MÓDULO

FUNDAMENTOS DE ESTADÍSTICA

ANNA RENAU

Graduada en Matemáticas (Universitat Autònoma de Barcelona).
Máster en *Modelling in Science and Engineering*, especialidad en *Statistics* (Universitat Autònoma de Barcelona).
Data Scientist en SCRM-Lidl.

Institut de Formació Contínua-IL3
UNIVERSITAT DE BARCELONA

ÍNDICE

OBJETIVOS

- Ofrecer nociones básicas de estadística que permitan, dadas unas observaciones, analizar por encima los datos y extraer información relevante.
- Introducir el concepto de probabilidad como elemento fundamental de la inferencia estadística.
- Comentar ejemplos de cálculos de probabilidad para ver su aplicación en el día a día.
- Explicar cómo tratar cada variable, cómo extraer información de ella, y cómo sacar datos que nos puedan aportar dos o más variables conjuntamente.
- Determinar las diferencias entre variables categóricas y numéricas, y aprender a aplicar los principales estimadores univariables: recuento, media, mediana y desviación estándar.
- Tratar las relaciones entre dos variables, enfocar cómo representarlas y trabajar con los principales métodos de análisis bivariables: tablas de contingencia, correlación y regresión.
- Aprender a trabajar con los principales métodos de análisis multivariables: árboles de decisión, agrupación (*clustering*), factorial y regresión.

1. PRINCIPIOS BÁSICOS DE PROBABILIDAD

1.1. INTRODUCCIÓN

IMPORTANTE

Una variable aleatoria es una función que asocia a cada resultado de un experimento un número real.

Existen dos tipos de variables aleatorias:

- **Discretas**: *X* es una variable aleatoria discreta si toma valores en un conjunto numerable $x_1,...,x_n$. Como tiene un número finito de puntos, se representa gráficamente mediante un **gráfico de barras**.
- **Continuas**: *X* es una variable aleatoria continua si toma valores en un conjunto infinito no numerable. Como tiene un número infinito de puntos, se representa gráficamente mediante una **línea continua**.

Las variables aleatorias siguen una **distribución de probabilidad**. Esta depende de si la variable es discreta o continua y se definen mediante funciones (función de probabilidad, función de densidad y función de distribución).

Aparte de la distribución, las variables aleatorias tienen medidas que ayudan a entender su comportamiento (**esperanza y varianza**).

FUNCIÓN DE PROBABILIDAD

Sea *X* una variable aleatoria **discreta** con valores $x_1,...,x_n$. Se define la función de probabilidad de *X*, *P(X)*, como la que asocia una probabilidad $p_1,...,p_n$ a cada valor posible de *X*.

$$P(X = x_i) = p_i$$

En una función de probabilidad se cumplen las dos condiciones siguientes:

$$0 \leq p_i \leq 1$$

$$\sum_{i=1}^{n} p_i = 1$$

FUNCIÓN DE DENSIDAD

Sea *X* una variable aleatoria **continua**. Se define la función de densidad de *X*, *f(X)*, como la probabilidad relativa según la cual *X* tomará un determinado valor.

$$P(a \leq x \leq b) = \int_a^b f(x)\, dx$$

Fíjate que miramos la probabilidad en un **intervalo**. En una variable aleatoria continua no tiene sentido calcular la probabilidad en un punto concreto, ya que $P(X = x) = 0$.

En una función de densidad se cumplen las dos condiciones siguientes:

$$f(x) \geq 0$$

$$\int_{-\infty}^{\infty} f(x)\, dx = 1$$

FUNCIÓN DE DISTRIBUCIÓN

Sea *X* una variable aleatoria. Se define la función de distribución de *X*, *F(X)*, como la que proporciona, en cada punto, la probabilidad acumulada hasta dicho valor:

$$F(x) = p(X \leq x)$$

Si *X* es una variable aleatoria **discreta**, se cumple que:

$$F(x) = \sum_{i=1}^{x} P(X = x_i)$$

Si *X* es una variable aleatoria **continua**, se cumple que:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

En una función de distribución se cumplen las dos condiciones siguientes:

$$\lim_{x \to -\infty} F(x) = 0$$

$$\lim_{x \to +\infty} F(x) = 1$$

ESPERANZA

La esperanza de una variable aleatoria es el valor que se espera que esta vaya a tomar.

