graph-gallery.com/

https://www.data-to-viz.com/

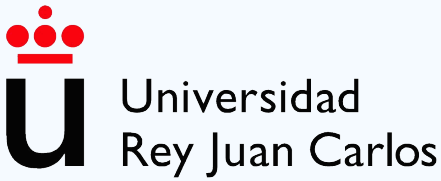
Bibliografía de utilidad para el lector

- "Fundamentos de ciencia de datos con R" coordinado por Gema Fernández-Avilés y José-María Montero: https://cdr-book.github.io/
- Weiss, N. A., & Weiss, C. A. (2017). *Introductory statistics*. London: Pearson.
- "Estadística Aplicada a las Ciencias y la Ingeniería" escrito por Emilio L. Cano. https://emilopezcano.github.io/estadistica-ciencias-ingenieria/index.html
- R for Data Science: https://r4ds.hadley.nz/eda
  - Primera versión en castellano: https://es.r4ds.hadley.nz/

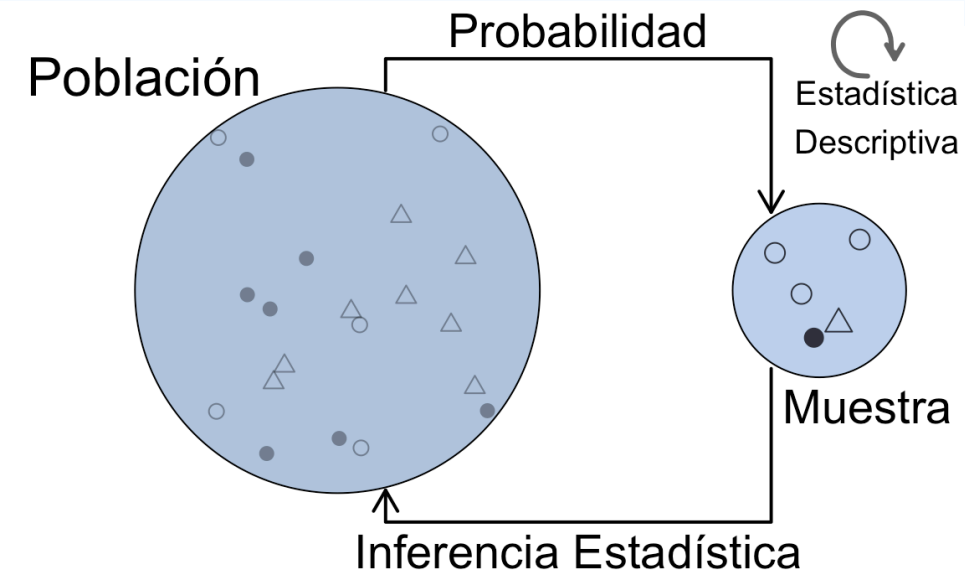
Inferencia Estadística. Estimación puntual

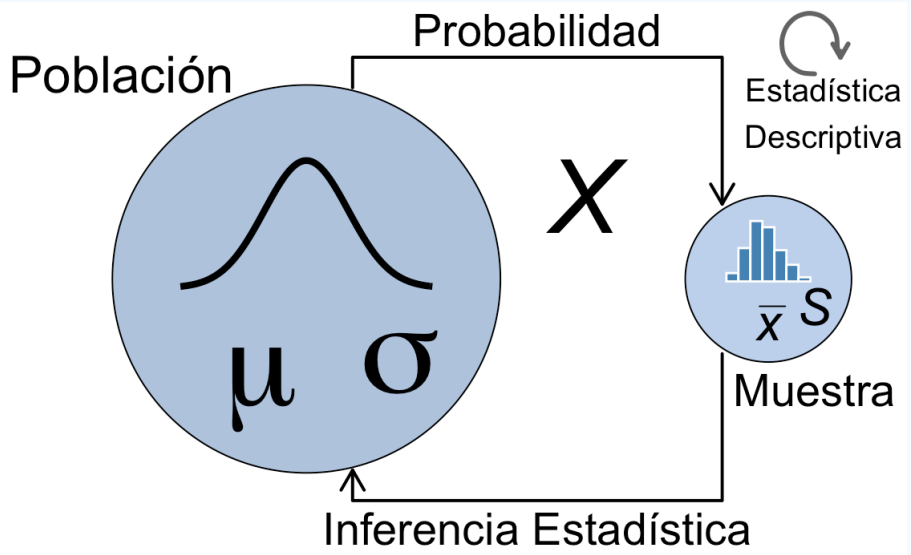
Inferencia Estadística - Grado en Ciencia e Ingeniería de Datos

Curso académico 2024-2025



Un poco de repaso





- X es una variable aleatoria (¡de interés!) cuya función de distribución es conocida a excepción de determinados paramétros (Por ejemplo, μ y σ en la distribución Normal)
- "Conocidos" (estimados) los paramétros (por ejemplo, $\overline{x}$ y S para la Normal) –> Función de distribución de probabilidad totalmente determinada y ¡a extraer información y sacar conclusiones!

Ejemplo Instagram

- Sabemos que el número de horas al día que pasan las personas en Instagram sigue una distribución Normal. Estamos interesados en conocer la media de horas.
- Variable aleatoria X: horas que pasan al día los estudiantes de la URJC en Instagram
- Población: Estudiantes de la URJC
- Para estimar la media μ, se pregunta a 10 estudiantes al azar el número de horas que pasan al día $x_1, \ldots, x_{10}$:

$$2.5, 3.0, 1.5, 4.0, 2.0, 3.5, 2.8, 1.0, 4.5, 3.2$$

Inferencia estadística paramétrica

- X variable aleatoria (v.a.) con función de distribución F_θ conocida, que depende de un parámetro $\theta \in \Theta \subseteq \mathbb{R}$ desconocido, siendo Θ su espacio paramétrico
- El objetivo de la inferencia estadística es lograr buenas estimaciones/aproximaciones del parámetro θ en base a una muestra aleatoria simple (m.a.s.) $X_1, \ldots, X_n$ de la población. $X_i, i = 1, \ldots, n$ son v.a. independientes e igualmente distribuidas F_θ
- Con cada observación $x_1, \ldots, x_n$ de la muestra aleatoria simple se puede hacer una estimación del parámetro desconocido θ –> estadístico muestral o estimador

Estadístico muestral

- Formalmente, dada una variable aleatoria X con función de distribución F_θ, con θ desconocido, un estadístico o estimador $T = T(X_1, \ldots, X_n)$ es una función real de la m.a.s. $(X_1, \ldots, X_n)$ que estima el valor del parámetro desconocido

$$T = T(X_1, \ldots, X_n) = \hat{\theta}$$

- Ejemplos:
  - Media, mediana
  - Varianza, desviación estándar
- Un estadístico es una variable aleatoria, y por lo tanto, tiene asociada una distribución que se denomina **distribución muestral**

Dado el mismo estadístico, la aplicación del mismo a distintas muestras concretas dará lugar a diferentes **estimaciones**

Ejemplo Instagram:

- Estadístico media muestral $\overline{X}$
- Muestra A: $2.5, 3.0, 1.5, 4.0, 2.0, 3.5, 2.8, 1.0, 4.5, 3.2$ –> $\overline{x}_A = 2.8$
- Muestra B: $3.1, 2.7, 1.8, 4.2, 3.3, 2.4, 1.6, 4.0, 3.5, 2.9$ –> $\overline{x}_B = 2.95$
- Muestra C: $2.9, 3.6, 1.4, 4.1, 2.3, 3.0, 1.9, 4.4, 3.8, 2.2$ –> $\overline{x}_C = 2.96$

La media muestral varía en las distintas muestras –> es una variable aleatoria con una distribución de probabilidad. Caracterizaremos dichas distribuciones muestrales mediante su media y su varianza.

Ejemplo Instagram: estimadores

- Media: $T_1(X_1, \ldots, X_n) = \overline{X} = \frac{(X_1 + \cdots + X_n)}{n}$
- Mediana: $T_2(X_1, \ldots, X_n) = \frac{X_{(n/2)} + X_{(n/2+1)}}{2}$
- Media del máximo y el mínimo: $T_3(X_1, \ldots, X_n) = \frac{max(X_1,\ldots,X_n) + min(X_1,\ldots,X_n)}{2}$
- etc

Distintos estadísticos pueden estimar el valor del mismo parámetro

No todos los estadísticos sirven para estimar el mismo parámetro

Ejemplo Instagram: estimaciones

La aplicación de cada estimador a la muestra ofrece una estimación diferente de la media poblacional μ:

- Media:
  $T_1(x_1, \ldots, x_{10}) = \overline{x} = \frac{(2.5+3.0+1.5+4.0+2.0+3.5+2.8+1.0+4.5+3.2)}{10} = 2.8$
- Mediana: $T_2(x_1, \ldots, x_{10}) = \frac{2.8+3.0}{2} = 2.9$
- Media del máximo y el mínimo: $T_3(x_1, \ldots, x_{10}) = \frac{4.5+1}{2} = 2.75$

Así, se obtienen 3 estimaciones de la media poblacional: $\hat{\mu}_1 = 2.8$, $\hat{\mu}_2 = 2.9$, $\hat{\mu}_3 = 2.75$

Error de los estimadores

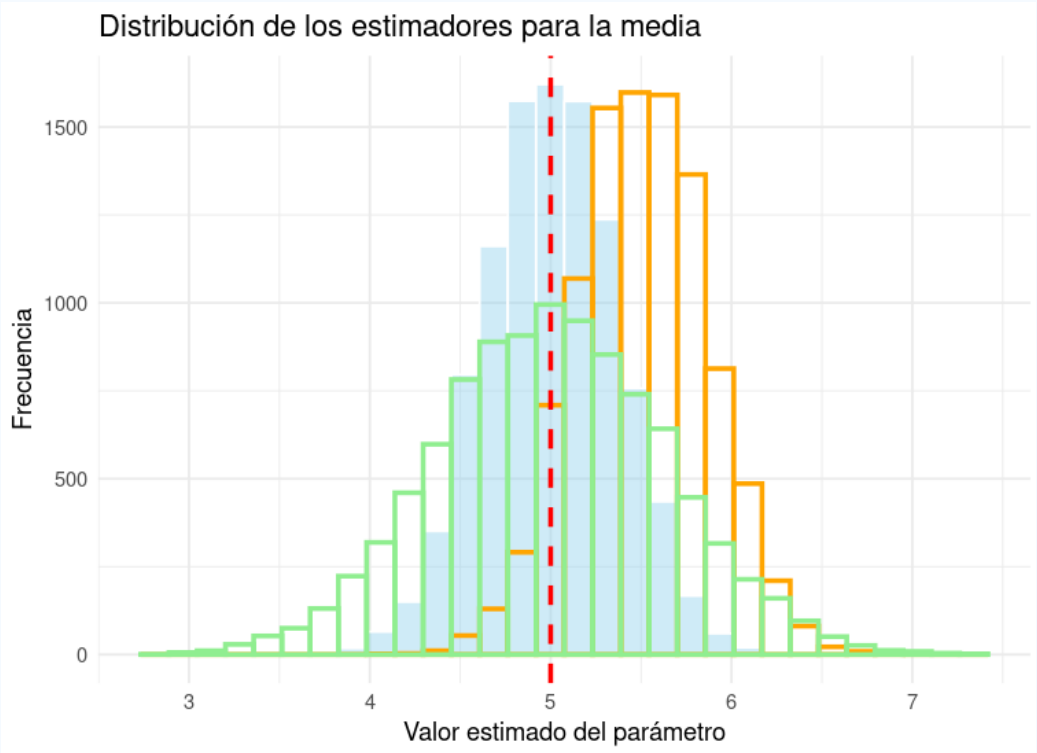
- En el tema 1 dijimos: $DATOS = MODELO + ERROR$
- Particularmente: $ERROR = |T(X_1, \ldots, X_n) - \theta|$
- Queremos que el error sea lo mínimo posible, pero desconocemos θ
  –> Pedimos ciertas propiedades a los estimadores para que sean adecuados para estimar un parámetro

Propiedades de los estimadores

Supongamos una población Normal con media conocida $\mu = 5$

Se obtiene una muestra de 10000 observaciones y se calculan los 3 estadísticos que se muestran a continuación (azul, naranja y verde).

¿Qué opinas?



Propiedades de los estimadores

- **Insesgadez**. Un estimador es insesgado (o centrado) si, en promedio, coincide con el valor verdadero del parámetro que se estima. Es decir, el valor esperado del estimador es igual al parámetro poblacional: $E[\hat{\theta} = T(X_1, \ldots, X_n)] = \theta$. Esto quiere decir que la distribución del estadístico muestral está centrada en el verdadero valor del parámetro
- **Eficiencia**. La varianza de un estimador debe ser lo más pequeña posible. Entre dos estimadores insesgados, el más eficiente es el que tiene menor varianza, es decir, el que proporciona estimaciones más precisas.

Propiedades de los estimadores

- **Consistencia**: Un estimador es consistente si, a medida que el tamaño de la muestra aumenta, la estimación se aproxima al valor verdadero del parámetro. Es decir:

$$lim_{n\to\infty} P(|\hat{\theta} - \theta| \geq \delta) = 0, \forall \delta > 0$$

  donde n es el tamaño muestral

- **Suficiencia**: Un estimador del parámetro θ es suficiente si utiliza toda la información contenida en la muestra sobre el parámetro que se está estimando. Es decir, si la distribución de la muestra $X_1, \ldots, X_n$ dado el estadístico T es independiente del valor del parámetro desconocido. Esto es, se tiene la misma información sabiendo el valor del estadístico T que conociendo todas las observaciones de la muestra

Repaso: Teorema Central del Límite

El Teorema Central del Límite (TCL) establece que, bajo ciertas condiciones, la distribución de la suma (o el promedio) de un número (lo suficientemente grande) de variables aleatorias independientes e identicamente distribuidas tiende a seguir una distribución Normal, independientemente de la distribución original de las variables.

Formalmente, el TCL establece que si, $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes e idénticamente distribuidas, con media μ y varianza $\sigma^2 < \infty$, entonces para n suficientemente grande se verifica

$$\overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

Distribuciones muestrales

- Usamos estadísticos para estimar los parámetros desconocidos
- Queremos que estos estadísticos tengan buenas propiedades
- Estos estadísticos son variables aleatorias con distribución de probabilidad

--> Vamos a estudiar algunos de los estadísticos más comunes, sus distribuciones y características

¡Ojo! Por el TCL, teniendo el tamaño muestral adecuado, sabremos que la distribución de los estadísticos será una Normal

Distribución media muestral

Dadas $X_1, \ldots, X_n$ variables aleatorias independientes e idénticamente distribuidas con media μ y varianza σ^2 conocida,

la media muestral

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n}$$

tiene las siguiente esperanza (media) y varianza:

- $E[\overline{X}] = \mu$
- $V[\overline{X}] = \frac{\sigma^2}{n}$

Entonces, con n suficientemente grande, $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

Distribución varianza muestral

La varianza muestral, que sirve para estimar la varianza poblacional, es:

$$V^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

es una variable aleatoria con media:

$E[V^2] = \frac{n-1}{n}\sigma^2$ –>¡No es un estimador insesgado de la varianza poblacional!

¿Ideas?

Distribución cuasivarianza muestral

La cuasivarianza muestral, que sirve para estimar la varianza poblacional, es:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

es una variable aleatoria con media:

$E[S^2] = \sigma^2$ –>Estimador insesgado de la varianza poblacional

$V[S^2] = \frac{2\sigma^4}{n-1}$

Estimación puntual

La **estimación puntual** es una técnica en estadística que consiste en utilizar los datos de una muestra para calcular un valor único, denominado **estimador puntual**, que se usa como mejor aproximación de un parámetro desconocido de la población. Nótese que la función de distribución de la población es conocida a excepción de dicho parámetro.

La estimación puntual proporciona una forma simple y directa de hacer inferencias sobre parámetros poblacionales a partir de una muestra, aunque su simplicidad también implica que no proporciona información sobre la precisión o variabilidad de la estimación, aspectos que se abordan mediante la **estimación por intervalos** y otras técnicas inferenciales.

Estimador puntual de la media

- Media muestral $\overline{X}$ es un estimador insesgado para la media poblacional μ, cuya distribución cumple que:
  - $E[\overline{X}] = \mu$
  - $V[\overline{X}] = \sigma^2/n$
- Dos casos:
  - Varianza conocida y n suficientemente grande ($n \geq 30$): $\overline{X} \sim N(\mu, \sigma/\sqrt{n}) \Leftrightarrow \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
  - Varianza desconocida y n suficientemente grande (se requiere normalidad), se utiliza V^2 y entonces $\frac{\overline{X}-\mu}{V/\sqrt{n}} \sim t_{n-1}$ siendo t_{n-1} la distribución t-Student con $n-1$ grados de libertad

Estimador puntual de la media

- Supongamos que se quiere estimar el gasto promedio que hacen los estudiantes en cultura
- Se selecciona una muestra aleatoria de 5 estudiantes cuyos gastos mensuales (en euros) son: 100, 50, 125, 20 y 40
- ¿Estimación puntual de la media?

Estimador puntual de la varianza

- Se usa la cuasivarianza porque es insesgado
- Cuasivarianza muestral $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$
- Su distribución cumple:
  - $E[S^2] = \sigma^2$
  - $V[S^2] = \frac{2\sigma^4}{n-1}$
- En poblaciones normales o con n suficientemente grande cumple

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

siendo χ^2_{n-1} la distribución chi-cuadrado con $n-1$ grados de libertad

Estimador puntual de la varianza

- Supongamos que se quiere estimar el gasto promedio que hacen los estudiantes en cultura
- Se selecciona una muestra aleatoria de 5 estudiantes cuyos gastos mensuales (en euros) son: 100, 50, 125, 20 y 40
- ¿Estimación puntual de la varianza?
- Supongamos que la muestra es de tamaño 500 y tenemos la media y la varianza estimadas. ¿Qué más sabemos?

Estimador puntual de la proporción

Dadas $X_1, \ldots, X_n$ variables aleatorias independientes e idénticamente distribuidas con una distribución Bernoulli de parámetro p. En este caso $\hat{p} = \overline{X}$ es un estimador insesgado para p y su distribución muestral cumple:

- $E[\hat{p}] = p$
- $V[\hat{p}] = \frac{p(1-p)}{n}$

Si $n \geq 30$, por el TCL se tiene que

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right) \Leftrightarrow \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \sim N(0,1)$$

Estimación puntual de la proporción

- En una encuesta a 100 personas, 70 dicen que prefieren estudiar de noche y 30 de día
- Se quiere estimar la proporción de personas que prefieren estudiar de noche
- ¿Estimación puntual de la proporción?

¿Cómo hemos llegado a los estadísticos usados?

- Método de los momentos
- Método de la máxima verosimilitud

Método de los momentos

El método de los momentos es una técnica utilizada en estadística para estimar los parámetros desconocidos $(\theta_1, \ldots, \theta_k)$ de una distribución de probabilidad. Fue introducido por el estadístico *Karl Pearson* en 1984.

Este método se basa en igualar los momentos muestrales (calculados a partir de los datos observados) con los momentos teóricos (expresados en términos de los parámetros de la distribución poblacional) y, con el sistema de ecuaciones resultantes, despejar los parámetros desconocidos. Así se logran los estimadores de los parámetros poblacionales desconocidos.

Momentos poblaciones

En estadística, los **momentos poblaciones** de una distribución son medidas que describen diversas características de la misma, como su media, varianza, simetría y curtosis. Los momentos más comunes son:

1. Primer momento (media): $\mu = E[X]$
2. Segundo momento (varianza): $\mu_2 = E[X^2]$
3. Tercer momento (asimetría): $\mu_3 = E[X^3]$
4. Cuarto momento (curtosis): $\mu_4 = E[X^4]$

k. k-ésimo momento: $\mu_k = E[X^k]$

Recordemos que:

- $\mu_k = E[X^k] = \sum_{i=1}^{n} x_i^k p_i$, siendo p_i la probabilidad de cada valor del dominio de la v.a. en el caso discreto (función de masa)
- $\mu_k = E[X^k] = \int x^k f(x) dx$ siendo $f(x)$ la función de densidad en el caso continuo

Momento muestral

Los momentos muestrales se definen como:

$m_k = \frac{1}{n} \sum_{i=1}^{n} x_i^k$

Pasos del método de los momentos

1. **Calcular Momentos Muestrales**: Se calculan los momentos muestrales de los datos observados. El (k)-ésimo momento muestral se define como $m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$ donde n es el tamaño de la muestra y x_i son los valores de la muestra
2. **Igualar Momentos Muestrales y Teóricos**: Se igualan los momentos muestrales con los momentos teóricos de la distribución. Los momentos teóricos se expresan en términos de los parámetros desconocidos que se desean estimar
3. **Resolver el Sistema de Ecuaciones**: Se resuelve el sistema de ecuaciones resultante para encontrar los estimadores de los parámetros desconocidos. Fíjate que tenemos k ecuaciones y k parámetros $(\theta_1, \ldots, \theta_k)$.De modo que es posible despejar los parámetros de estas ecuaciones, quedando estos parámetros en función de los momentos. En estas ecuaciones se sustituyen los momentos poblacionales por sus correspondientes momentos poblacionales. Esto da como resultado estimaciones de esos parámetros.

Método de los momentos

¡Veámos cómo obtener estimadores para los parámetros μ y σ^2 de la distribución Normal con el método de los momentos!

Método de los momentos

Ventajas

- **Simplicidad**: El método de los momentos es relativamente sencillo de aplicar y no requiere técnicas complejas de optimización.
- **Intuición**: Ofrece una interpretación intuitiva de los parámetros en términos de momentos.

Limitaciones

- **Precisión**: Los estimadores de los momentos no siempre son los estimadores más eficientes (no tienen la mínima varianza posible).
- **Aplicabilidad**: En algunas distribuciones complejas, los momentos pueden no existir o ser difíciles de calcular.
- **Consistencia**: Los estimadores de momentos no siempre son consistentes, especialmente en muestras pequeñas.

Método de la máxima verosimilitud

- El **método de la máxima verosimilitud** es una técnica estadística ampliamente utilizada para estimar los parámetros desconocidos de una distribución de probabilidad.
- Este método se basa en encontrar los valores de los parámetros que maximicen la función de verosimilitud, la cual mide la probabilidad de observar los datos dados los parámetros. Es decir, la idea básica es seleccionar el valor del parámetro que hace que los datos sean más probables.
- El método de máxima verosimilitud es el método más popular para obtener un estimador.

Método de la máxima verosimilitud

Dado un modelo estadiıstico (es decir, una familia de distribuciones $f(\cdot|\theta), \theta \in \Theta$ donde θ es el parámetro del modelo), el método de máxima verosimilitud encuentra el valor del parámetro del modelo θ que maximiza la función de verosimilitud:

$$\hat{\theta}(x) = \max_{\theta \in \Theta} L(\theta|\mathbf{x})$$

Para una muestra aleatoria $\mathbf{x} = (x_1, \ldots, x_n)$ de una variable aleatoria X, la **verosimilitud** es proporcional al producto de las probabilidades asociadas a los valores individuales:

$$\prod_j P(X = x_j)$$

Cuando X es una variable aleatoria continua, la verosimilitud es aproximadamente proporcional a

$$\prod_j f(x_j)$$

donde f es la función de densidad de X. Por lo tanto, la **verosimilitud** describe lo plausible que es un valor del parámetro poblacional, dadas unas observaciones concretas de la muestra

Función de Verosimilitud: La función de verosimilitud, $L(\theta|\mathbf{x})$, para un conjunto de datos $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ y un vector de parámetros θ, es el producto de las funciones de densidad (o de probabilidad) de los datos observados, dadas las posibles realizaciones de θ:

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = f(x_1, \ldots, x_n|\theta) =$$

$$= f(x_1|\theta)f(x_2|\theta)\ldots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

donde $f(x_i|\theta)$ es la función de densidad (o de probabilidad) de x_i dado θ.

Log-Verosimilitud: Debido a que la función de verosimilitud puede implicar productos de muchos términos, es más práctico trabajar con su logaritmo natural, conocido como la log-verosimilitud:

$$\ell(\theta|\mathbf{x}) = \log L(\theta|\mathbf{x}) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

Pasos del método de la máxima verosimilitud

1. **Especificar la Función de Verosimilitud**: Identificar la función de verosimilitud correspondiente a los datos observados y a la distribución supuesta.
2. **Calcular la Log-Verosimilitud**: Tomar el logaritmo natural de la función de verosimilitud para obtener la función de log-verosimilitud.
3. **Derivar y Resolver**: Derivar la función de log-verosimilitud con respecto a cada parámetro y resolver las ecuaciones obtenidas igualando a cero (puntos críticos) para encontrar los estimadores de máxima verosimilitud (EMV).
4. **Verificar Máximos**: Asegurarse de que las soluciones encontradas corresponden a máximos y no a mínimos o puntos de inflexión, típicamente verificando la segunda derivada.

Método de la máxima verosimilitud

¡Veámos cómo obtener estimadores para los parámetros μ y σ^2 de la distribución Normal con el método de la máxima verosimilitud!

Ventajas

- **Consistencia**: Los estimadores de máxima verosimilitud son consistentes
- **Eficiencia**: En muchos casos, los estimadores de máxima verosimilitud son eficientes, alcanzando la varianza mínima entre los estimadores insesgados.
- **Flexibilidad**: Se puede aplicar a una amplia gama de distribuciones y modelos complejos.
- **Invariantes**: Si T es el estimador de máxima verosimilitud para θ, entonces $\tau(T)$ es el estimador de máxima verosimilutd para $\tau(\theta)$ para cualquier función τ.

Limitaciones

- **Complejidad Computacional**: Encontrar los estimadores de máxima verosimilitud puede implicar resolver ecuaciones no lineales, lo cual puede ser complejo y requerir técnicas numéricas.
- **Existencia y Unicidad**: Los estimadores de máxima verosimilitud no siempre existen y, si existen, no siempre son únicos. En problemas reales, la derivada de la función de verosimilitud es, a veces, analíticamente intratable. En esos casos, se utilizan métodos iterativos para encontrar soluciones numéricas para las estimaciones de los parámetros.
- **Sesgo en Muestras Pequeñas**: Los estimadores pueden ser sesgados en muestras pequeñas, aunque el sesgo disminuye a medida que el tamaño de la muestra aumenta.

Referencias

Gomez Villegas, M. A. (2005). *Inferencia estadística*. Ediciones Díaz de Santos.

Spiegel, M., & Stephens, L. (2009). Estadística–Serie Schaum. *Mc Graw-Hill*.

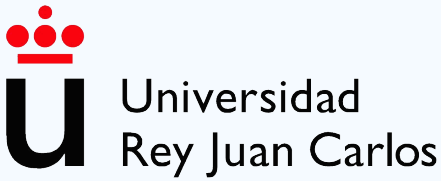
Wasserman, L. (2013). All of statistics: A concise course in statistical inference.

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

Estadística paramétrica. Intervalos de confianza

Inferencia Estadística - Grado en Ciencia e Ingeniería de Datos

Curso académico 2024-2025



¿De dónde venimos?

- Usamos estimación puntual para obtener una estimación del parámetro poblacional
  - El número medio de horas/día que pasan los estudiantes en Instagram es 3.7
- Dicha estimación es un único valor. ¡Nuestra mejor apuesta! Pero un único valor
- Los estimadores tienen distribución muestral. Con cada muestra, obtendremos una estimación puntual distinta
- Luego, cuando damos una estimación del parámetro poblacional en base a una muestra, vamos a fallar seguro...basta con coger otra muestra y ver que la estimación cambia/puede cambiar
- ¿Solución? Acompañar la estimación de su variabilidad (error estándar)

Estimación por intervalo

- ¿Solución? Acompañar la estimación de su variabilidad (error estándar)
- Con la estimación puntual y su error estándar —> construiremos **intervalos de confianza** para el parámetro poblacional de interés θ

$$\hat{\theta} \pm \text{error estándar}$$

- Estos intervalos tendrán un **nivel de confianza** $1 - \alpha$ asociado que indicará con qué seguridad el intervalo contiene el verdadero valor del parámetro

Estimación por intervalo

- La estimación puntual proporciona una aproximación razonable para un parámetro de la población, pero no tiene en cuenta la variabilidad debido al tamaño muestral, la variabilidad en la población, el conocimiento de otros parámetros, etc.
- La **estimación por intervalo** es una técnica en estadística que, a diferencia de la estimación puntual que proporciona un único valor, ofrece un rango de valores dentro del cual se espera que se encuentre el parámetro poblacional desconocido con un cierto nivel de confianza. Este rango se denomina **intervalo de confianza (IC)**
- La estimación por intervalos es una herramienta esencial en la inferencia estadística, ya que no solo ofrece una estimación del parámetro poblacional, sino que también proporciona un marco para entender la precisión y confiabilidad de esa estimación. Esto la convierte en una técnica poderosa para hacer inferencias más robustas y útiles basadas en datos muestrales

Conceptos clave

- **Intervalo de Confianza (IC)**: Es un rango de valores calculado a partir de los datos de la muestra, que se utiliza para estimar el parámetro poblacional θ desconocido. Se expresa comúnmente como (Límite Inferior, Límite Superior).
  Dada una muestra aleatoria simple $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ de una población X con función de distribución F que depende de un parámetro desconocido θ, diremos que un estimador por intervalos de confianza del parámetro θ con un nivel de confianza de $(1-\alpha) = 100 * (1-\alpha)\%$ es un intervalo de la forma $(T_{inf}(\mathbf{X}), T_{sup}(\mathbf{X}))$ que satisface:

$$P(\theta \in (T_{inf}(\mathbf{X}), T_{sup}(\mathbf{X}))) = 1 - \alpha$$

  Nótese que el parámetro θ es un valor fijo, pero $T_{inf}(\mathbf{X})$ y $T_{sup}(\mathbf{X})$ son cantidades aleatorias que variarán con cada muestra

Conceptos clave

- **Nivel de Confianza**: Es la probabilidad teórica de que el intervalo de confianza contenga el verdadero valor del parámetro poblacional. Se denota como $1 - \alpha$, donde α es el nivel de significancia.

Cuando tomamos una muestra, los valores ya son fijos, por lo que no podemos seguir hablando en términos de probabilidad

Conceptos clave: nivel de confianza

- Un **nivel de confianza** común es el 95%, lo que significa que estamos un 95% seguros de que el intervalo contiene el parámetro verdadero. Si repetimos el experimento N veces, en el 95% de las ocasiones el verdadero valor del parámetro estará incluido en el intervalo proporcionado. Sin embargo es importante señalar que, dado que el experimento solo suele realizarse en una ocasión, no podemos estar seguros de que el verdadero valor del parámetro está incluido en nuestro intervalo. Estará incluido o no estará incluido, pero no podemos saber en qué situación nos encontramos. Estar seguro sería tanto como decir que conocemos el verdadero valor del parámetro. En ese caso, obviamente, no necesitaríamos estimación ninguna.

Conceptos clave

- **Error Estándar (SE)**: Es una medida de la variabilidad de un estimador o estadístico muestral en las distintas muestras
- No debe confundirse con la desviación típica, que se refiere a la variabilidad de las observaciones individuales.
- Se utiliza para calcular los límites del intervalo de confianza.

Cálculo del Intervalo de Confianza

- El cálculo de un intervalo de confianza generalmente sigue la fórmula:

$$\text{Estimación Puntual} \pm (\text{Valor Crítico} \times \text{Error Estándar})$$

- Para alcanzar el intervalo de confianza, generalmente se busca una cantidad (aleatoria) $C(\mathbf{X}, \theta)$ relacionada con el parámetro desconocido θ y con la muestra $\mathbf{X}$, cuya distribución sea conocida y no dependa del valor del parámetro
- Esta cantidad recibe el nombre de *pivote* o *cantidad pivotal* para θ (¡aquí entran en juego las distribuciones muestrales de los estadísticos!)
- El valor crítico dependerá de la distribución teórica

Cálculo del Intervalo de Confianza

Dado que conocemos la distribución del pivote (conocemos la distribución del estimador), podemos usar los cuartiles $1-\alpha/2$ y $\alpha/2$ de dicha distribución, y el error estándar del estimador por intervalos de confianza, para plantear la siguiente ecuación:

$$P(\text{ cuantil}_{1-\alpha/2} < C(\mathbf{X},\theta) < \text{ cuantil}_{\alpha/2}) = 1-\alpha$$

Para obtener los extremos (inferior y superior) del estimador por intervalos de confianza $T_{inf}(\mathbf{X})$ y $T_{sup}(\mathbf{X})$, se resuelve la doble desigualdad en θ. De este modo el intervalo de confianza al $100(1-\alpha)\%$ para θ es $(T_{inf}(\mathbf{X}), T_{sup}(\mathbf{X}))$

IC para la media con var conocida

Sea una muestra aleatoria simple $(X_1, \ldots, X_n)$ de tamaño n obtenida de X, siguiendo una distribución normal con parámetros (μ desconocido) y varianza conocida (σ^2), $N(\mu, \sigma^2)$. Queremos obtener un IC para la media μ con un nivel de confianza $1 - \alpha$.

Como ya hemos visto, el estadístico $\bar{X}$ tiene una distribución normal:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Además, sabemos que:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

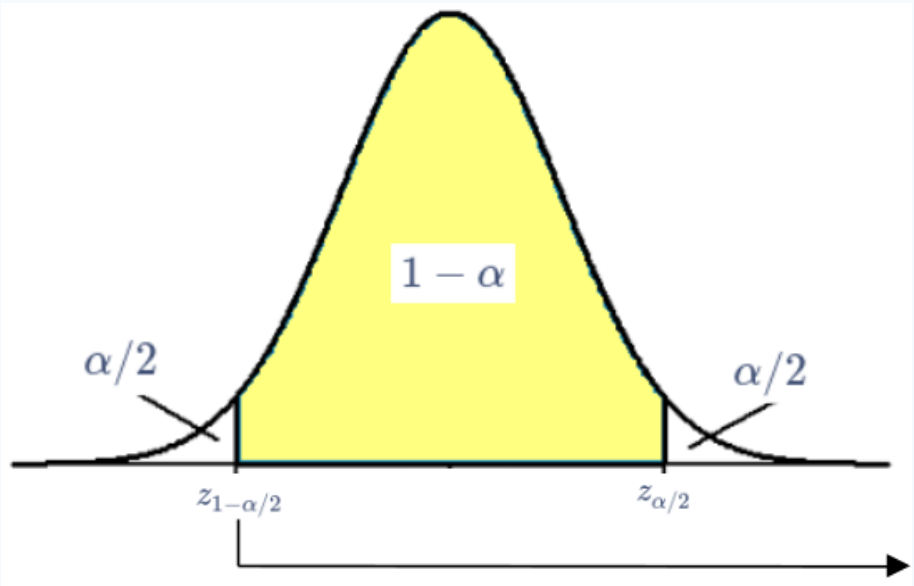
que es la cantidad pivotal para el IC para μ

IC para la media con var conocida

Ahora, si $z_{1-\alpha/2}$ y $z_{\alpha/2}$ son los cuartiles $1-\alpha/2$ y $\alpha/2$ de la distribución $N(0,1)$, entonces tenemos:

$$P\left(z_{1-\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha$$

Nótese que, como la distribución Normal es simétrica: $z_{1-\alpha/2} = -z_{\alpha/2}$



Resolvemos la doble desigualdad para μ:

$$-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}$$

$$-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} - \bar{X} < -\mu < -\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} + \bar{X} > \mu > \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

De modo que el estimador por intervalos de confianza es:

$$\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

y por tanto, el intervalo de confianza para la media se calcula como:

$$IC_{1-\alpha}(\mu) = \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \left(\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

donde $\bar{X}$ es la media muestral, $z_{\alpha/2}$ es el valor crítico del estadístico z para el nivel de confianza deseado y $\sigma/\sqrt{n}$ es el error estándar

¿Qué puede afectar al IC?

Pensemos juntos...¿qué factores pueden influir en la longitud intervalo de confianza?

¿Qué puede afectar al IC?