En una variable aleatoria **discreta** se define como:

$$E(x) = \sum_{i=1}^{n} x_i \, P(X = x_i)$$

En una variable aleatoria **continua** se define como:

$$E(x) = \int_{-\infty}^{\infty} x \, f(x) \, dx$$

VARIANZA

Se define la varianza de una variable aleatoria *X* como:

$$Var(X) = E(X^2) - E(X)^2$$

Se cumple siempre que $Var(X) \geq 0$.

1.2. PRINCIPALES FUNCIONES DE PROBABILIDAD: VARIABLES ALEATORIAS DISCRETAS

1.2.1. DISTRIBUCIÓN UNIFORME DISCRETA

IMPORTANTE

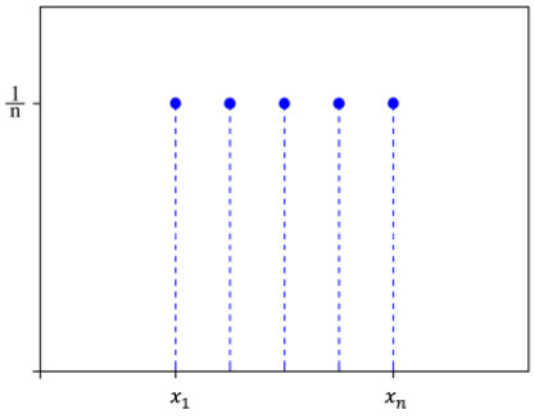
Sea *X* una variable aleatoria discreta que toma valores $x_1,...,x_n$, diremos que esta tiene una distribución uniforme discreta cuando puede tomar cualquiera de los *n* valores con la misma probabilidad.

Es la distribución discreta más sencilla que existe:

$$X \sim U(n)$$

Su **función de probabilidad** es:

$$P(X = x_i) = \frac{1}{n}$$



Ejemplo de distribución uniforme discreta.

Su **esperanza** es:

$$E(x) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Y su **varianza**:

$$Var(X) = \frac{1}{n} \sum_{i=1}^{n} (x_i)^2 - E(x)^2$$

EJEMPLO

Si tiramos un dado, ¿cuál es la probabilidad de que nos salga un 5?

Un dado puede tomar los valores 1, 2, 3, 4, 5 o 6 con la misma probabilidad, por lo tanto, tiene una distribución uniforme discreta con $n = 6$.

$$X \sim U(6)$$

La probabilidad de obtener un 5 es la siguiente:

$$P(X = 5) = \frac{1}{6}$$

1.2.2. DISTRIBUCIÓN DE BERNOUILLI

IMPORTANTE

Una variable aleatoria discreta X tiene una distribución de Bernouilli si solo puede tomar dos valores (éxito: $x_1 = 1$, o fracaso: $x_2 = 0$ con probabilidades $p \in [0,1]$ y $q = 1 - p$.

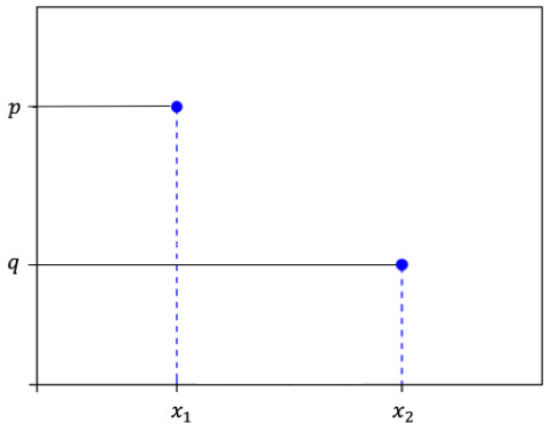
Siendo *p* la probabilidad de éxito y *q*, la probabilidad de fracaso:

$$X \sim B(p)$$

Su **función de probabilidad** es:

$$P(\text{éxito}) = P(X = x_1) = P(X = 1) = p$$

$$P(\text{fracaso}) = P(X = x_2) = P(X = 0) = 1 - p = q$$



Ejemplo de distribución de Bernouilli.

Su **esperanza** es:

$$E(X) = p$$

Y su **varianza**:

$$Var(X) = pq$$

EJEMPLO

El 40 % de los trabajadores de un país tiene estudios universitarios. ¿Cuál es la probabilidad de seleccionar un trabajador al azar y que este haya cursado estudios universitarios?