- Tamaño muestral
- Variabilidad
- Nivel de confianza

Aclaración cuantiles

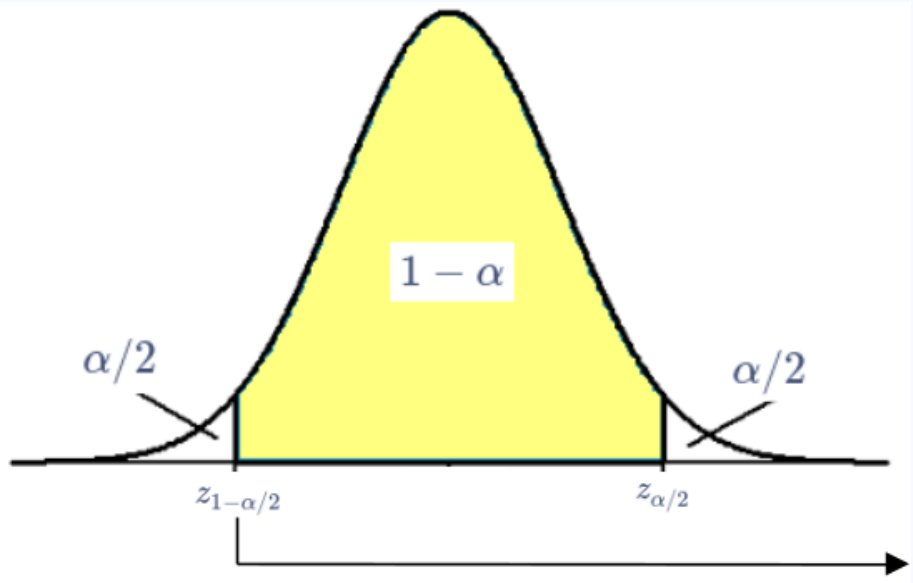


Figure 1: Cola superior

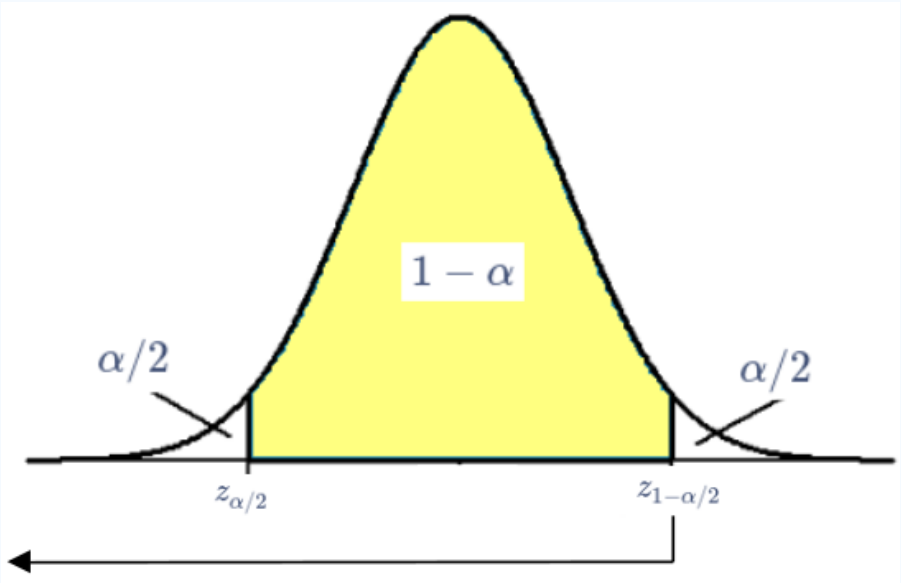


Figure 2: Cola inferior (en R: lower.tail = TRUE)

En R, el parámetro `lower.tail` (en funciones del tipo `pnorm()`, `qnorm()`) nos permite elegir la cola:

- Para la cola superior, `lower.tail = FALSE`
- Para la cola inferior, `lower.tail = TRUE`

IC para la media con var conocida

Se ha probado que la altura de las alumnas de primer curso de la URJC se puede aproximar mediante una variable aleatoria con distribución normal con desviación típica $\sigma = 10$ cm pero la media μ desconocida. En un estudio con 50 alumnas se obtiene una media de 166 cm. Vamos a construir un intervalo de confianza al 95% para μ.

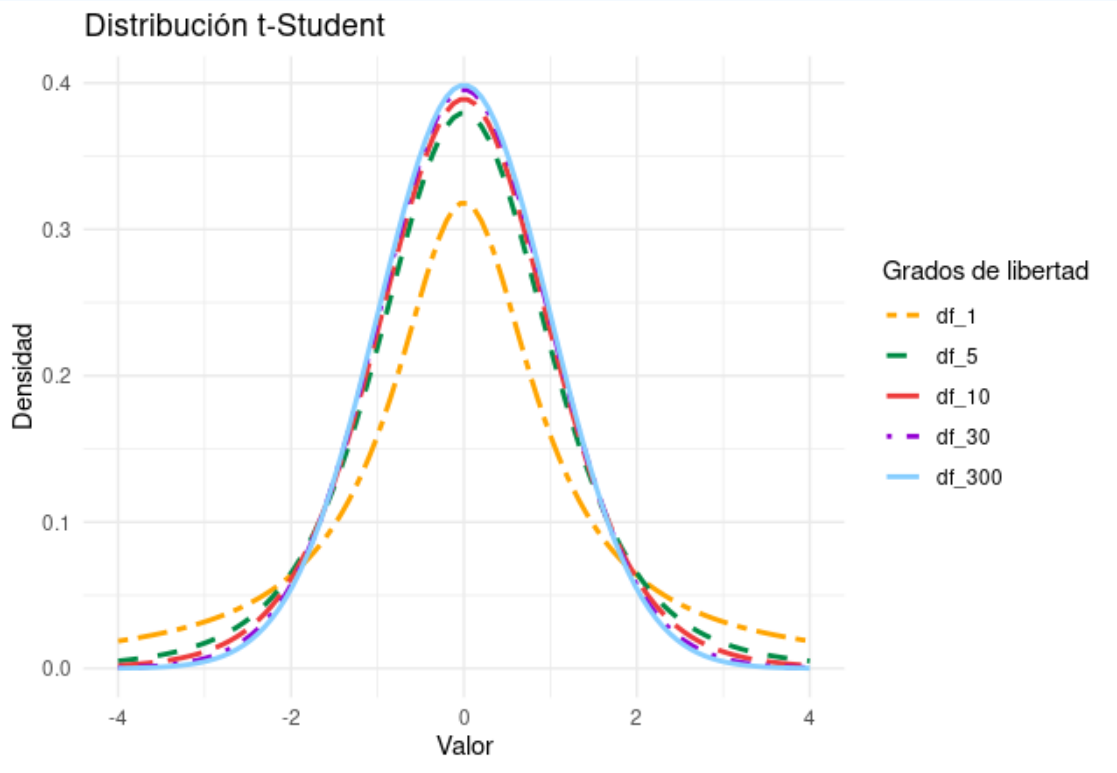
IC para la media con var desconocida

- Concepto previo: Distribución t-Student

Distribución t-Student

La distribución t-student t_n tiene un único parámetro $n > 0$ llamado grados de libertad

- Se parece a la distribución Normal pero con colas "más pesadas"
- Según $n \longrightarrow \infty$, la distribución es una $N(0, 1)$
- Data $T \sim t_n$ para $n > 1$, $E[T] = 0$ y para $n > 2$, $V[T] = \frac{n}{n-2}$



Distribución t-Student

- Sea $X_1, \ldots, X_n$ una muestra aleatoria simple (v.a.i.i.d.) de una población $N(\mu, \sigma^2)$, entonces el estadístico T:

$$T = \frac{\bar{X} - \mu}{\sqrt{S^2/n - 1}} = \frac{\bar{X} - \mu}{\sqrt{\hat{s}^2/n}} \sim t_{n-1}$$

  sigue la distribución t de Student con $n - 1$ grados de libertad.

- La distribución t de Student se puede definir como

$$T = \frac{N(0,1)}{\sqrt{\chi^2_n/n}}$$

  siendo $N(0,1)$ la normal estándar, χ^2_n es la distribución chi-cuadrado con n grados de libertad, siendo la normal y la chi-cuadrado independientes

Distribución t-Student

En R las funciones de la t-Student son:

- `dt()`: Función de densidad de probabilidad
- `qt()`: Función cuantil de la distribución
- `pt()`: Función de distribución acumulada
- `rt()`: Generación de números pseudoaleatorios

IC para la media con var desconocida

Sea una muestra aleatoria simple $(X_1, \ldots, X_n)$ de tamaño n obtenida de X, siguiendo una distribución normal con parámetros μ y varianza σ^2 desconocidos . Queremos obtener un IC para la media μ con un nivel de confianza $1 - \alpha$.

La cantidad pivotal para μ es:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

donde S^2 es la es la cuasi-varianza muestral y t_n es la distribución t de Student con n grados de libertad

IC para la media con var desconocida

Si $t_{n-1;1-\alpha/2}$ y $t_{n-1;\alpha/2}$ son los cuantiles $1-\alpha/2$ y $\alpha/2$ respectivamente de una distribución t de Student con $n-1$ grados de libertad:

$$P(t_{n-1;1-\alpha/2} < T < t_{n-1;\alpha/2}) = 1-\alpha$$

Como la distribución t-Student es simétrica: $t_{n-1;\alpha/2} = t_{n-1;1-\alpha/2}$

$$P(-t_{n-1;\alpha/2} < \frac{\bar{X}-\mu}{S/\sqrt{n}} < t_{n-1;\alpha/2}) = 1-\alpha$$

IC para la media con var desconocida

Se resuelve la doble desigualdad para μ y se obtiene el estimador por intervalos de confianza:

$$\left(\bar{X} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}\right)$$

Resultando el intervalo de confianza:

$$IC_{1-\alpha}(\mu) = \left(\bar{\mathbf{X}} - t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}, \bar{\mathbf{X}} + t_{n-1;\alpha/2}\frac{S}{\sqrt{n}}\right)$$

IC para la media con var desconocida

Se ha medido la temperatura media de una muestra aleatoria de 10 soluciones salinas, obteniendo los siguiente resultados:

$$37.2, 34.1, 35.5, 34.5, 32.9, 37.3, 32.0, 33.1, 42.0, 34.8$$

Se nos pide calcular el IC al 90% para la temperatura media, suponiendo que la temperatura de la solución salina se puede aproximar mediante una variable aleatoria con distribución normal

IC para la media con n grande

Sea $\mathbf{X} = (X_1, \ldots, X_n)$ una muestra aleatoria simple (v.a.i.i.d) de tamaño n de una variable aleatoria X. Supongamos que X sigue una distribución (conocida o no) con parámetros μ y σ^2. Además, supongamos que $n \geq 30$. Entonces, por el *Teorema Central del Límite* se tiene que la cantidad pivotal para μ cumple la siguiente propiedad:

$$Z = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} \sim N(0,1)$$

Si $z_{1-\alpha/2}$ y $z_{\alpha/2}$ son los cuantiles $1-\alpha/2$ y $\alpha/2$ de $N(0,1)$, tenemos:

$$P(z_{1-\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

IC para la media con n grande

Obtenemos el estimador por intervalos de confianza resolviendo la doble desigualdad para μ:

$$\left(\bar{X} - z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right)$$

El intervalo de confianza es:

$$IC_{1-\alpha}(\mu) = \left(\bar{\mathbf{x}} - z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}, \bar{\mathbf{x}} + z_{\alpha/2}\frac{\hat{\sigma}}{\sqrt{n}}\right)$$

IC para la media para la proporción

Sea una muestra aleatoria simple $(X_1, \ldots, X_n)$ de tamaño n obtenida de X, siguiendo una distribución de Bernoulli con parámetro p. Esto es:

$$\mu = E[X] = p \qquad \sigma^2 = Var[X] = p(1-p)$$

Además, supongamos que $n \geq 30$. Entonces, por el TCL se tiene que la cantidad pivotal para $\hat{p} = \bar{X}$ cumple la siguiente propiedad:

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0,1)$$

EL intervalo de confianza con nivel de confianza $1-\alpha$ para estimar una proporción poblacional p es:

$$IC_{1-\alpha}(p) = \left(\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

IC para la media para la proporción

- Si el tamaño muestral no es lo suficientemente grande y no se puede aplicar el Teorema Central del Límite –> hay que comprobar la normalidad de los datos
- Si los datos son normales, el intervalo de confianza anterior es apropiado
- En caso contrario, tendrían que buscarse alternativas

IC para la media para la proporción

Supongamos que estamos realizando una encuesta para determinar la proporción de personas que apoyan una nueva política ambiental en una ciudad. Hemos encuestado a 1000 personas, y 560 de ellas han respondido que apoyan la nueva política.

Queremos calcular un intervalo de confianza del 95% para la proporción de apoyo en toda la población.

Interpretación de los IC

- Si calculamos un intervalo de confianza del 95% para la media poblacional y, por ejemplo, obtenemos un intervalo de $(5, 10)$, esto no significa que hay un 95% de probabilidad de que la media poblacional esté en ese intervalo en un caso particular, sino que, si repetimos este procedimiento muchas veces, el 95% de los intervalos construidos contendrán la verdadera media poblacional.
- Podríamos decir que estamos un 95% seguros de que la media poblacional se encuentra entre 5 y 10, pero ¡ojo!, la media poblacional (cuyo valor desconocemos) estará o no estará en ese intervalo.

Interpretación de los IC

Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability - J. Neyman

It will be noticed that in the above description the probability statements refer to the problems of estimation with which the statistician will be concerned in the future. In fact, I have repeatedly stated that the frequency of correct results *will* tend to α.* Consider now the case when a sample, E′, is already drawn and the calculations have given, say, $\underline{\theta}\,(E') = 1$ and $\overline{\theta}\,(E') = 2$. Can we say that in this particular case the probability of the true value of θ_1 falling between 1 and 2 is equal to α ?

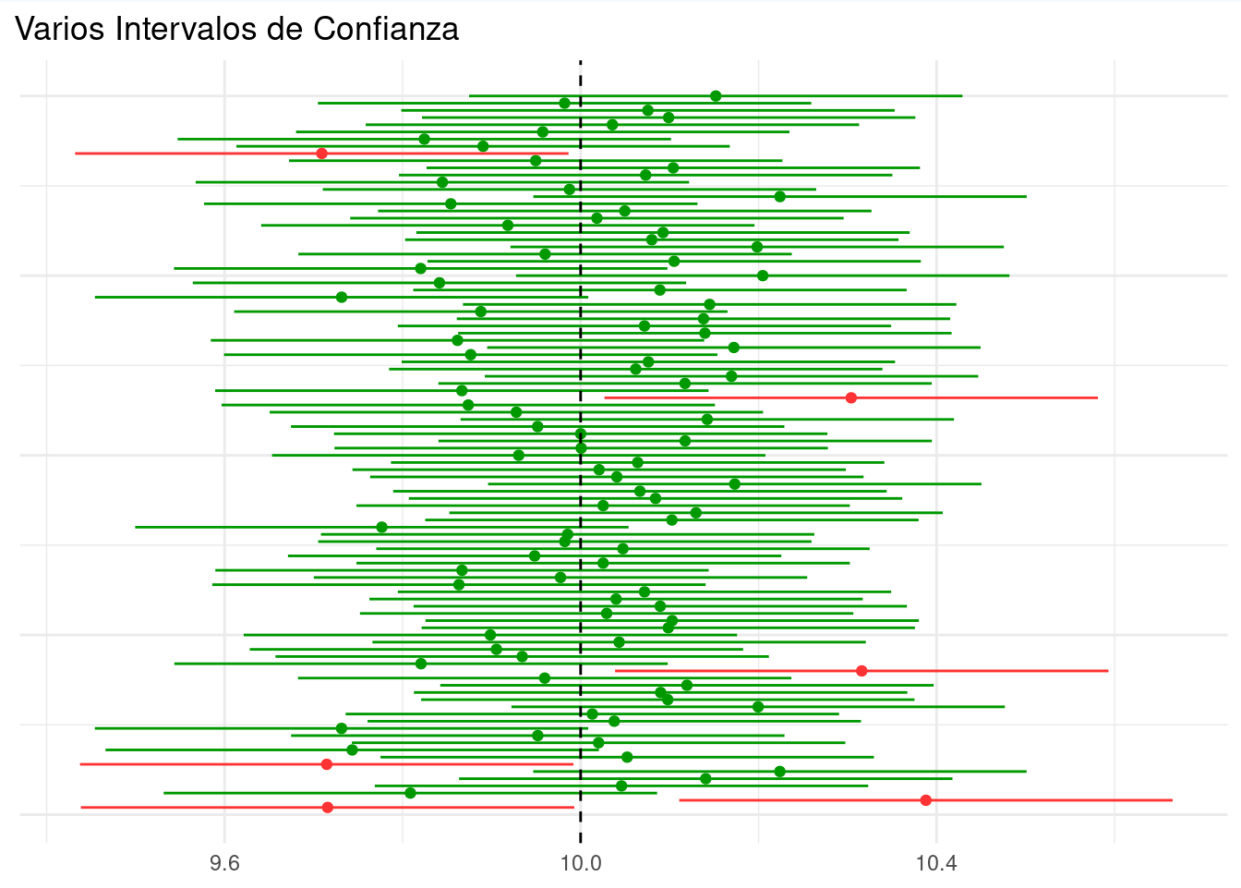
The answer is obviously in the negative. The parameter θ_1 is an unknown constant and no probability statement concerning its value may be made, that is except for

Notación en este artículo: $$P\{\underline{\theta}\,(E) \leq \theta_1{}^0 \leq \overline{\theta}\,(E) | \theta_1{}^0\} = \alpha$$

Simulación para interpretar correctamente el concepto frecuentista de intervalo de confianza

- Generamos 100 muestras de tamaño $n = 50$ de una distribución $X \sim N(\mu = 10, \sigma^2 = 1)$
- Para cada muestra, se construye un IC para la media con $\alpha = 0.05$.
- Representamos todos esos intervalos de confianza en un único gráfico. En verde se pintan los intervalos de confianza que incluyen el verdadero valor del parámetro 10. En rojo los que no

Interpretación de los IC



- ¿Cuántos intervalos de confianza, de entre los 100 contienen al verdadero valor del parámetro?

Interpretación de los IC

- ¿Cómo crees que afecta a la longitud del intervalo de confianza los siguientes aspectos?
  - Tamaño muestral
  - Nivel de confianza

Tamaño muestral

- Un tamaño muestral adecuado es crucial en la inferencia estadística
- Garantiza que los intervalos de confianza sean precisos y que las conclusiones obtenidas sean representativas de la población
- Las técnicas para determinar el tamaño muestral están relacionadas directamente con los intervalos de confianza y se basan en varios factores:
  - Nivel de confianza deseado
  - La precisión (o margen de error) deseada
  - La variabilidad esperada en la población

Tamaño muestral: factores clave

1. **Nivel de confianza** $1-\alpha$: indica el grado de certeza con el que intervalo de confianza contiene al parámetro poblacional. Niveles de confianza comunes son 90%, 95% y 99%. Un nivel de confianza más alto requiere una muestra más grande para asegurar la misma precisión
2. **Margen de Error (E)**: Es la máxima diferencia tolerable entre la estimación muestral y el valor real del parámetro poblacional. Un margen de error más pequeño requiere una muestra más grande para asegurar una estimación precisa
3. **Variabilidad poblacional** σ: La variabilidad en la población, medida por la desviación estándar, afecta directamente al tamaño muestral. Una mayor variabilidad requiere una muestra más grande para obtener una estimación precisa

Tamaño muestral para media poblacional (Normal)

El tamaño muestral n necesario para estimar una media poblacional con un margen de error E y un nivel de confianza $1-\alpha$ se puede calcular usando la fórmula:

$$n = \left(\frac{z_{\alpha/2} \cdot \sigma}{E} \right)^2$$

donde:

- $z_{\alpha/2}$ es el valor crítico del estadístico z correspondiente al nivel de confianza deseado
- σ es la desviación estándar de la población (si es desconocida, se puede usar la desviación estándar de la muestra S)

Tamaño muestral para media poblacional (Normal)

¿De dónde sale esta fórmula?

Efectivamente, teníamos que el intervalo de confianza para la media era:

$$\left(\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Y buscamos el n para que el margen de error sea menor que E, es decir:

$$z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < E$$

$$n > \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2$$

Ejemplo

Supongamos que deseamos estimar la media de una población con un nivel de confianza del 95%, un margen de error de 5 unidades y se estima que la desviación estándar de la población es 15 unidades. ¿Qué tamaño muestral se necesita?

Tamaño muestral para proporción poblacional

El tamaño muestral n necesario para estimar una proporción poblacional p con un margen de error E y un nivel de confianza $1-\alpha$ se puede calcular usando la fórmula:

$$n = \frac{z_{\alpha/2}^2 \cdot p \cdot (1-p)}{E^2}$$

donde

- p es la proporción esperada (si no se conoce, se usa $p = 0.5$ para maximizar el tamaño muestral)
- $z_{\alpha/2}$ es el valor crítico del estadístico z correspondiente al nivel de confianza deseado

Ejemplo

Supongamos que deseamos estimar la proporción de personas que aprueban una nueva ley con un nivel de confianza del 95%, un margen de error del 3% (0.03) y se estima que la proporción esperada es $p = 0.5$. ¿Qué tamaño muestral se necesita?

Referencias

Wasserman, L. (2013). All of Statistics: a Concise Course in Statistical Inference.

Spiegel, M., & Stephens, L. (2009). Estadística–Serie Schaum. *Mc Graw-Hill*.

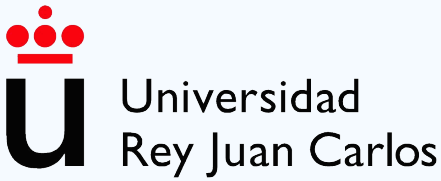
Gomez Villegas, M. A. (2005). *Inferencia estadística*. Ediciones Díaz de Santos.

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

Contrastes de hipótesis

Inferencia Estadística - Grado en Ciencia e Ingeniería de Datos

Curso académico 2024-2025



Contrastes de hipótesis

- Estimación puntual. Conocemos la distribución a excepción de un parámetro. Estimamos el valor de dicho parámetro.
- Intervalos de confianza. Construimos un intervalo que contiene el valor del parámetro con una confianza de $1 - \alpha$
- Contrastes de hipótesis. Sirven para responder preguntas del tipo:
  - Dados los datos, ¿el valor del parámetro puede ser a? Por ejemplo, ¿la altura media de las alumnas de la URJC es 1.65m?
  - En base a la muestra, ¿estos dos medicamentos son igual de efectivos ($\mu_1 = \mu_2$) para tratar la ansiedad?

Contrastes de hipótesis

- Permiten evaluar si los datos disponibles proporcionan suficiente evidencia en contra de una hipótesis previamente establecida sobre una población
- Es un proceso estructurado para evaluar afirmaciones sobre parámetros poblacionales utilizando datos muestrales. Mediante la formulación de hipótesis, selección de niveles de significancia, elección de estadísticas de prueba y evaluación del p-valor, podemos tomar decisiones informadas y cuantitativamente justificadas
- Este enfoque es fundamental en muchas áreas de investigación y análisis de datos, proporcionando un marco riguroso para la inferencia estadística

Conceptos básicos: hipótesis nula

Hipótesis nula (H_0):

La hipótesis nula es una afirmación sobre parámetros poblacionales que se asume verdadera hasta que se presente suficiente evidencia en contra. Se asume inicialmente que la hipótesis nula es correcta (semejante a suponer inocencia a menos que se pruebe la culpa). Habitualmente corresponde al estatus quo. Esto es, generalmente, la hipótesis nula representa un estado de "no efecto" o "no diferencia".

Ejemplo:

- $H_0 : \mu = 50$ Se contrasta si la media poblacional es 50
- $H_0 : \mu \leq 50$ Se contrasta si la media poblacional es ≤ 50
- $H_0 : \mu_1 = \mu_2$ Se contrasta si la media de las dos muestras es igual

Se rechaza H_0 cuando los datos apoyan mucho más otra hipótesis, llamada hipótesis alternativa H_1

Hipótesis alternativa H_1:

La hipótesis alternativa es una afirmación que contrasta con la hipótesis nula y representa el efecto o diferencia que se desea detectar

Ejemplo:

- $H_1 : \mu \neq 50$ La media poblacional no es 50
- $H_1 : \mu > 50$ La media poblacional es mayor que 50

Ejemplo H0 vs H1 media poblacional

Supongamos que una empresa de educación en línea afirma que sus estudiantes pasan en promedio al menos 4 horas diarias estudiando en su plataforma. Queremos comprobar si esta afirmación es cierta basándonos en una muestra de estudiantes

- La hipótesis nula es la afirmación que queremos poner a prueba y que asumimos verdadera inicialmente. En este caso, la hipótesis nula es que la media del tiempo de estudio diario es de al menos 4 horas.

$$H_0 : \mu \geq 4 \text{ horas}$$

- La hipótesis alternativa es lo que queremos demostrar y se contrapone a la hipótesis nula. En este caso, queremos ver si el tiempo de estudio diario es menor de 4 horas. Fíjate que la empresa podría estar "inflando" sus resultados y lo "intersante" en este caso es "demostrar" que realmente los alumnos pasan menos tiempo en la plataforma.

$$H_1 : \mu < 4 \text{ horas}$$

Ejemplo H0 vs H1 proporción poblacional

Supongamos que el rectorado de la URJC afirma que menos del 20% de los estudiantes de sus grados, fuman. Queremos verificar si la proporción de fumadores es mayor al 20%

- La hipótesis nula es que la proporción de fumadores es menor o igual al 20%

$$H_0 : p \leq 0.20$$

- La hipótesis alternativa es si la proporción de fumadores es mayor al 20%

$$H_1 : p > 0.20$$

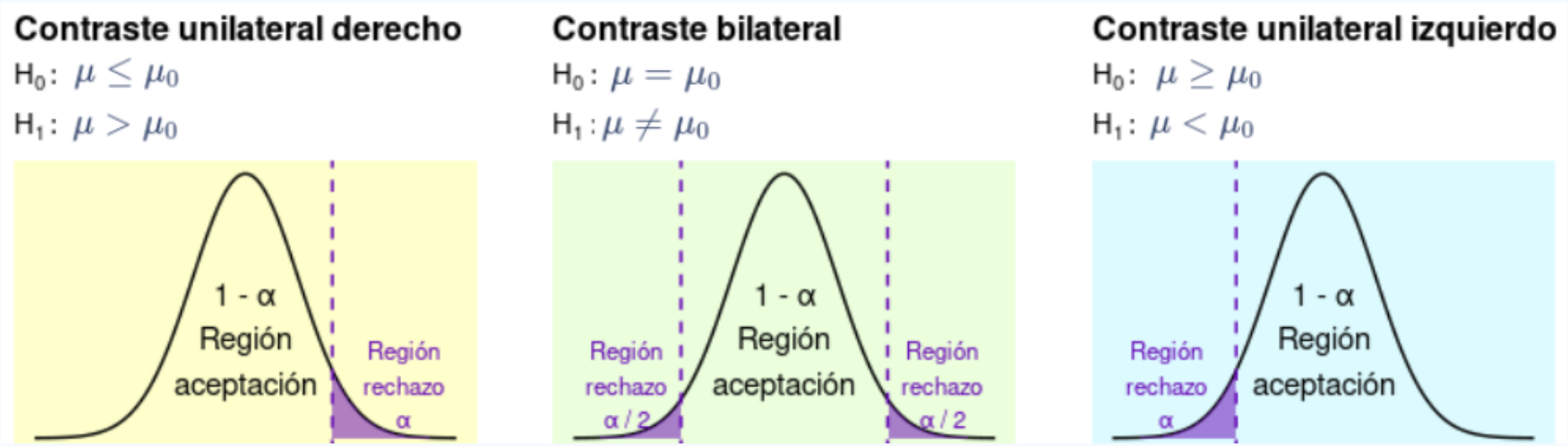
Tipos de contrastes

- Unilarerales. Se contrasta si el valor es mayor o menor, es decir, si queda a a izquierda o a la derecha en la distribución

$$H_0 : p \leq 0.20 \quad \text{vs} \quad H_1 : p > 0.20$$

- Bilaterales. Se contrasta la igualdad y se rechaza si queda a la derecha o a la izquierda en la distribución del estadístico bajo H_0

$$H_0 : p = 0.20 \quad \text{vs} \quad H_1 : p \neq 0.20$$



Pasos en un contraste de hipótesis

1. **Formular las hipótesis**:
   - Definir H_0 y H_1 claramente
2. **Seleccionar el nivel de significatividad estadística** (α):
   - El nivel de significatividad estadística es la probabilidad de rechazar H_0 cuando es verdadera. Comúnmente, se utilizan $\alpha = 0.05, 0.01, 0.10$.
3. **Elegir el estadístico de prueba**:
   - Seleccionar un estadístico que resuma la información de interés sobre los datos (ejemplo: media muestral para la estatura media) y que siga una distribución conocida bajo H_0 (por ejemplo, la distribución Normal)
4. **Calcular el** $p - valor$:
   - El $p - valor$ es la probabilidad de observar un valor tan extremo o más extremo que el observado, bajo la suposición de que H_0 es verdadera

5. **Tomar una decisión**:
   - La regla de decisión de un contraste de hipótesis se basa en la "distancia" entre los datos muestrales y los valores esperados si H_0 es cierta
   - Esta distancia se calcula a partir del estadístico del contraste y se considera "grande" o no, en base a la distribución del mismo y a la probabilidad de observar realizaciones "más extremas" de dicho estadístico. Para tomar la decisión, comparamos el $p - valor$ con α:
     - Si $p - valor \leq \alpha$, se rechaza H_0. Hay suficiente evidencia en la muestra como para rechazar la hipótesis nula. El valor del parámetro establecido en H_0 es poco creíble dada la muestra observada.
     - Si $p - valor > \alpha$, no se rechaza H_0. **Muy importante**: esto no significa que la hipótesis nula sea cierta. La interpretación es que no existe, en la muestra que hemos observado, suficiente evidencia en contra de la hipótesis nula como para recharzarla.

Tenemos por tanto que el $p - valor$ es una medida que nos dice cuán probable sería obtener nuestros datos observados si la hipótesis nula fuera verdadera. En otras palabras, mide la evidencia en contra de H_0. Si el $p - valor$ es pequeño (generalmente menor que 0.05), tenemos razones para rechazar H_0. Si es grande, no tenemos suficiente evidencia para rechazarla

Ejemplo para la media

- Queremos probar si, después de una campaña de concienciación, los jóvenes de 20 años de GCID pasan, de media, menos de 3 horas diarias en Instagram o siguen pasando 3 horas o más
- Se sabe que la v.a. X : horas que pasan los jóvenes en Instagram sigue una distribución Normal de media desconocida μ y varianza $\sigma^2 = 0.25$
- Para contrastar si la campaña de concienciación ha tenido efecto, se toma una muestra de $n = 50$ estudiantes

Hipótesis:

- Hipótesis nula H_0: $\mu \geq 3$ horas
- Hipótesis alternativa H_1: $\mu < 3$ horas

Ejemplo para la media

- Si la media muestral de los 50 estudiantes es $1'5$ horas, ¿qué haríamos?
- ¿Y si fuera de 4 horas?
- Como la media muestral (en este caso) sigue una distribución $N(\mu, \sigma^2/n) = N(3, 0.25/50)$ bajo H_0, trabajamos en términos de probabilides
- Con el contraste lo que hacemos es, dada la distribución de la población bajo H_0, estudiar cómo de probable es obtener una media muestral como la lograda con los 50 estudiantes
- Si la probabilidad es muy baja –> existen evidencias para rechazar la hipótesis nula de que los jóvenes pasan 3 o más horas en Instagram. El número medio de horas sería significativamente menor
- Si la probabilidad es muy alta –> no existen evidencias para rechazar la hipótesis nula

Ejemplo para la media

- Bajo H_0:
  - La población $X \sim N(\mu = 3, \sigma^2 = 0.25)$
  - De dicha población, se toma una m.a.s de v.a. independientes e igualmente distribuidas $X_1, \ldots, X_{50} \sim N(\mu = 3, \sigma^2 = 0.25)$
  - Sabemos que $\overline{X} \sim N(\mu = 3, \sigma^2/n = 0.25/50)$
  - Calculamos el valor de la media muestral en nuestra muestra concreta: $\overline{x} = 2.75$
  - Suponiendo dicha distribución, calculamos la probabilidad de que el número medio de horas sea 2.75: $P(\overline{X} < 2.75) = 0.0002$ ¿Interpretación?
- $p - valor = 0.0002$ ¿Decisión con $\alpha = 0.05$?

Errores

- Una vez especificadas las hipótesis nula H_0 y alternativa H_1 y recogida la información muestral, se toma una decisión sobre la hipótesis nula (rechazar o no rechazar H_0)
- Existe la posibilidad de llegar a una conclusión equivocada, porque solo se dispone de una muestra aleatoria y no se puede tener la certeza de que sea correcta o no
- Cuando realizamos un contraste de hipótesis, pueden ocurrir 2 errores: el **error de tipo I** o α y el **error de tipo II** o β
- Entender estos errores es fundamental para interpretar correctamente los resultados de cualquier prueba estadística
- El balance entre α y β, así como el tamaño de la muestra, juegan un papel importante en la fiabilidad de los resultados obtenidos

Error de tipo I

- El error de tipo I ocurre cuando rechazamos la hipótesis nula H_0 siendo esta verdadera. En otras palabras, concluimos que hay un efecto o una diferencia cuando, en realidad, no la hay
- El nivel de significancia α es la probabilidad de cometer un error de tipo I:

$$\alpha = P(\text{rechazar } H_0 \mid H_0 \text{ es correcta})$$

- Este valor se establece de antemano, comúnmente 0.05, 0.01 o 0.10.
- Si el $p - valor$ de nuestra prueba es menor o igual a α, rechazamos H_0. Por ejemplo, si $\alpha = 0.05$, esto significa que estamos dispuestos a aceptar un 5% de probabilidad de rechazar H_0 cuando es verdadera

Error de tipo II

- El error de tipo II ocurre cuando no rechazamos la hipótesis nula H_0 siendo esta falsa. En otras palabras, concluimos que no hay un efecto o una diferencia cuando, en realidad, sí la hay:

$$\beta = P(\text{No rechazar } H_0 \mid H_0 \text{ es incorrecta})$$

- **Potencia del test**: La potencia de una prueba estadística es la probabilidad de rechazar H_0 cuando H_0 es falsa:

$$Potencia = 1 - \beta = P(\text{Rechazar } H_0 \mid H_1 \text{ es correcta})$$

- Una alta potencia es deseable –> mayor probabilidad de detectar un efecto o diferencia cuando realmente existe. Por ejemplo, si $\beta = 0.20$, esto significa que hay un 20% de probabilidad de no rechazar H_0 cuando es falsa. La potencia de la prueba, probabilidad de detectar un efecto cuando realmente existe, sería 0.80

Ejemplo. Errores Tipo I y II

Supongamos que estamos evaluando la efectividad de un nuevo medicamento.

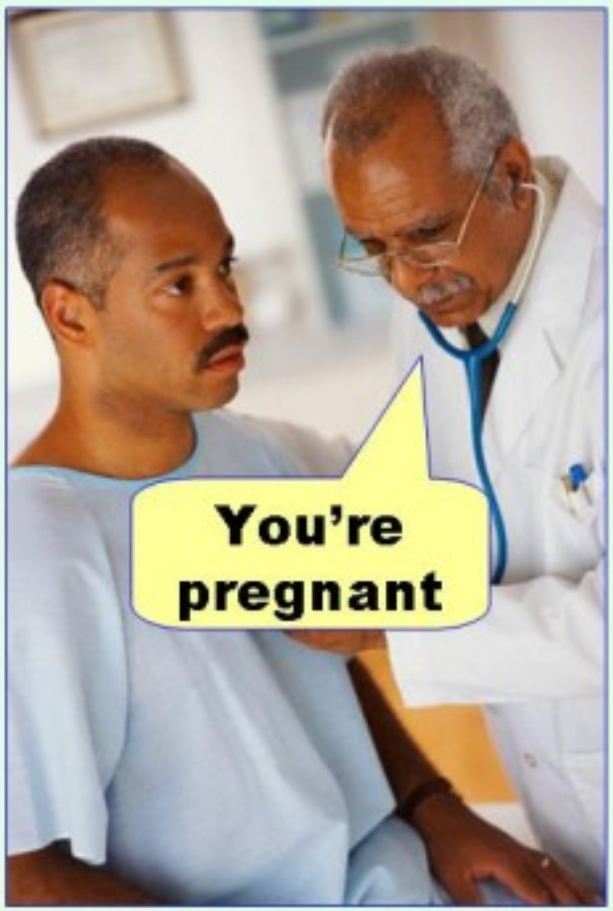
- Hipótesis nula H_0: El medicamento no tiene efecto $(\mu = 0)$.
- Hipótesis alternativa H_1: El medicamento tiene un efecto $(\mu \neq 0)$.