En este caso, tenemos una variable aleatoria discreta que solo puede tomar dos valores: éxito = tiene estudios universitarios, fracaso = no tiene estudios universitarios. Por lo tanto:

$$X \sim B(0,4)$$

La probabilidad de seleccionar a un trabajador al azar y que este tenga estudios universitarios es de:

$$P(\text{éxito}) = P(X = 1) = 0,4$$

1.2.3. DISTRIBUCIÓN BINOMIAL

Es una distribución de Bernouilli en la cual un experimento se repite *n* veces de forma independiente:

$$X \sim B(n,p)$$

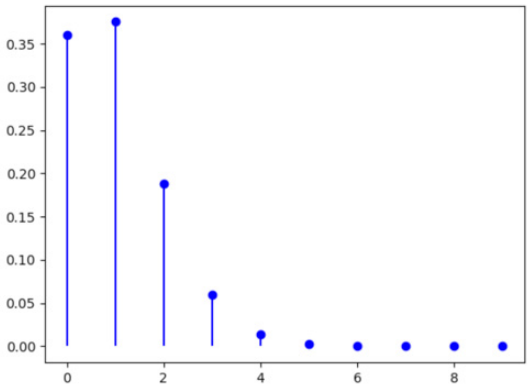
IMPORTANTE

Es necesario verificar siempre que $p \in [0,1]$ y que $n > 1$, si $n = 1$ tenemos una distribución de Bernouilli.

Su **función de probabilidad** es:

$$P(X = x_i) = \binom{n}{x_i} p^{x_i} (1-p)^{n-x_i} \quad x_i = \{1,\ldots,n\}$$

Donde $\binom{n}{x_i} = \frac{n!}{x_i!(n-x_i)!}$ son las combinaciones de n en x_i.



Ejemplo de distribución binomial.

Su **esperanza** es:

$$E(X) = np$$

Y su **varianza**:

$$Var(X) = np(1-p) = npq$$

Recordamos que $q = 1 - p$.

EJEMPLO

Lanzamos una moneda 20 veces. Si la probabilidad de que salga cara es de ½, ¿cuál es la probabilidad de que salga cara 14 veces?

En este caso tenemos una variable aleatoria discreta que solo puede tomar dos valores: éxito = cara, fracaso = cruz. Por lo tanto, una distribución Bernouilli. Como repetimos el experimento 20 veces, tenemos una distribución binomial:

$$X \sim B(20,½)$$

La probabilidad de que salga cara 14 veces es de:

$$P(X = 14) = \binom{20}{14} (½)^{14} (1 - ½)^{20-14} = 0,04$$

1.2.4. DISTRIBUCIÓN DE POISSON

IMPORTANTE

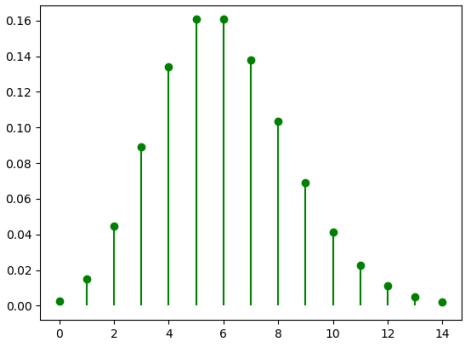
Una distribución aleatoria discreta *X* tiene una distribución de Poisson si expresa, a partir de una frecuencia de ocurrencia media, la probabilidad de que ocurra un determinado número de eventos durante cierto período de tiempo.

$$X \sim P(\lambda)$$

Donde *λ* representa el número promedio de ocurrencias en un intervalo de tiempo o en un espacio.

Su **función de probabilidad** es:

$$P(X = x_i) = e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}$$



Ejemplo de distribución de Poisson.

Su **esperanza** es:

$$E(X) = \lambda$$

Y su **varianza**:

$$Var(X) = \lambda$$

EJEMPLO

Una central telefónica recibe una media de 480 llamadas por hora. Si el número de llamadas se distribuye según una Poisson y la central tiene una capacidad para atender, a lo sumo, 11 llamadas por minuto, ¿cuál es la probabilidad de que en un minuto determinado no sea posible dar línea a todos los clientes porque se reciben 12 llamadas?

Es un ejemplo claro de una distribución de Poisson, a partir de una frecuencia de ocurrencia media (480 llamadas por hora), buscamos la probabilidad de que ocurra un determinado número de eventos durante cierto período de tiempo (12 llamadas por minuto).