Error de tipo I: Si el medicamento no tiene ningún efecto pero el estudio concluye que sí lo tiene, hemos cometido un error de tipo I. Esto podría llevar a la aprobación y uso de un medicamento ineficaz

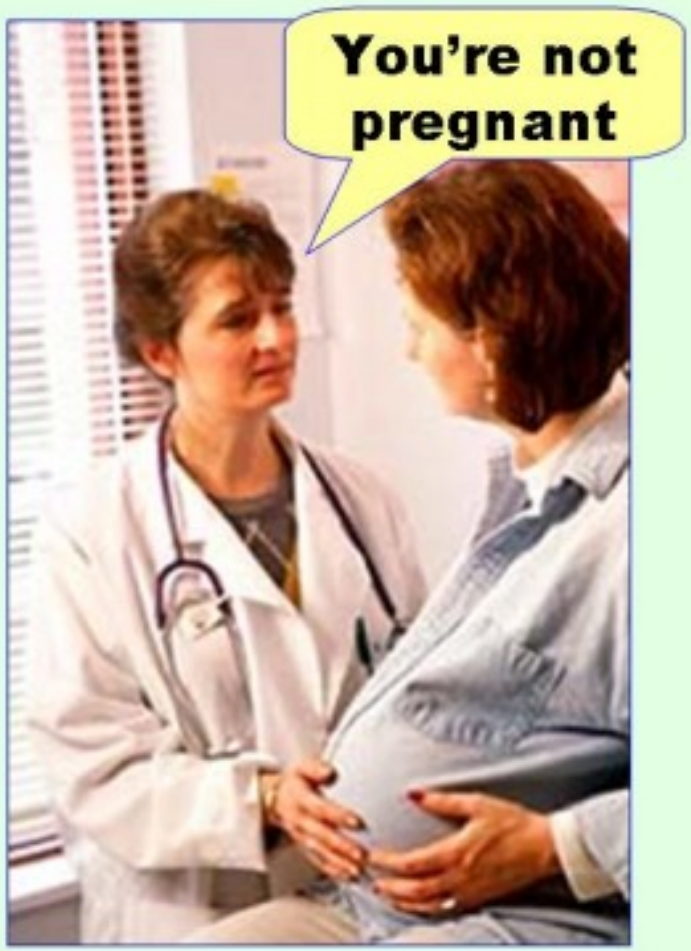
Error de tipo II: Si el medicamento tiene un efecto, pero el estudio concluye que no lo tiene, hemos cometido un error de tipo II. Esto podría llevar a la no aprobación de un medicamento potencialmente beneficioso

Ejemplo. Errores Tipo I y II

Type I error
(false positive)



Type II error
(false negative)



Relación entre errores de tipo I y II

- **Inversamente proporcionales**: Reducir α (haciendo la prueba más conservadora y menos propensa a rechazar H_0) generalmente aumenta β (haciendo la prueba más propensa a no detectar un efecto cuando realmente existe), y viceversa. Los errores de tipo I y de tipo II no se pueden comenter simultáneamente:
  - El error de tipo I solo puede darse si H_0 es correcta
  - El error de tipo II solo puede darse si H_0 es incorrecta
- **Tamaño de la muestra**: Aumentar el tamaño de la muestra puede reducir ambos tipos de errores, incrementando la precisión de la prueba

Relación entre errores de tipo I y II

La siguiente tabla refleja la relación entre los dos tipos de errores en relación con la decisión del contraste y la verdadera situación en la población:

	Verdadera situación	
Decisión	H_0 correcta	H_0 incorrecta
No rechazar H_0	Sin error $(1-\alpha)$	Error de Tipo II (β)
Rechazar H_0	Error de Tipo I (α)	Sin error $(1-\beta$=potencia)

Relación entre errores de tipo I y II

Es importante notar que, si todo lo demás no cambia, entonces la potencia del contraste disminuye cuando:

- La diferencia entre el valor supuesto para el parámetro y el valor real disminuye
- La variabilidad de la población aumenta
- El tamaño muestral disminuye

CH para la media (var conocida, Normal)

- El contraste para la media de una población normal con varianza conocida es un procedimiento estadístico utilizado para determinar si la media de una población difiere de un valor específico (hipótesis nula)
- El parámetro de estudio es la media de la variable aleatoria: $X \sim N(\mu, \sigma^2)$
- Distintas opciones de hipótesis:
  - Hipótesis nula: H_0: $\mu = \mu_0$ (La media de la población es igual a μ_0)
  - Tenemos varias opciones para la hipótesis alternativa:
    - H_1: $\mu \neq \mu_0$ (Contraste bilateral)
    - H_1: $\mu > \mu_0$ (Contraste unilateral derecho)
    - H_1: $\mu < \mu_0$ (Contraste unilateral izquierdo)

CH para la media (var conocida, Normal)

- Nivel de significación α
- El estadístico de prueba se calcula utilizando la distribución Normal estándar $Z \sim N(0,1)$, dado que la varianza (σ^2) es conocida. La fórmula para el estadístico de prueba es:

$$Z = \frac{\bar{\mathbf{X}} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

donde $\bar{\mathbf{X}}$ es el estadístico media muestral

- Calculamos el valor observado del estadístico:

$$z = \frac{\bar{\mathbf{x}} - \mu_0}{\sigma/\sqrt{n}}$$

donde $\bar{\mathbf{x}}$ es la media muestral calculada con la muestra concreta

CH para la media (var conocida, Normal)

- Calculamos el $p - valor$ como sigue:
  - Contraste bilateral: $p - valor = P(|Z| \geq |z|)$
  - Contraste unilateral derecho: $p - valor = P(Z \geq z)$
  - Contraste unilateral izquierdo: $p - valor = P(Z \leq z)$
- Si esta probabilidad es menor o igual que el valor de referencia α entonces, rechazamos la hipótesis nula en favor de la alternativa.

CH para la media (var conocida, Normal)

Otra forma de plantear el contraste de hipótesis es determinando el rechazo de H_0. Para ello, debemos comparar el valor del estadístico Z con los valores críticos de la distribución Normal estándar.

- Para un contraste bilateral (dos colas): Rechaza H_0 si $|z| > z_{\alpha/2}$
- Para un contraste unilateral derecho (una cola): Rechaza H_0 si $z > z_\alpha$
- Para un contraste unilateral izquierdo (una cola): Rechaza H_0 si $z < -z_\alpha$

Aquí, z_α y $z_{\alpha/2}$ son los valores críticos (cuantiles) de la distribución Normal estándar correspondientes al nivel de significación α

Es decir, decidimos si rechazamos o no la hipótesis nula en función del valor del estadístico de prueba:

- Si su valor está en la región crítica, rechaza H_0
- Si su valor no está en la región crítica, no rechaces H_0

Ejemplo

Supón que queremos probar si la edad media de los profesores de la URJC es igual a 50 años con una desviación estándar conocida de 10 años, y tienes una muestra de 36 observaciones con una media muestral de 52

Contraste de hipótesis para la media de una población Normal con varianza desconocida

- Cuando la varianza poblacional es desconocida, se utiliza la desviación estándar muestral (S) y el estadístico de prueba se basa en la distribución t de Student con $n-1$ grados de libertad

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

donde $\bar{\mathbf{X}}$ es el estadístico media muestral

- Calculamos el valor observado del estadístico en base a la muestra concreta

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

CH para la media (var desconocida, Normal)

En función del tipo de contraste, se calculará el $p - valor$

- H_1: $\mu \neq \mu_0$ (Contraste bilateral), $p - valor = P(|T_{n-1}| > |t|)$
- H_1: $\mu > \mu_0$ (Contraste unilateral derecho), $p - valor = P(T_{n-1} > t)$
- H_1: $\mu < \mu_0$ (Contraste unilateral izquierdo), $p - valor = P(T_{n-1} < t)$

Aquí, T_{n-1} es una variable aleatoria con una distribución t de Student con $n - 1$ grados de libertad.

La decisión asociada al contraste es:

- Si el $p - valor \leq \alpha$, rechazar H_0
- Si el $p - valor > \alpha$, no rechazar H_0

Ejemplo

Supón que una empresa quiere verificar si el tiempo promedio de entrega de sus pedidos es mayor de 30 minutos. Toma una muestra aleatoria de 16 entregas y encuentra que el tiempo promedio de entrega es de 32 minutos con una desviación estándar muestral de 4 minutos. Realiza un contraste de hipótesis con un nivel de significación del 0.05 para ver si el tiempo promedio de entrega es mayor de 30 minutos.

CH igualdad de medias (muestras indep.)

- Cuando se desea comparar las medias de dos muestras independientes asumiendo que los datos siguen una distribución normal, se puede usar el contraste de hipótesis paramétrico conocido como la prueba t de Student para muestras independientes
- Este método es robusto y se basa en suposiciones claras acerca de la normalidad de las distribuciones subyacentes:
  - Las dos muestras son independientes
  - Los datos de cada muestra se distribuyen normalmente (con n grande, por el TCL, esto se cumplirá)
  - Las varianzas poblacionales son desconocidas, pero se pueden asumir iguales para una versión específica del test t (si esta suposición es razonable).

CH igualdad de medias (muestras indep.)

- Supongamos dos m.a.s. independientes con medias, desviaciones típicas y tamaño muestral igual a: $\bar{\mathbf{x}}_1$, $\bar{\mathbf{x}}_2$, s_1^2, s_2^2, n_1 y n_2, respectivamente.
- Formulamos las hipótesis:
  - Hipótesis nula H_0: Las medias de las dos poblaciones son iguales $\mu_1 = \mu_2$
  - Hipótesis alternativa H_1 Las medias de las dos poblaciones son diferentes $mu_1 \neq \mu_2$
- Consideramos un nivel de significancia estadística α

- Calculamos el estadístico muestral, en este caso

$$t = \frac{\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2}{SE}$$

donde:

$$SE = \sqrt{S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

siendo S_p la desviación típica combinada:

$$S_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

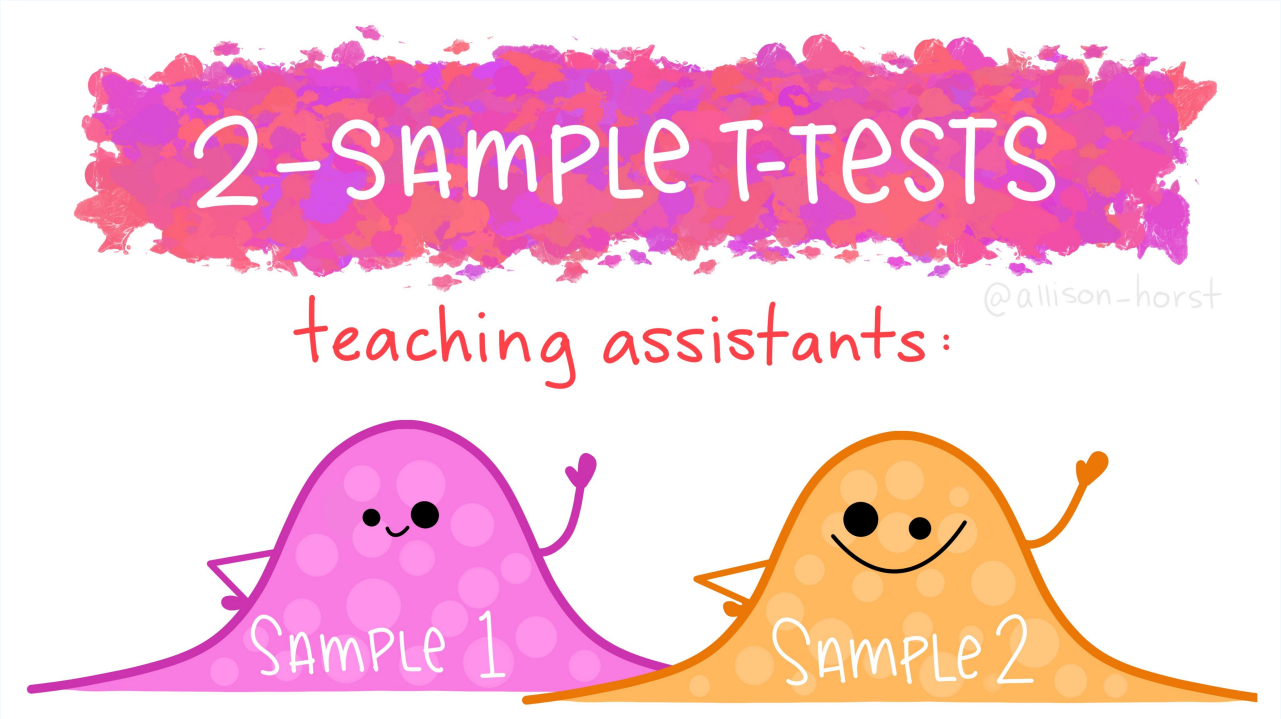
CH igualdad de medias (muestras indep.)

- Para un nivel de significancia $\alpha = 0.05$ y grados de libertad $df = n_1 + n_2 - 2$, buscamos el valor crítico $T_{df,1-\alpha/2}$ para la distribución t de Student para una prueba de dos colas.
- Comparamos el valor del estadístico t calculado con las muestras con el valor crítico:
  - Si $|t| > T_{df,1-\alpha/2}$, rechazamos H_0
  - Si $|t| \leq T_{df,1-\alpha/2}$, no rechazamos H_0

CH igualdad de medias (muestras indep.)

El t-test de forma gráfica (¡y divertida!) gracias a las ilustraciones de Allison Horst:

`https://twitter.com/allison_horst/status/1216411185240690688`



Ejemplo

Supongamos que un investigador quiere comparar la efectividad de dos métodos de enseñanza de matemáticas. Se seleccionan dos grupos de estudiantes al azar, uno para cada método. Después de un semestre, se mide el puntaje de un examen final de matemáticas.

Datos:

- Grupo A (Método 1):
  - Tamaño de la muestra $(n_1) = 12$
  - Puntajes: 85, 78, 92, 88, 75, 84, 90, 91, 83, 79, 87, 86
- Grupo B (Método 2):
  - Tamaño de la muestra $(n_2) = 10$
  - Puntajes: 82, 77, 85, 80, 79, 81, 83, 78, 82, 76

Consideramos un nivel de significancia $\alpha = 0.05$.

CH igualdad de proporciones

El contraste de hipótesis para la diferencia de proporciones se utiliza para determinar si hay una diferencia significativa entre las proporciones de éxito en dos grupos independientes. Supongamos dos variables aleatorias X e Y que siguen una distribución binomial de parámetros p_1 y p_2 respectivamente.

Formulamos las hipótesis:

- Hipótesis nula H_0: Las proporciones de las dos poblaciones son iguales $p_1 = p_2$
- Hipótesis alternativa H_1: Las proporciones de las dos poblaciones son diferentes $p_1 \neq p_2$

CH igualdad de proporciones

- Consideremos dos m.a.s. de tamaño n_1 y n_2, siendo $\mathbf{x}$ y $\mathbf{y}$ el número de observaciones que cumplen un criterio, de modo que:

$$\hat{p}_1 = \frac{\mathbf{x}}{n_1}, \hat{p}_2 = \frac{\mathbf{y}}{n_1}$$

  son los estimadores de máxima verosimilutd de p_1 y p_2, respectivamente
- Consideramos un nivel de significancia estadística α
- Calculamos el estadístico muestral, en este caso:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{SE}$$

  donde

$$SE = \sqrt{\hat{p}(1-\hat{p})}\left(\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

  siendo $\hat{p} = \frac{\mathbf{x}+\mathbf{y}}{n_1+n_2}$

CH igualdad de proporciones

- Para valores grandes de n_1 y n_2, la distribución de Z es $N(0,1)$
- Para un nivel de significancia $\alpha = 0.05$, buscamos el valor crítico $Z_{1-\alpha/2}$ para la distribución Normal
- Comparamos el valor del estadístico $Z = z$ calculado con el valor crítico:
  - Si $|Z| > Z_{1-\alpha/2}$, rechazamos H_0
  - Si $|Z| \leq Z_{1-\alpha/2}$, no rechazamos H_0

Ejemplo

Supongamos que una empresa de marketing quiere evaluar la efectividad de dos campañas publicitarias diferentes (Campaña A y Campaña B) para atraer clientes. La empresa desea saber si hay una diferencia significativa en la proporción de clientes que responden positivamente a cada campaña.

- Campaña A:
  - Número de personas que recibieron la campaña: 500
  - Número de personas que respondieron positivamente: 75
- Campaña B:
  - Número de personas que recibieron la campaña: 600
  - Número de personas que respondieron positivamente: 120

Referencias

Wasserman, L. (2013). All of Statistics: a Concise Course in Statistical Inference.

Spiegel, M., & Stephens, L. (2009). Estadística–Serie Schaum. *Mc Graw-Hill*.

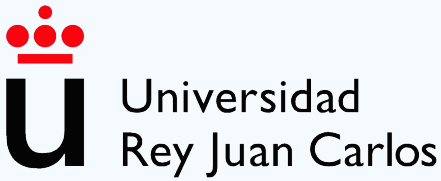
Gomez Villegas, M. A. (2005). *Inferencia estadística*. Ediciones Díaz de Santos.

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

Inferencia no paramétrica

Inferencia Estadística - Grado en Ciencia e Ingeniería de Datos

Curso académico 2024-2025



Inferencia no paramétrica

¿Qué son? Los contrastes no paramétricos son métodos estadísticos utilizados para probar hipótesis cuando no se cumplen los supuestos necesarios para los contrastes paramétricos (por ejemplo, la normalidad)

- **Características Principales:**
  - **Flexibilidad**: No requieren que los datos sigan una distribución específica (se usa la Función de Distribución Empírica (EDF, por sus siglas en inglés)).
  - **Robustez**: Menos sensibles a valores atípicos y a desviaciones de supuestos de normalidad.
  - **Usos frecuentes**: Cuando las muestras son pequeñas, los datos son ordinales o tienen distribuciones asimétricas.

- **Aplicaciones**:
  - Comparación de medianas entre grupos.
  - Evaluación de relaciones de orden entre variables.
  - Pruebas de independencia entre variables categóricas.
- **Ejemplos**:
  - Prueba de Kolmogorov-Smirnov: Compara la EDF de dos muestras o de una muestra con una distribución teórica.
  - Prueba de Mann-Whitney: Utiliza la posición o el orden de los datos
  - Prueba de Kruskal-Wallis: Comparación de varias muestras
  - Prueba de Chi-cuadrado: Análisis de independencia para variables categóricas.

Función de distribución empírica

- **Definición**: Es una función que estima la distribución de una muestra de datos observados. Se construye como una función escalonada que aumenta en 1/n en cada observación de la muestra.
- **Uso en contrastes no paramétricos**: La FDE es fundamental en contrastes no paramétricos, pues permite trabajar con los datos de manera directa, sin asumir una forma teórica para su distribución, proporciona una estimación de la función de distribución acumulada de una muestra de datos

Función de distribución empírica

- Dada una muestra de datos $(X_1, X_2, \ldots, X_n)$, la función de distribución empírica $F_n(x)$ se define como:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$$

  donde $I(X_i \leq x)$ es una función indicadora que toma el valor 1 si $X_i \leq x$ y 0 en caso contrario
- En otras palabras, $F_n(x)$ es la proporción de valores en la muestra que son menores o iguales a x

Propiedades de la Función de Distribución Empírica

1. Escalonada: La EDF es una función escalonada que incrementa en pasos de $1/n$ en cada punto de datos
2. No decreciente: La EDF nunca disminuye a medida que x aumenta.
3. Límites La EDF varía entre 0 y 1. Específicamente, $F_n(x) = 0$ para x menor que el valor mínimo de la muestra y $F_n(x) = 1$ para x mayor que el valor máximo de la muestra

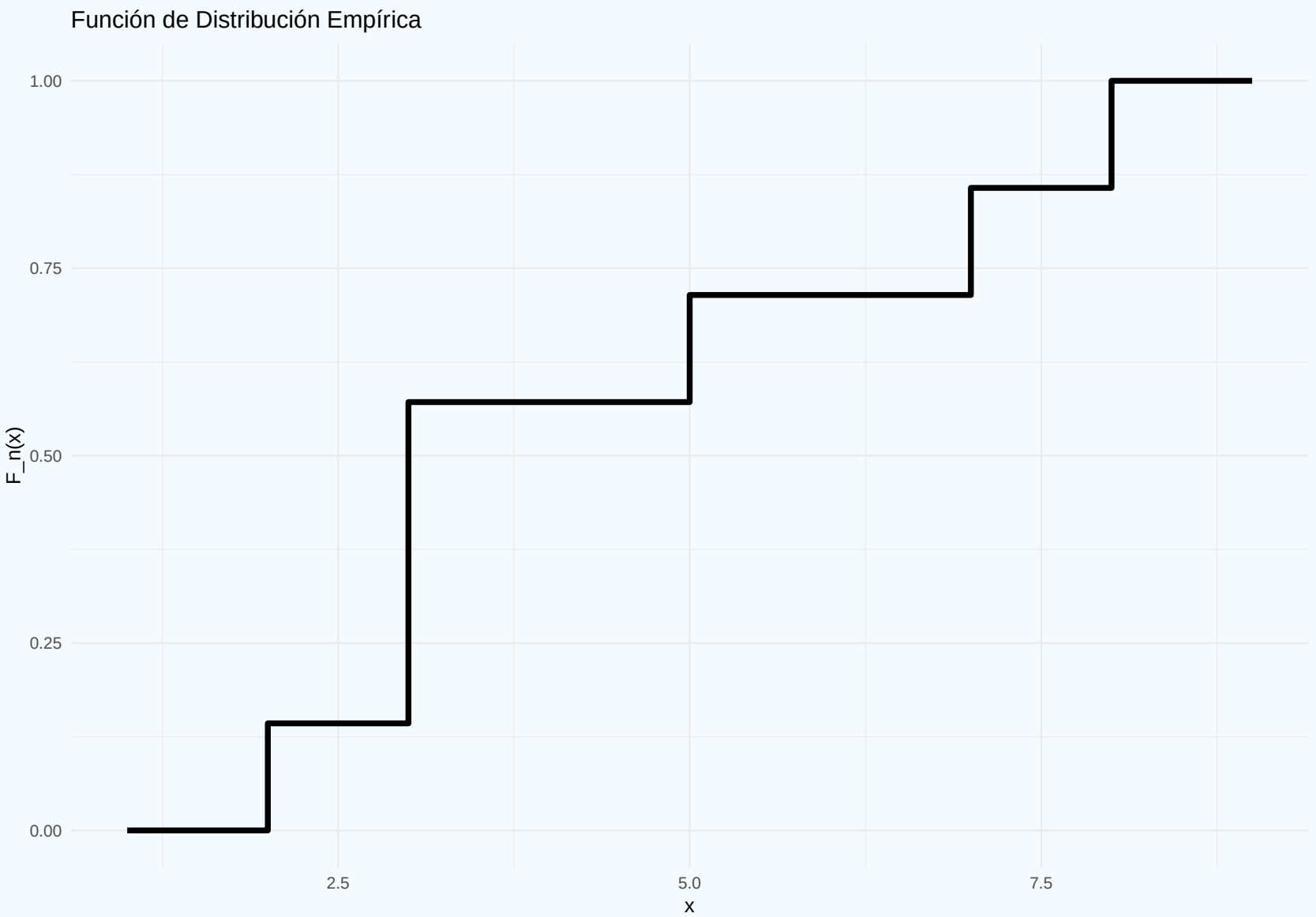
Función de distribución empírica

Calculemos la función de distribución empírica de los datos $X = \{8, 3, 5, 3, 7, 3, 2\}$.

1. Ordenador los datos: $\{2, 3, 3, 3, 5, 7, 8\}$
2. Calcular la función de distribución empírica $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$ para nuestras $n = 7$ observaciones

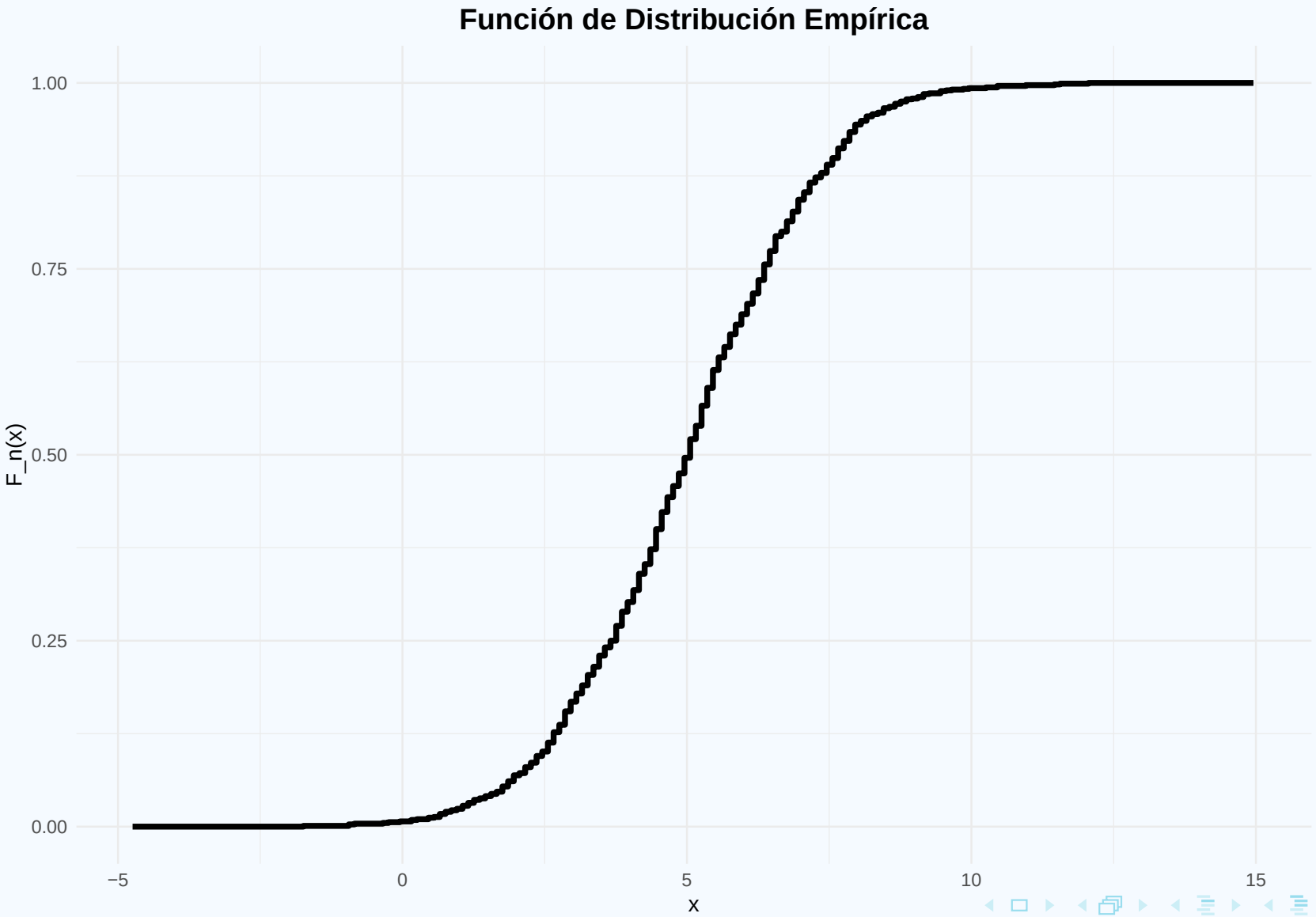
$$F_n(x) = \begin{cases} 0 & \text{si } x < 2 \\ \frac{1}{7} = 0.143 & \text{si } 2 \leq x < 3 \\ \frac{4}{7} = 0.571 & \text{si } 3 \leq x < 5 \\ \frac{5}{7} = 0.714 & \text{si } 5 \leq x < 7 \\ \frac{6}{7} = 0.857 & \text{si } 7 \leq x < 8 \\ 1 & \text{si } x \geq 8 \end{cases}$$

Función de distribución empírica



Función de distribución empírica

Con muchos datos:



Prueba Chi-cuadrado de independencia

- La prueba chi-cuadrado de independencia se utiliza para determinar si hay una asociación significativa entre dos variables categóricas X con categorías $X_1, X_2, \ldots, X_r$ e Y con categorías $Y_1, Y_2, \ldots, Y_c$
- Esta prueba compara las frecuencias observadas en la tabla de contingencia con las frecuencias esperadas bajo la hipótesis de independencia

	Y_1	$\ldots$	Y_j	$\ldots$	Y_c	
X_1	$n_{1,1}$	$\ldots$	$n_{1,j}$	$\ldots$	$n_{1,c}$	$n_{1.}$
$\ldots$						$\ldots$
X_i	$n_{i,1}$	$\ldots$	$n_{i,j}$	$\ldots$	$n_{i,c}$	$n_{i.}$
$\ldots$						$\ldots$
X_r	$n_{r,1}$	$\ldots$	$n_{r,j}$	$\ldots$	$n_{r,c}$	$n_{r.}$
	$n_{.1}$		$n_{.j}$		$n_{.c}$	$n_{..}$

Prueba Chi-cuadrado de independencia

La hipótesis nula H_0 de esta prueba es que no hay asociación entre las variables, esto es que las variables implicadas son independientes:

- **Hipótesis nula** H_0: No hay asociación entre las variables (son independientes)
- **Hipótesis alternativa** H_1: Hay una asociación entre las variables (son dependientes)

Las frecuencias esperadas se calculan como sigue:

$$E_{ij} = \frac{(n_{i.} \times n_{.j})}{N}$$

donde:

- E_{ij} es la frecuencia esperada en la celda (i, j)
- $n_{i.}$ es el total de la fila i
- $n_{.j}$ es el total de la columna j
- $N = n_{..}$ es el total general

Prueba Chi-cuadrado de independencia

Ahora, comparamos las frecuencias esperadas con las frecuencias observadas, definiendo con ellos el estadístico chi-cuadrado:

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

donde O_{ij} es la frecuencia observada en la celda (i, j).

Bajo la hipótesis nula, el estadístico de prueba sigue una distribución chi-cuadrado con $(r-1)(c-1)$ grados de libertad, donde r es el número de filas y c es el número de columnas. Podemos calcular el $p-valor$ como:

$$p - valor = P(\chi^2_{(r-1)(c-1)} \geq \chi^2_{observado})$$

Como ocurría en los contrastes de hipótesis paramétricos, comparamos el $p-valor$ con el nivel de significancia α, generalmente 0.05. Si $p-valor < \alpha$, se rechaza la hipótesis nula.

Ejemplo

Supongamos que un investigador desea determinar si hay una asociación entre el tipo de dispositivo usado (Laptop, Tablet, Smartphone) y la satisfacción del usuario (Satisfecho, No Satisfecho).

Recolectamos datos de una muestra de 150 usuarios y construimos la siguiente tabla de contingencia:

	Satisfecho	No Satisfecho	Total
Laptop	30	10	40
Tablet	20	20	40
Smartphone	50	20	70
Total	100	50	150

¿Existe asociación entre las variables?

Prueba Chi-cuadrado de independencia

Cuando la tabla de contingencia es pequeña (2x2), el estadístico chi-cuadrado puede devolver valores demasiados pequeños (lo que aumenta la probabilidad de error de tipo I).

Soluciones planteadas:

- Corrección de continuidad de Yates

$$\chi^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij} - 0.5)^2}{E_{ij}}$$

- Test exacto de Fisher (tabla 2x2 y alguna celda con menos de 5 obs)

Prueba de bondad de ajuste

Esta prueba se utiliza para determinar si una distribución de frecuencias observadas $F(x)$ sigue una distribución teórica esperada $F_0(x)$ (Normal, exponencial, Poisson, etc.). Este tipo de pruebas se llaman pruebas de bondad de ajuste (test of goodness of fit) y contrastan:

$$H_0 : F(x) = F_0(x) \qquad \forall x$$

$$H_1 : F(x) \neq F_0(x)$$

En particular, veremos la prueba Chi-cuadrado de bondad de ajuste

Prueba Chi-cuadrado de bondad de ajuste

Dada la muestra aleatoria simple $X_1, \ldots, X_n$ de n observaciones se pretende analizar si concuerdan con una distribución específica conocida $F_0(x)$:

- Hipótesis nula H_0: Las frecuencias observadas siguen la distribución esperada $F_0(x)$
- Hipótesis alternativa H_1: Las frecuencias observadas no siguen la distribución esperada

Prueba Chi-cuadrado de bondad de ajuste

Para ello, vamos a dividir el dominio en k trocitos y vamos a comparar las frecuencias de ambas distribuciones en ellos. Por ejemplo, en el caso real, se divide la recta real en: $(-\infty, b_1], (b_1, b_2], \ldots, (b_{k-1}, \infty)$

Como la distribución $F_0(x)$ es conocida, sabemos cuáles son las probabilidades de dichos intervalos: $p_1, \ldots, p_k$, $\sum_{i=1}^{k} p_i = 1$. Por tanto, según la distribución teórica, el número esperado de observaciones en cada intervalo es $E_i = np_i$, $\forall i = 1, \ldots, k$.

Prueba Chi-cuadrado de bondad de ajuste

Por otro lado, sabemos las observaciones O_i de la muestra que caen en cada intervalo (siendo $\sum_{i=1}^{k} O_i = n$).

Finalmente, se comparan las frecuencias observadas O_i con las frecuencias esperadas $E_i = np_i$ con el estadístico Chi-cuadrado

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{k-1,1-\alpha}$$

Ejemplo

Supongamos que un investigador quiere determinar si los resultados de un dado son uniformemente distribuidos. El dado se lanza 60 veces y los resultados son los siguientes:

- 1: 8 veces
- 2: 10 veces
- 3: 9 veces
- 4: 11 veces
- 5: 12 veces
- 6: 10 veces

Queremos comprobar si estos resultados siguen una distribución uniforme, es decir, cada número tiene la misma probabilidad de 1/6.

Pruebas de homogeneidad

- La prueba no paramétrica de homogeneidad se utiliza para determinar si dos o más muestras independientes provienen de la misma distribución o de distribuciones similares
- Ejemplo de pruebas no paramétricas de homogeneidad:
  - **Prueba de Kolmogorov-Smirnov** para dos muestras: Compara dos muestras para verificar si provienen de la misma distribución
  - **Prueba de Mann-Whitney U** (o Wilcoxon Rank-Sum Test): Compara dos muestras independientes para determinar si tienen la misma distribución
  - **Prueba de Kruskal-Wallis**: Extiende la prueba de Mann-Whitney U a más de dos muestras independientes

Prueba de Kolmogorov-Smirnov para dos muestras

- La prueba de Kolmogorov-Smirnov (K-S) para dos muestras independientes es una prueba no paramétrica utilizada para determinar si dos muestras independientes provienen de la misma distribución
- A diferencia de otras pruebas que se centran en comparar medias o varianzas, la prueba K-S compara las distribuciones acumuladas de dos muestras
- Las hipótesis de la prueba son:
  - H_0: Las dos muestras provienen de la misma distribución
  - H_1: Las dos muestras provienen de distribuciones diferentes

Prueba de Kolmogorov-Smirnov para dos muestras

- Sea $X_1, \ldots, X_{n_1}$ una m.a.s de una población 1 donde las X_i son independientes e idénticamente distribuidas y sea $Y_1, \ldots, Y_{n_2}$ otra m.a.s de una población 2 donde las Y_j son independientes e idénticamente distribuidas
- Para cada muestra, se construyen las funciones de distribución empírica (EDF)
- Calculamos el estadístico D de la prueba K-S que es la máxima diferencia absoluta entre las dos funciones de distribución empírica:

$$D = \sup_x |F_{n_1}(x) - F_{n_2}(x)|$$

donde, $F_{n_1}(x)$ y $F_{n_2}(x)$ son las funciones de distribución empírica de las dos muestras

Prueba de Kolmogorov-Smirnov para dos muestras

- El $p - valor$ se determina utilizando la distribución del estadístico D bajo la hipótesis nula de que ambas muestras provienen de la misma distribución
- El cálculo exacto del $p - valor$ para la prueba de K-S no es trivial y generalmente se realiza mediante métodos numéricos o tablas pre-calculadas. Sin embargo, se puede aproximar utilizando la distribución asintótica del estadístico D

Prueba de Kolmogorov-Smirnov para dos muestras

- Para muestras grandes, el $p - valor$ se puede aproximar usando la fórmula:

$$p \approx Q_{KS}(\sqrt{n}D)$$

donde:

  - $n = \frac{n_1 \cdot n_2}{n_1 + n_2}$ es el número efectivo de muestras
  - D es el valor del estadístico K-S
  - Q_{KS} es una función que representa la cola superior de la distribución de Kolmogorov-Smirnov

- La función Q_{KS} para grandes valores de n se puede aproximar usando la siguiente fórmula:

$$Q_{KS}(\lambda) = 2 \sum_{k=1}^{\infty} (-1)^{k-1} e^{-2k^2\lambda^2}$$

donde $\lambda = \sqrt{n}D$

Prueba de Mann-Whitney U

- La prueba de Mann-Whitney U, también conocida como prueba de Wilcoxon para muestras independientes, es una prueba no paramétrica que se utiliza para contrastar si dos muestras independientes provienen de la misma distribución
- Es una alternativa a la prueba t de Student cuando no se cumplen los supuestos de normalidad. En lugar de trabajar con los valores originales, la prueba utiliza los rangos de los datos
- Esta prueba se basa en combinar y ordenar juntas ambas muestras. Si en dicha ordenación:
  1. Los valores de ambas muestras se mezclan de forma aleatoria –> entenderemos que las muestras no son distintas
  2. Los valores de cada muestra quedan claramente agrupados –> las muestras son distintas

Prueba de Mann-Whitney U

Queremos contrastar si el tratamiento A y el tratamiento B tienen el mismo efecto

Tx	A	A	A	A	B	A	B	B	B	B
Rango	1	2	3	4	5	6	7	8	9	10

(a)

Tx	B	A	A	B	B	A	B	A	B	A
Rango	1	2	3	4	5	6	7	8	9	10

(b)

Prueba de Mann-Whitney U

- En primer lugar se combinan los datos de ambas muestras y se ordena el total de los valores de menor a mayor
- A continuación se asignan rangos a estos valores. Los empates se gestionan otorgando a los valores iguales el rango promedio
- Se calcula el estadístico del contraste tal y como sigue:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

donde:

  - n_1 y n_2 son los tamaños de las dos muestras
  - R_1 y R_2 son la suma de los rangos de las muestras 1 y 2, respectivamente

- El estadístico U final es $U = min(U_1, U_2)$

Prueba de Mann-Whitney U

- El $p-valor$ se determina comparando el estadístico U con una distribución de referencia para U (tabla de Mann-Whitney)
- Si el tamaño muestral es grande, se puede determinal la significatividad mediante la aproximación normal para grandes tamaños de muestra:

$$Z = \frac{U - E(U)}{\sqrt{Var(U)}} \sim N(0,1)$$

siendo

$$E(U) = \frac{n_1 n_2}{2} \qquad Var(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

Y podemos calcular el $p-valor$ como hemos hecho en métodos anteriores: $p-valor = P(Z > z)$. Siendo z el valor del estadístico calculado en las muestras

- Comparamos el $p-valor$ p con el nivel de significancia α, generalmente 0.05. Si $p-valor < \alpha$, se rechaza la hipótesis nula.