Si definimos X = N.º de llamadas por minuto, entonces λ es el número promedio de llamadas por minuto. Teniendo en cuenta que la central telefónica recibe una media de 480 llamadas por hora y que una hora tiene 60 minutos, $\lambda = 480/60 = 8$. Así, la distribución de X es:

$$X \sim P(8)$$

Usando su función de probabilidad, la probabilidad de que en un minuto determinado se realicen 12 llamadas es de:

$$P(X = 12) = e^{-8}\frac{8^{12}}{12!} = 0{,}048$$

1.3. PRINCIPALES FUNCIONES DE PROBABILIDAD: VARIABLES ALEATORIAS CONTINUAS

1.3.1. DISTRIBUCIÓN UNIFORME CONTINUA

IMPORTANTE

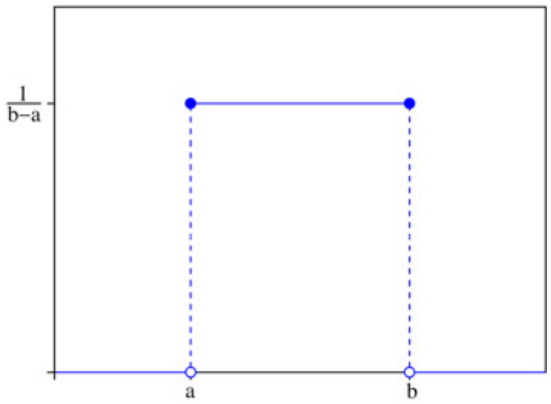
Sea X una variable aleatoria continua, diremos que esta tiene una distribución uniforme continua si toma valores $x_i \in [a,b]$ de manera equiprobable, donde $-\infty < a < b < \infty$.

Es la distribución continua más sencilla que existe.

$$X \sim U(a,b)$$

Su **función de densidad** es:

$$f(x) = \begin{cases} 0 & x < a \\ \dfrac{1}{b-a} & a \le x \le b \\ 0 & x > b \end{cases}$$



Ejemplo de distribución uniforme continua.

Su **esperanza** es:

$$E(X) = \frac{a+b}{2}$$

Y su **varianza**:

$$Var(X) = \frac{(b-a)^2}{12}$$

EJEMPLO

Sea *X* un número real seleccionado al azar en el intervalo [1, 10], ¿cuál es la probabilidad de que este número sea menor o igual a 8?

Como *X* puede tomar cualquier número en el intervalo [1, 10] con la misma probabilidad, se trata de un claro caso de distribución uniforme continua:

$$X \sim U(1,10)$$

Queremos saber la probabilidad de que el número elegido sea menor de 8, es decir: *P(X ≤ 8) = F(8)*.

Recordamos que en una variable aleatoria continua:

$$F(x) = \int_{-\infty}^{x} f(t)\,dt$$

Con lo cual:

$$F(x) = \begin{cases} 0 & x < a \\ \dfrac{x-a}{b-a} & a \le x \le b \\ 1 & x > b \end{cases}$$

Según la función de distribución, la probabilidad buscada es de:

$$P(X \le 8) = F(8) = \frac{8-1}{10-1} = 0{,}78$$

1.3.2. DISTRIBUCIÓN NORMAL

Es la distribución más común, ya que en la vida real podemos encontrar multitud de fenómenos que se comportan según esta distribución.

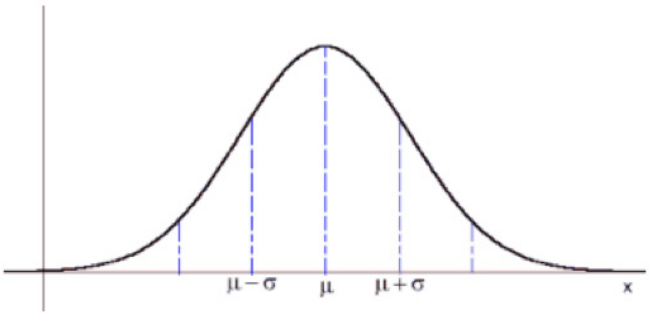
IMPORTANTE

La distribución normal se caracteriza porque los valores se distribuyen formando una campana de Gauss en torno al valor medio de la distribución $\mu \in R$, con una dispersión $\sigma > 0$.

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

Su **función de densidad** es:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} = e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Ejemplo de distribución normal.