Ejemplo

Supongamos que queremos comparar los tiempos de entrega (en días) de dos proveedores distintos:

- Proveedor A: $2, 3, 5, 6, 8$
- Proveedor B: $1, 4, 4, 7, 9$

Prueba de Kruskal-Wallis

- La prueba de Kruskal-Wallis es una prueba no paramétrica utilizada para comparar tres o más muestras independientes para determinar si provienen de la misma distribución
- Es una extensión de la prueba de Mann-Whitney U a más de dos grupos y una alternativa robusta a la ANOVA cuando no se cumplen los supuestos de normalidad y homogeneidad de varianzas
- Dado que es una prueba no paramétrica, no requiere que los datos provengan de una distribución normal. Al igual que la prueba de Mann-Whitney, la prueba de Kruskal-Wallis trabaja con los rangos de los datos en lugar de los valores originales

Las hipótesis son:

- H_0: Todas las muestras provienen de la misma distribución
- H_1: Al menos una de las muestras proviene de una distribución diferente

Prueba de Kruskal-Wallis

- Se comienza combinando los datos de todas las muestras y se ordenan los valores de menor a mayor
- A continuación se asignan rangos R_i a estos valores. Los empates se gestionan otorgando a los valores iguales el rango promedio
- El estadístico Kruskal-Wallis H se define como:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

donde N es el tamaño total de la muestra (suma de los tamaños de las k muestras), R_i es la suma de los rangos de la i-ésima muestra, n_i es el tamaño de la i-ésima muestra y k es el número de muestras

-

Prueba de Kruskal-Wallis

- El $p - valor$ se determina comparando el estadístico H con una distribución de referencia para H (tabla de Kruskal-Wallis)
- Para tamaños de muestra relativamente grandes el estadístico H sigue una distribución χ^2 con $k - 1$ grados de libertad. El $p - valor$ se determina como

$$p - valor = P(H > h)$$

siendo h el valor que toma H en las muestras en cuestión

Ejemplo

Unos investigadores estaban interesados en estudiar la interacción social de distintos adultos para estudiar si la interacción social puede vincularse a la confianza en uno mismo

Para ello, clasificaron a 17 participantes en tres grupos en función de la interacción social exhibida: alta, media, baja

Después de clasificar a los participantes en los tres grupos, se les pidió que completaran una autoevaluación de la autoconfianza en una escala de 25 puntos:

Alta	Media	Baja
21, 23, 18, 12, 19, 20	19, 5, 10, 11, 9	7, 8, 15, 3, 6, 4

Se quiere determinan si existe diferencia entre alguno de los 3 grupos.

Prueba sobre muestras pareadas

- **Muestras pareadas**: las muestras no son independientes, están relacionadas, emparejadas entre sí
- Ejemplos:
  - Nivel de colesterol en un grupo de personas antes y después de tomar un medicamento
  - Medidas tomadas en los dos pies
- Contrastes:
  - Prueba del signo (sign test)
  - Prueba de rangos de signos de Wilcoxon (Wilcoxon signed-rank test)

Prueba del signo

- Prueba no paramétrica utilizada para comparar dos muestras relacionadas o emparejadas
- Se emplea cuando se tienen dos conjuntos de datos dependientes y se desea determinar si hay una diferencia significativa en sus medianas
- Alternativa a la prueba de la t-Student cuando no se cumplen las hipótesis
- Simplifica la comparación entre las muestras mediante una binarización de la misma: convierte los resultados en "+" y "-" y lo compara en esa versión

Prueba del signo

- Sean dos muestras relacionadas $X_1 \ldots, X_n$ e $Y_1, \ldots, Y_n$ de tamaño n
- Comenzamos calculando las diferencias entre pares: Para cada par (X_i, Y_i), calcular la diferencia $D_i = X_i - Y_i$
- Contar los signos
  - S_+ es el número de diferencias positivas $D_i > 0$
  - S_- es el número de diferencias negativas $D_i < 0$
  - Ignorar las diferencias que son cero $D_i = 0$
- Estadístico de Prueba: $S = min(S_+, S_-)$

Prueba del signo

- Determinar el valor crítico. Consultar una tabla de la distribución binomial para obtener el valor crítico correspondiente al nivel de significancia α y el tamaño de la muestra efectiva n (número de pares no nulos). Podemos calcular el valor crítico como sigue:

$$p - valor = P(S \leq s)$$

  donde $S \sim Binomial(n, p = 0.5)$

- Decisión:
  - Rechazar H_0 si el estadístico de la prueba es menor o igual al valor crítico
  - No rechazar H_0 si el estadístico de la prueba es mayor que el valor crítico

Prueba del signo

- Si $S_+ + S_- \geq 25$, se puede usar la siguiente formula

$$z = \frac{max(S_+, S_-) - 1/2(S_+ + S_-) - 1/2}{1/2\sqrt{S_+ + S_-}} \sim N(0,1)$$

Así, $p - value = P(Z \leq z)$ en el contraste unilateral y se multiplicará por 2 en el bilateral.

Prueba de rangos de signos de Wilcoxon

- La prueba de Wilcoxon de rangos de signos se utiliza para comparar muestras pareadas
- Es una alternativa no paramétrica a la prueba t de muestras pareadas. En lugar de comparar medias, esta prueba compara las medianas de las diferencias entre las dos muestras pareadas
- Es una prueba ideal para muestras pequeñas cuando la normalidad no puede ser asumida

Prueba de rangos de signos de Wilcoxon

Procedimiento:

1. **Calcular las diferencias** entre cada par de observaciones
2. **Ordenar las diferencias** en valor absoluto y asignarles rangos, ignorando las diferencias que sean cero
3. **Asignar signos** a los rangos de acuerdo con el signo de las diferencias originales
4. **Calcular la suma de los rangos positivos** y la suma de los rangos negativos

5. **Determinar el estadístico de prueba**: El estadístico de Wilcoxon es el menor de las dos sumas de rangos:

$$T = min(\sum R_+, \sum R_-)$$

siendo R_+ la suma de los rangos con diferencias positivas y R_- la suma de los rangos con diferencias negativas

6. **Significatividad.** Se compara el estadístico de prueba$T$ con los valores críticos de la tabla de Wilcoxon para determinar la significancia estadística

Prueba de rangos de signos de Wilcoxon

Si el tamaño muestral es grande, se puede determinal la significatividad mediante la aproximación normal para grandes tamaños de muestra:

$$Z = \frac{T - E(T)}{Var(T)} - \sim N(0,1)$$

siendo

$$E(T) = \frac{n(n+1)}{4} \qquad Var(T) = \frac{n(n+1)(2n+1)}{24}$$

con n el número de muestras pareadas

Con ello, $p - valor = P(Z > z_{\alpha/2})$

Referencias

Corder, G. W., & Foreman, D. I. (2014). *Nonparametric statistics: A step-by-step approach*. John Wiley & Sons.

Deshpande, J. V., Naik-Nimbalkar, U., & Dewan, I. (2017). *Nonparametric statistics: theory and methods*. World Scientific.

Gomez Villegas, M. A. (2005). *Inferencia estadística*. Ediciones Díaz de Santos.

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

Análisis de la varianza

Inferencia Estadística - Grado en Ciencia e Ingeniería de Datos

Curso académico 2024-2025

Universidad
Rey Juan Carlos

Introducción

- En temas previos hemos comparado las medias de dos poblaciones
  $H_0 : \mu_1 = \mu_2$
  $H_1 : \mu_1 \neq \mu_2$
- ¿Y si queremos comprobar si hay diferencia entre 3 o más medias muestrales? Es decir, comprobar la hipótesis de que todas las medias (3 o más) son iguales –> **ANOVA**

ANOVA

- **Análisis de la varianza**, conocido como **ANOVA** (del inglés *Analysis of Variance*)
- Técnica estadística utilizada para comparar las medias de dos o más grupos y determinar si existen diferencias significativas entre ellos
- Desarrollado por Fisher en las primeras décadas del siglo XX.
- La idea central es analizar la variabilidad de los datos y dividirla en componentes atribuibles a diferentes fuentes de variación

- En su forma más simple, el ANOVA se utiliza para probar hipótesis sobre las diferencias entre las medias de grupos
- Por ejemplo, si quisiéramos comparar el rendimiento de tres tipos diferentes de métodos educativos, podríamos usar ANOVA para determinar si el método educativo tiene un efecto significativo en el rendimiento académico:

Tipo de método	Rendimiento académico
Método 1	85, 78, 90, 82
Método 2	88, 79, 91, 85
Método 3	80, 75, 88, 83

Procedimiento general

1. **Formulación de hipótesis**: Se establece una hipótesis nula que indica que no hay diferencias entre las medias de los grupos y una hipótesis alternativa que sugiere que al menos una media es diferente
   $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
   H_1 : al menos una de las medias es distinta
2. **Cálculo de la varianza**: Se calculan dos tipos de varianza: la varianza dentro de los grupos (variabilidad debido a diferencias dentro de los mismos grupos) y la varianza entre los grupos (variabilidad debido a diferencias entre los grupos)

Procedimiento general

3. **F-test**: Se realiza una prueba F de Fisher para evaluar la relación entre las varianzas. Si la varianza entre los grupos es significativamente mayor que la varianza dentro de los grupos, esto sugiere que hay diferencias significativas entre las medias de los grupos
4. **Análisis de resultados**: Si la prueba F indica que hay diferencias significativas, se pueden realizar pruebas adicionales para identificar entre qué grupos existen estas diferencias.

- Es una herramienta poderosa porque permite comparaciones múltiples mientras controla la tasa de error tipo I
- Es ampliamente utilizado en experimentos donde se comparan tratamientos o condiciones en diferentes grupos o en diferentes momentos

Distribución F (Fisher-Snedecor)

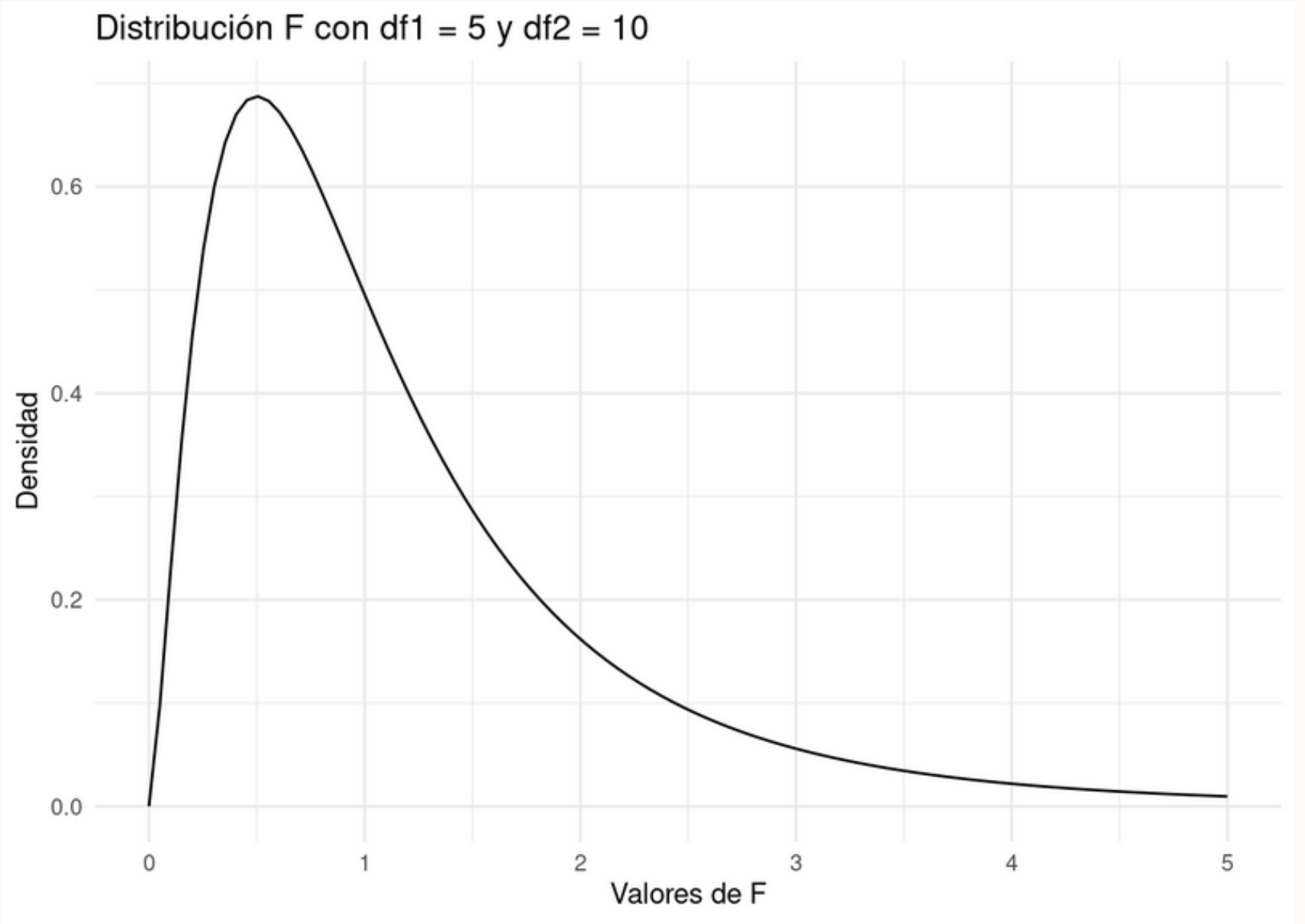
- Necesitamos una distribución muestral para comparar varianzas
- Se toma el estadístico cociente S_X^2/S_Y^2. Cuanto más próximo a 1, más parecidas las varianzas
- La distribución muestral de S_X^2/S_Y^2 es la distribución F
- Dadas dos muestras $\mathbf{X} = (X_1, \ldots, X_{n+1})$, $\mathbf{Y} = (Y_1, \ldots, Y_{m+1})$ de tamaño n y m, procedentes de dos poblaciones Normales con varianzas σ_X^2 y σ_Y^2. Entonces

$$\frac{nS_X^2}{\sigma_X^2} \sim \chi_n^2 \qquad \frac{mS_Y^2}{\sigma_Y^2} \sim \chi_m^2$$

y

$$\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma^2 Y} \sim F_{m,n}$$

Distribución F



Modelo con un factor

- Se utiliza cuando se estudia el efecto de un solo factor (variable independiente) en una variable dependiente continua
- Permite comparar las medias de varios grupos para determinar si existen diferencias significativas entre ellos
- **Hipótesis**:
  - **Hipótesis Nula (H_0)**: Todas las medias de los grupos son iguales ($\mu_1 = \mu_2 = \cdots = \mu_k$).
  - **Hipótesis Alternativa (H_1)**: Al menos una de las medias de los grupos es diferente.

Modelo con un factor

Tenemos una variable aleatoria Y que toma valores reales y una variable cualitativa o factor X con k niveles $1, 2, \ldots, i, \ldots, k$. La variable Y toma valores $Y_{ij}, j = 1, \ldots, n_i$ en el nivel i del factor X, siendo n_i el número de observaciones en el nivel i del factor X

X_1	$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$
$\vdots$	$\vdots$
X_i	$Y_{i1}, Y_{i2}, \ldots, Y_{in_i}$
$\vdots$	$\vdots$
X_k	$Y_{k1}, Y_{k2}, \ldots, Y_{kn_k}$

Modelo con un factor

Tenemos los siguientes supuestos:

- Normalidad: Las distribuciones de las poblaciones de las que provienen las muestras son normales. Se supone que los errores ϵ_{ij} están distribuidos como una Normal de media 0 y varianza σ^2
- Homogeneidad de varianzas: Las varianzas de las poblaciones son iguales
- Independencia: Las observaciones son independientes entre sí

Modelo con un factor

El modelo teórico es como sigue:

$$Y_{ij} = \mu + \tau_i + \epsilon_{ij}$$

Donde:

- Y_{ij} es la observación j-ésima del grupo i-ésimo
- μ es la media general
- ϵ_{ij} es el término de error aleatorio, $\epsilon_{ij} \sim N(0, \sigma^2)$
- τ_i es el efecto del grupo i-ésimo en la media de la variable respuesta Y. Esto es, cuánto aumenta o disminuye la media de Y por pertenecer la observación a la categoría i. De modo que podemos llamar

$$Y_i = \mu + \tau_i$$

al efecto medio del grupo i-ésimo.

La suma de las diferencias al cuadrado de cada dato respecto a la media general se calcula como sigue:

$$\text{SST} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\cdot\cdot})^2$$

donde $\bar{Y}_{\cdot\cdot}$ es la media general de todas las observaciones.