Su **esperanza** es:

$$E(X) = \mu$$

Y su **varianza**:

$$Var(X) = \sigma^2$$

Diremos que la distribución normal es **tipificada o estandarizada** cuando $\mu = 0$ y $\sigma = 1$. En este caso:

Su **función de densidad** es:

$$f(x) = \frac{1}{\sqrt{2\pi}} = e^{-\frac{x^2}{2}}$$

Su **esperanza** es:

$$E(X) = 0$$

Y su **varianza**:

$$Var(X) = 1$$

Dada la función de densidad de una variable con distribución normal, como puedes suponer no es muy práctico calcular la función de distribución. Por este motivo, existen tablas donde ya viene calculada $P(X \leq x)$. Esto ocurre con muchas distribuciones continuas, no solo con la distribución normal.

En el caso de la distribución normal, como no se puede tener una tabla para cada μ y σ, se usa la tabla con distribución normal tipificada (encontrarás una en el anexo).

IMPORTANTE

Lo que se debe hacer cuando tengamos una distribución normal X es convertirla en una distribución normal tipificada Z, y luego usar la tabla para encontrar la probabilidad en cuestión.

Pasar de una a la otra es muy sencillo, si partimos de $X \sim N(\mu,\sigma)$, para encontrar Z solo tenemos que restarle μ y dividir por σ.

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

EJEMPLO

Las notas de los alumnos de una clase se distribuyen normalmente con una media de 6 y dispersión de 3: ¿cuál es la probabilidad de que un alumno haya suspendido? (Entendemos que se suspende si la nota es inferior a 5).

Partimos de una distribución normal:

$$X \sim N(6,3)$$

Y queremos encontrar $P(X \leq 5)$. Si lo pasamos a una distribución normal tipificada, tenemos que:

$$P(X \leq 5) = P\left(Z \leq \frac{5 - 6}{3}\right) = P(Z \leq -0{,}33)$$

Ahora solo se debe buscar el valor en la tabla. Para ello, buscamos en los índices de fila el primer decimal (0,3) y en las columnas, el segundo (0,03).

z	0,00	0,01	0,02	0,03
0,0	0,5000	0,5040	0,5080	0,5120
0,1	0,5398	0,5438	0,5478	0,5517
0,2	0,5793	0,5832	0,5871	0,5910
0,3	0,6179	0,6217	0,6255	0,6293

Ejemplo de uso de la tabla de distribución normal tipificada.

Esto nos da el valor buscado: 0,6293. Como $P(Z \leq 0{,}33) = 1 - P(Z \leq -0{,}33)$, tenemos que:

$$P(Z \leq -0{,}33) = 1 - 0{,}6293 = 0{,}37$$

1.3.3. DISTRIBUCIÓN EXPONENCIAL

IMPORTANTE

Diremos que X tiene una distribución exponencial si estudia el tiempo que pasa entre dos sucesos consecutivos de una distribución de Poisson.

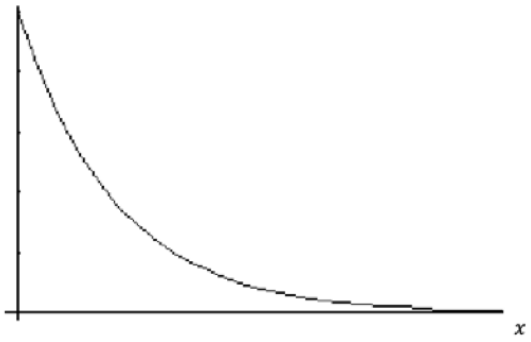
Mientras que la distribución de Poisson describe las llegadas por unidad de tiempo, la distribución exponencial estudia el **tiempo** entre cada una de estas llegadas. La distribución de Poisson es discreta, pero la distribución exponencial es **continua** porque el tiempo entre llegadas no tiene que ser un número entero:

$$X \sim Exp(\lambda)$$

Donde λ representa el número promedio de ocurrencias de la distribución de Poisson.

Su **función de densidad** es:

$$f(x) = \begin{cases} 0 & x \leq 0 \\ \lambda e^{-\lambda x} & x > 0 \end{cases}$$



Ejemplo de distribución exponencial.

Su **esperanza** es:

$$E(X) = \frac{1}{\lambda}$$

Y su **varianza**:

$$Var(X) = \frac{1}{\lambda^2}$$

EJEMPLO

El tiempo durante el cual cierta marca de batería trabaja en forma efectiva hasta que falla (tiempo de falla) se distribuye según el modelo exponencial con un tiempo promedio de fallas igual a 360 días. ¿Qué probabilidad hay de que el tiempo de falla sea mayor a 400 días?

Sabemos que *X* = tiempo que trabaja la batería hasta que falle tiene una distribución exponencial, y que el tiempo promedio de falla es de 360 días. Con lo cual:

$$X \sim Exp\left(\frac{1}{360}\right)$$

Queremos encontrar $P(X > 400) = 1 - P(X \leq 400) = F(400)$.

Recordamos que en una variable aleatoria continua:

$$F(x) = \int_{-\infty}^{x} f(t)\, dt$$

Con lo cual:

$$F(x) = \begin{cases} 0 & x \leq 0 \\ 1 - e^{-\lambda x} & x > 0 \end{cases}$$

Según la función de distribución:

$$P(X > 400) = 1 - P(X \leq 400) = 1 - F(400) = 1 - (1 - e^{\frac{400}{360}}) = 0{,}33$$

1.4. APROXIMACIÓN DE FUNCIONES DE PROBABILIDAD

Hay distribuciones que, bajo ciertas condiciones, son muy parecidas entre ellas; diremos entonces que una distribución se aproxima a otra.

SABÍAS QUE...

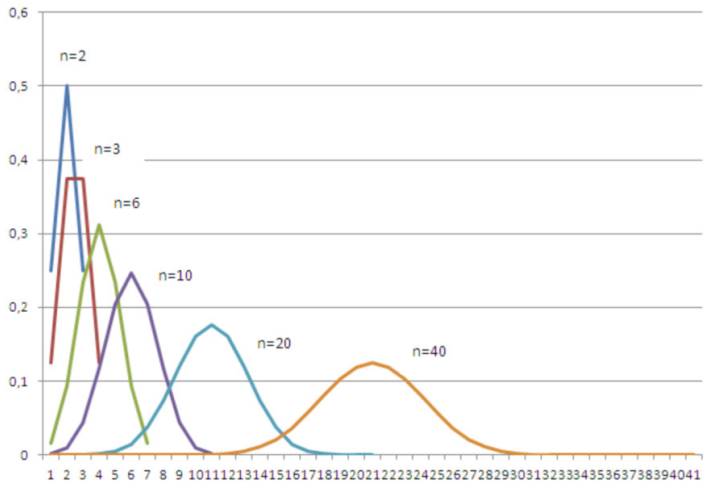
Cuando se cumplen las condiciones para que una distribución se aproxime a otra, se pueden hacer los cálculos de probabilidad con cualquiera de las dos.

1.4.1. APROXIMACIÓN DE LA DISTRIBUCIÓN BINOMIAL A LA DISTRIBUCIÓN NORMAL

Una distribución binomial con $p > 0{,}1$ y $q > 0{,}1$ tiende a una distribución normal cuando el número de experimentos tiende a infinito. Como en la vida real no tenemos infinitos experimentos, consideraremos que una binomial se puede aproximar a una normal cuando $n \geq 30$.

Vamos a verlo. Recordamos que la distribución binomial es discreta y, por lo tanto, se representa gráficamente mediante barras. Sin embargo, como en este caso queremos aproximarla a una distribución normal que es continua, lo que haremos será **representarla mediante puntos** y unir los puntos para simular una variable continua.

- Primero tenemos que seleccionar una p, $q > 0{,}1$, vamos a usar $p = 0{,}5$ (y en consecuencia $q = 1 - p = 0{,}5$).
- Luego el número de experimentos n. Veremos varios casos de $n < 30$, así se entenderá por qué se debe superar este número de experimentos. Vamos a ver $n = 2, 3, 6, 10, 20, 40$.



Distribución binomial con p = *0,5 y* n = *2, 3, 6, 10, 20 y 40.*

En este gráfico podemos observar claramente que, cuando $n \geq 30$ (en este caso $n = 40$), la distribución es igual a la de una normal. En este caso, ya a partir de $n = 10$ se puede comprobar la similitud entre ambas distribuciones, pero eso no pasa siempre, depende de la p elegida.