Teniendo en cuenta que: $Y_{ij} - \bar{Y}_{..} = Y_i + \epsilon_{ij} - \bar{Y}_{..}$

Podemos descomponer la suma de cuadrados, como sigue:

$$\text{SST} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{..})^2 = \underbrace{\sum_{i=1}^{k} n_i(\bar{Y}_i - \bar{Y}_{..})^2}_{SSB} + \underbrace{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2}_{SST}$$

¿Qué son SSB y SST? ¿Qué interpretación le dais?

- SSB: La varianza entre grupos se calcula como la suma de las diferencias al cuadrado de las medias de los grupos respecto a la media general, ponderada por el tamaño de los grupos:

$$\text{SSB} = \sum_{i=1}^{k} n_i (\bar{Y}_i - \bar{Y}_{..})^2$$

  donde $\bar{Y}_i$ es la media del grupo i.
- SSW: La varianza dentro de los grupos es la suma de las diferencias al cuadrado de cada dato respecto a la media de su grupo

$$\text{SSW} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

Esto es, se descompone la variabilidad total de los datos en dos componentes, SSB que refleja la diferencia de cada grupo respecto a la media global y SSW que refleja la variabilidad intrínseca dentro de cada grupo:

$$\text{SST} = \text{SSB} + \text{SSW}$$

Nota:

- SST: Sum of squares total
- SSB: Sum of squares between
- SSW: Sum of squares within

Cálculo del Estadístico F:

$$F = \frac{\text{Varianza Entre Grupos (MSB)}}{\text{Varianza Dentro de los Grupos (MSW)}}$$

Donde:

- MSB (Mean Square Between): Media cuadrática entre grupos.
- MSW (Mean Square Within): Media cuadrática dentro de los grupos.

Esto es:

$$F = \frac{SSB/df_B}{SSW/df_W}$$

siendo $df_B = k - 1$ son los grados de libertad entre los grupos y $df_W = N - k$ son los grados de libertad dentro de los grupos y N es el número total de observaciones.

Una vez se dispone de toda esta información, es común representarla en forma de tabla, en la llamada *Tabla ANOVA*:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado Medio
Disferencias entre grupos	SSB	k-1	MSB
Diferencias dentro de los grupos, Residual o Error	SSW	N-k	MSW
Total	SST	N-1	

El estadístico de prueba $F \sim F_{df_B, df_W}$ bajo la hipótesis nula de igualdad de medias.

El $p - valor$ se obtiene a partir de la distribución F, considerando los grados de libertad de los numeradores y denominadores. Esto es:

$$p - valor = P(F_{df_b, df_W} > F_{muestral})$$

Como en otros contrastes, si el $p - valor$ es menor que el nivel de significancia α, se rechaza la hipótesis nula, concluyendo que al menos una de las medias de los grupos es diferente.

Modelo con un factor: Ejemplo

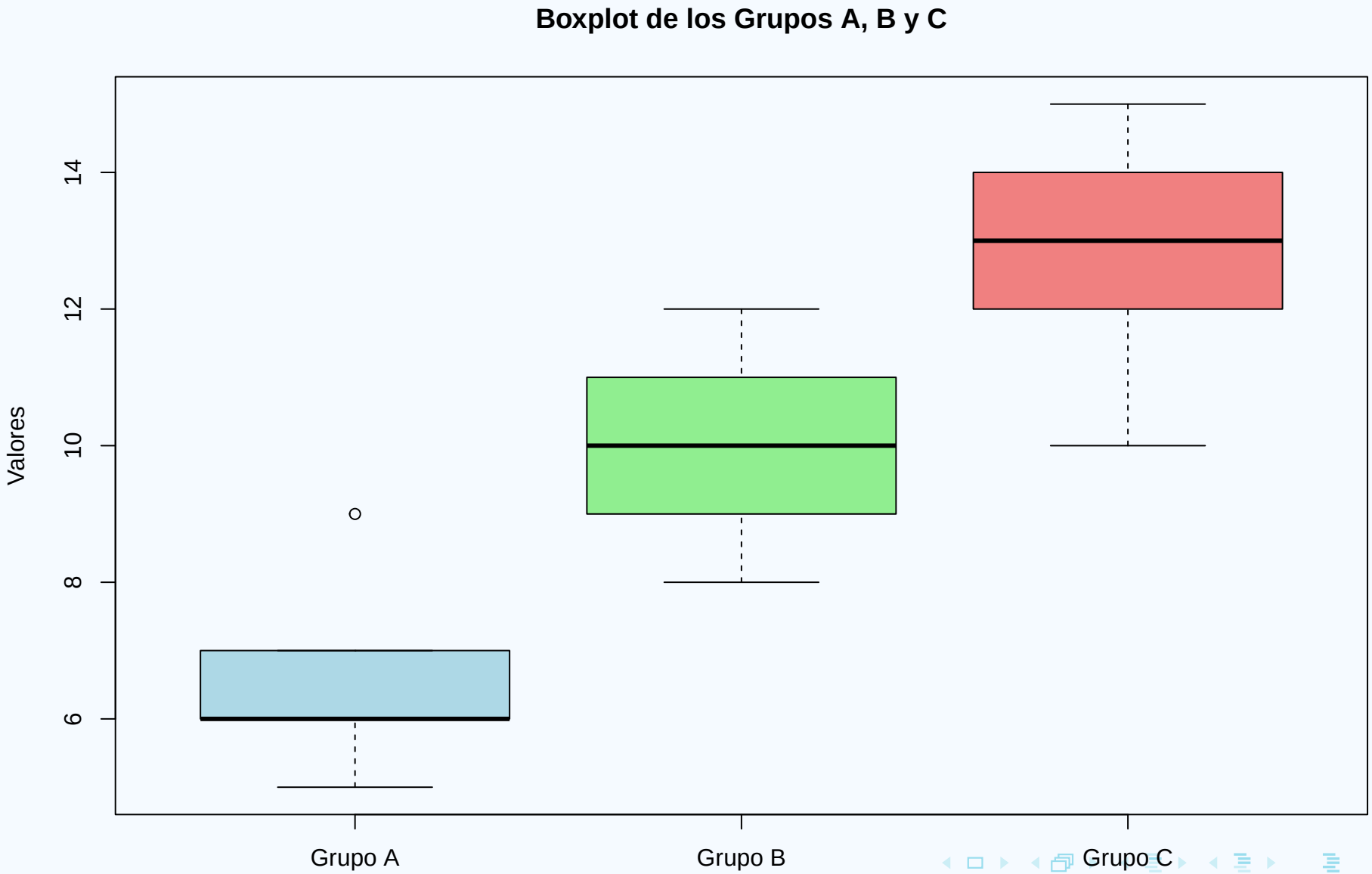
Supongamos que tenemos tres tratamientos (A, B y C) y sus correspondientes muestras de datos son:

- Grupo A: $[5, 7, 6, 9, 6]$
- Grupo B: $[8, 12, 9, 11, 10]$
- Grupo C: $[14, 10, 13, 15, 12]$

Nuestro objetivo es determinar si existe una diferencia significativa entre las medias de estos tres grupos.

Modelo con un factor: Ejemplo

Hagamos un boxplot con los tres tratamientos:



Modelo con un factor: Ejemplo

Comenzamos calculando la media de cada grupo y la media general:

- Media de Grupo A $\bar{Y}_A = \frac{5+7+6+9+6}{5} = \frac{33}{5} = 6.6$
- Media de Grupo B $\bar{Y}_B = \frac{8+12+9+11+10}{5} = \frac{50}{5} = 10$
- Media de Grupo C $\bar{Y}_C = \frac{14+10+13+15+12}{5} = \frac{64}{5} = 12.8$
- Media General $\bar{Y}$:

$$\bar{Y} = \frac{6.6 + 10.0 + 12.8}{3} = \frac{29.4}{3} = 9.8$$

Modelo con un factor: Ejemplo

Calculemos ahora cada uno de los componentes de SST = SSB + SSW

Comencemos por SSB:

$$\text{SSB} = n_A(\bar{Y}_A - \bar{Y})^2 + n_B(\bar{Y}_B - \bar{Y})^2 + n_C(\bar{Y}_C - \bar{Y})^2$$

siendo $n_A = n_B = n_C = 5$ (número de observaciones en cada grupo).

$$\text{SSB} = 5(6.6 - 9.8)^2 + 5(10.0 - 9.8)^2 + 5(12.8 - 9.8)^2 = 96.4$$

Modelo con un factor: Ejemplo

Calculemos a continuación la suma de los cuadrados dentro de los grupos:

$$\text{SSW} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

Para cada grupo, calculamos la suma de las diferencias al cuadrado entre cada dato y la media del grupo:

- Grupo A:
  $(5-6.6)^2 + (7-6.6)^2 + (6-6.6)^2 + (9-6.6)^2 + (6-6.6)^2 = 9.2$
- Grupo B:
  $(8-10)^2 + (12-10)^2 + (9-10)^2 + (11-10)^2 + (10-10)^2 = 10$
- Grupo C:
  $(14-12.8)^2 + (10-12.8)^2 + (13-12.8)^2 + (15-12.8)^2 + (12-12.8)^2 = 14.8$

Con ello,

$$\text{SSW} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2 = 9.2 + 10.0 + 14.8 = 34.0$$

Por tanto que la Suma Total de Cuadrados (SST) es:

$$\text{SST} = \text{SSB} + \text{SSW} = 96.4 + 34.0 = 130.4$$

Modelo con un factor: Ejemplo

Para realizar el contraste, calculamos el estadístico F

$$F = \frac{\text{MSB}}{\text{MSW}} = \frac{\text{SSB}/df_B}{\text{SSW}/df_W}$$

Los grados de libertad son:

- Grados de libertad entre los grupos: $df_B = k - 1 = 3 - 1 = 2$
- Grados de libertad dentro de los grupos: $df_W = N - k = 15 - 3 = 12$

Luego, el valor del estadístico del contraste es

$$F = \frac{\text{MSB}}{\text{MSW}} = \frac{\text{SSB}/df_B}{\text{SSW}/df_W} = \frac{96.4/2}{34/12} = \frac{48.2}{2.83} = 17.01$$

El p-valor se obtiene utilizando la distribución F-Snedecor con $df_B = 2$ y $df_W = 12$.

El comando en R es `pf(17.01, df1 = 2, df2 = 12, lower.tail = FALSE)` que devuelve un p-valor de 0.00031. Como es menor que el grado de significancia $\alpha = 0.05$, indica una diferencia significativa entre los grupos.

Modelo con un factor: Ejemplo

```r
grupo_a <- c(5, 7, 6, 9, 6)
grupo_b <- c(8, 12, 9, 11, 10)
grupo_c <- c(14, 10, 13, 15, 12)

datos <- data.frame(
  valores = c(grupo_a, grupo_b, grupo_c),
  grupo = factor(rep(c("Grupo A", "Grupo B", "Grupo C"), each

anova = aov(datos$valores ~ datos$grupo)
summary(anova)
```

```
             Df Sum Sq Mean Sq F value   Pr(>F)
datos$grupo   2   96.4   48.20   17.01 0.000314 ***
Residuals    12   34.0    2.83
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Modelo ANOVA con dos factores

- Se utiliza cuando se estudian dos factores simultáneamente para evaluar su efecto individual y conjunto en una variable dependiente
- Puede verse como una generalización del caso de ANOVA con un único factor
- Este modelo es más complejo y permite entender no solo los efectos principales de cada factor, sino también si hay una interacción entre ellos

Modelo ANOVA con dos factores

Sean A y B dos factores que se desean estudiar, con m_A y m_B niveles. Trabajaremos con las siguientes hipótesis nulas:

Opciones de hipótesis:

- Hipótesis nula para los efectos principales H_0:
  - No hay efecto del primer factor
  - No hay efecto del segundo factor
- Hipótesis nula para la interacción H_0: No hay interacción entre los dos factores

Modelo con dos factores sin interacción

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

Donde:

- Y_{ijk} es la observación k-ésima del nivel j-ésimo del factor B y nivel i-ésimo del factor A
- μ es la media general
- α_i es el efecto del nivel i-ésimo del factor A
- β_j es el efecto del nivel j-ésimo del factor B
- ϵ_{ijk} es el término de error aleatorio, $\epsilon_{ijk} \sim N(0, \sigma^2)$

Modelo con dos factores sin interacción

En este caso, la tabla ANOVA queda como sigue:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado Medio
Diferencias entre niveles del factor A	SSB_A	$m_A - 1$	MSB_A
Diferencias entre niveles del factor B	SSB_B	$m_B - 1$	MSB_B
Error	SSW	$N - m_A - m_B + 1$	MSW
Total	SST	$N - 1$	

Modelo con dos factores sin interacción

Para estudiar la importancia de cada factor se calcula el estadístico F particular para cada uno de ellos como sigue:

$$F_A = \frac{MSB_A}{MSW} \sim F_{m_A-1, N-m_A-m_B+1}$$

y

$$F_B = \frac{MSB_B}{MSW} \sim F_{m_B-1, N-m_A-m_B+1}$$

A partir de estos estadísticos de prueba podemos contrastar las hipótesis nulas de no existencia de efectos asociados a los factores A y B respectivamente.

Modelo con dos factores con interacción

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

Donde $(\alpha\beta)_{ij}$ representa el efecto de interacción entre el nivel i-ésimo del factor A y el nivel j-ésimo del factor B

Modelo con dos factores con interacción

En este caso, la tabla ANOVA añade el factor de interacción:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado Medio
Diferencias entre niveles del factor A	SSB_A	$m_A - 1$	MSB_A
Diferencias entre niveles del factor B	SSB_B	$m_B - 1$	MSB_B
Diferencias debidas la interacción	SSB_{AB}	$(m_A - 1) * (m_B - 1)$	MSB_{AB}
Error	SSW	$N - m_A * m_B$	MSW
Total	SST	$N - 1$	

Modelo con dos factores con interacción

Para estudiar la importancia de cada de la interacción se calculan el estadístico F correspondiente:

$$F_{AB} = \frac{MSB_{AB}}{MSW} \sim F_{(m_A-1)*(m_B-1),N-m_A*m_B}$$

Con este estadístico, contrastamos la hipótesis nula de no existencia de interacción entre los dos factores A, B:

- Si podemos rechazar esa hipótesis, es decir, si existe interacción entre los factores, entonces hemos terminado. Es decir, no podemos eliminar ningún factor del modelo.
- En cambio, si no rechazamos la hipótesis nula, es decir, si no existe interacción entre los factores, podemos eliminar dicho efecto (la interacción) del modelo y pasar a un modelo sin interacción (como el previo)

Modelo con dos factores con interacción

Se ha realizado un estudio para evaluar el efecto de dos factores sobre la calidad del sueño:

- **Factor A**: Tipo de rutina antes de dormir (con tres niveles: "Leer", "Meditar" y "Uso de pantallas").
- **Factor B**: Ambiente de sueño (con dos niveles: "Silencio" y "Con ruido").

Cada combinación de los factores tiene dos observaciones para medir la calidad del sueño (con puntuaciones del 1 al 10, siendo 10 la mejor calidad). Los datos son los siguientes:

	Silencio	Con ruido
Leer	7, 8	5, 6
Meditar	9, 10	6, 7
Uso de pantallas	4, 5	3, 4

Comparaciones múltiples

- Con ANOVA podemos rechazar la siguiente hipótesis nula: $H_0 : \mu_1 = \mu_2 \ldots = \mu_k$, siendo k el número de niveles en el factor
- ¿En qué niveles del factor se encuentran las principales diferencias? Es decir, qué hipotesis (una o varias) de las siguientes son rechazadas:

$$H_0 : \mu_1 = \mu_2$$

$$H_0 : \mu_1 = \mu_3$$

$$\ldots$$

$$H_0 : \mu_{k-1} = \mu_k$$

Comparaciones múltiples

- En un caso con k niveles en el factor, hay $k * (k-1)/2$ posibles contrastes de igualda de medias
- Si realizamos todos esos contrastes, aumenta la probabilidad de cometer errores de tipo I (rechazar incorrectamente la hipótesis nula)
- Este fenómeno se conoce como el problema de las **comparaciones múltiples**

Comparaciones múltiples

- Cuando realizamos una sola prueba de hipótesis, establecemos un nivel de significancia predeterminado (ejemplo $\alpha = 0.05$)
- Esto es, aceptamos una probabilidad de error de tipo I del 5%, es decir, hay un 5% de probabilidad de rechazar incorrectamente la hipótesis nula cuando es verdadera
- Cuando se realizan múltiples test de hipótesis, la probabilidad acumulada de cometer al menos un error de tipo I aumenta con cada prueba adicional
  - Ejemplo: Con 10 tests de hipótesis independientes, cada uno con un nivel de significancia de $\alpha = 0.05$, la probabilidad de cometer al menos un error de tipo I aumenta a más del 40% $(1 - (1 - 0.05)^{10} \approx 0.40)$

Comparaciones múltiples

- Solución: Aplicar correcciones cuando se realizan comparaciones múltiples para controlar este aumento en el riesgo de error
- Existen varios métodos para controlar el problema de las pruebas múltiples:
  - Bonferroni, Holm-Bonferroni, LSD (Least Significant Differences), Tuley HSD (Honestly-significant-difference), entre otros
- Estos métodos controlan la tasa global de error de tipo I para todas las comparaciones realizadas, manteniendo un nivel de significancia general específico

Método de Bonferroni

- Este método es relativamente simple y conservador
- Idea: ajustar el nivel de significancia individual para cada prueba de hipótesis realizada. En lugar de utilizar un nivel de significancia estándar (por ejemplo, $\alpha = 0.05$), se divide el nivel de significancia global deseado por el número total de pruebas realizadas (m):

$$\alpha' = \frac{\alpha}{m}$$

  Esta división produce un nivel de significancia más estricto (α') para cada prueba individual, lo que ayuda a controlar el riesgo global de error de tipo I
- Se utiliza el nivel de significancia individual ajustado para cada prueba de hipótesis. Si el $p - valor$ de una prueba es menor que el nivel de significancia ajustado, se rechaza la hipótesis nula de la prueba

Método de Bonferroni

- Fácil de entender e implementar
- Proporciona un control conservador sobre el error de tipo I en comparaciones múltiples
- Puede ser un método demasiado conservador en situaciones donde se realizan muchas comparaciones, lo que puede resultar en una pérdida de potencia estadística

Referencias

Gomez Villegas, M. A. (2005). *Inferencia estadística*. Ediciones Díaz de Santos.

Spiegel, M., & Stephens, L. (2009). Estadística–Serie Schaum. *Mc Graw-Hill*.

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

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