En conclusión, si se cumplen las condiciones:

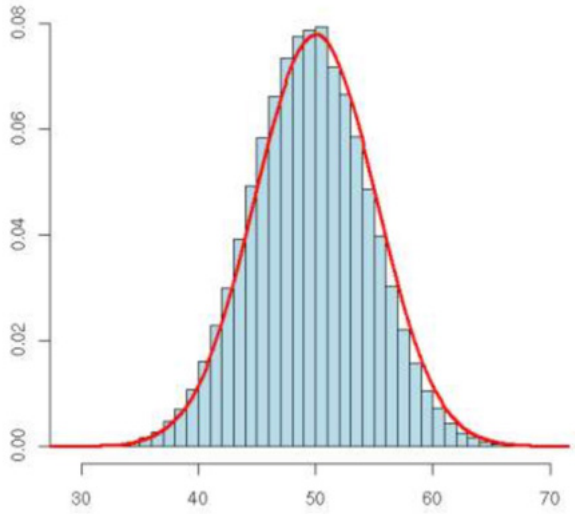
- $p, q > 0{,}1$.
- $n \geq 30$.

La distribución binomial puede aproximarse por una distribución normal de parámetros:

$$B(n,p) \cong N(np, \sqrt{npq})$$

Cuanto mayor sea n, mejor será la aproximación. Así que, si tenemos una distribución binomial donde se cumplen estas dos condiciones, podemos hacer los cálculos mediante la distribución normal, ya que son menos tediosos.

Veamos un último ejemplo de aproximación con $p = 0{,}5$ y $n = 100$. Si representamos gráficamente las distribuciones $B(100, 0{,}5)$ y $N(100 \cdot 0{,}5, \sqrt{100 \cdot 0{,}5 \cdot 0{,}5})$ vemos que son prácticamente iguales.

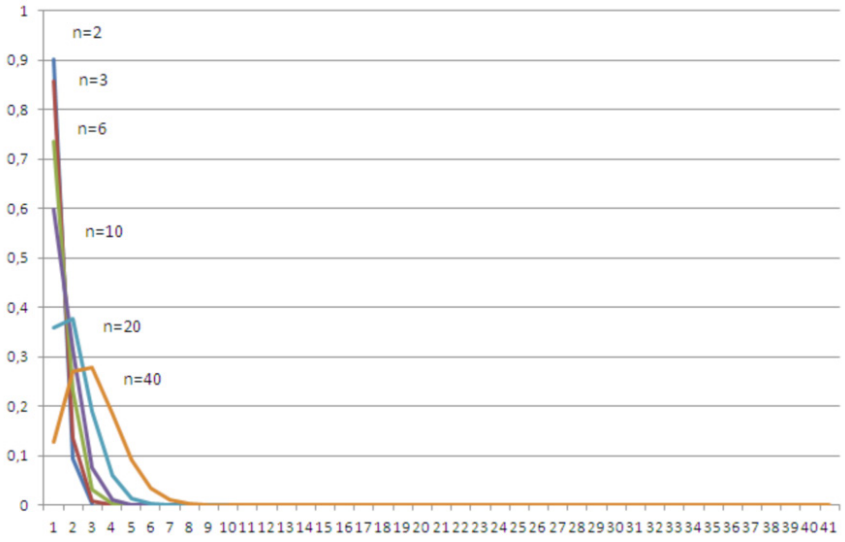


Distribución binomial vs. distribución normal.

1.4.2. APROXIMACIÓN DE LA DISTRIBUCIÓN BINOMIAL A LA DISTRIBUCIÓN DE POISSON

Una distribución binomial con $p \leq 0{,}1$ o $q \leq 0{,}1$ tiende a una distribución de Poisson cuando el número de experimentos tiende a infinito. Igual que en el caso anterior, consideraremos que esto pasa cuando $n \geq 30$.

Veamos un ejemplo. Consideramos $p = 0{,}05$ y el mismo número de experimentos que en la situación anterior $n = 2, 3, 6, 10, 20, 40$.



Distribución binomial con p = *0,05 y* n = *2, 3, 6, 10, 20 y 40.*

La similitud con una distribución de Poisson es clara; por $n = 40$ las dos distribuciones son **prácticamente equivalentes**.

Cuanto menor sea p (o q) y mayor sea n, mejor será la aproximación.

En conclusión, si se cumplen las condiciones:

- $p \leq 0{,}1$.
- $n \geq 30$.

La distribución binomial puede aproximarse mediante una distribución de Poisson de parámetros:

$$B(n,p) \cong P(np)$$

Si se cumplen las condiciones:

- $q \leq 0{,}1$.
- $n \geq 30$.

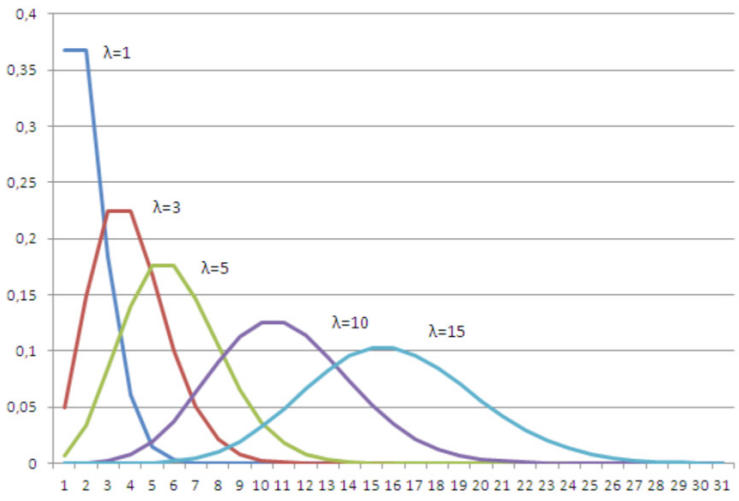
La distribución binomial puede aproximarse por una distribución de Poisson de parámetros:

$$B(n,p) \cong P(nq)$$

1.4.3. APROXIMACIÓN DE LA DISTRIBUCIÓN DE POISSON A LA DISTRIBUCIÓN NORMAL

Una distribución de Poisson con $\lambda \geq 10$ tiende a una distribución normal.

Veamos un ejemplo con diferentes λ.



Distribución de Poisson con λ = 1, 3, 5, 10 y 15.

En el gráfico se ve claramente que con $\lambda \geq 10$ la distribución de Poisson se aproxima a una normal.

En conclusión, si se cumple la condición:

- $\lambda \geq 10$.

La distribución de Poisson puede aproximarse por una distribución normal de parámetros:

$$P(\lambda) \cong N(\lambda, \sqrt{\lambda})$$

2. ESTIMADORES UNIVARIABLES

Cuando hablamos de *data* nos referimos a números o categorías con un significado. Siempre tiene que haber un contexto, sino los datos no tienen sentido.

Por ejemplo, considera los datos: 30, 23, 91, 72. Estos datos pueden significar muchas cosas: edades, velocidades tomadas por un radar, etc. Sin un contexto, los datos no tienen sentido.

Agrupamos los datos en variables. Una variable puede ser la edad de una persona, las calificaciones de un examen, el color de un coche, etc.

Hay **dos tipos de variables**:

- **Categóricas (cualitativas)**: los valores de una variable categórica son categorías o grupos mutuamente excluyentes. Los datos categóricos pueden tener o no tener un orden lógico. Por ejemplo: sexo, país de residencia, etc.
- **Cuantitativas**: los valores de una variable cuantitativa son números que suelen representar una medición. Por ejemplo: edad, altura, peso, etc.

2.1. VARIABLES CATEGÓRICAS

Lo primero que haremos cuando queramos analizar una variable categórica es contar el número de observaciones por categoría, eso se denomina ***tabla de frecuencias***. También se puede representar como porcentaje respecto al total: **tabla de frecuencias relativas**.

EJEMPLO

Imaginemos que la policía local nos proporciona datos de todos los robos que ha habido en el último año. Supongamos que tienen clasificados a los ladrones según las siguientes categorías: ladrones de coches, de carteras, de tiendas, de bancos y de casas.

CATEGORÍA DE LADRÓN	FRECUENCIA	FRECUENCIA RELATIVA
COCHES	78	0,03
CARTERAS	2.341	0,78
TIENDAS	332	0,11
BANCOS	4	0,00
CASAS	241	0,08
TOTAL	2.996	1

Tabla de frecuencias y frecuencias relativas-categorías de ladrones.

IMPORTANTE

Las frecuencias relativas son conocidas como la distribución de la variable. En el caso de una variable categórica, esto nos dice cuál es la probabilidad de que una futura observación de la variable pertenezca a cada categoría.

En nuestro ejemplo, cuando en un futuro se denuncie un robo nos informa con qué probabilidad este será de un coche, de una cartera, etc.

2.1.1. REPRESENTACIÓN GRÁFICA

Las representaciones gráficas más simples para representar una variable categórica son los **gráficos de barras** y los **gráficos circulares**.

GRÁFICO DE BARRAS

En los gráficos de barras cada barra es una categoría, y la longitud de estas representa la frecuencia.

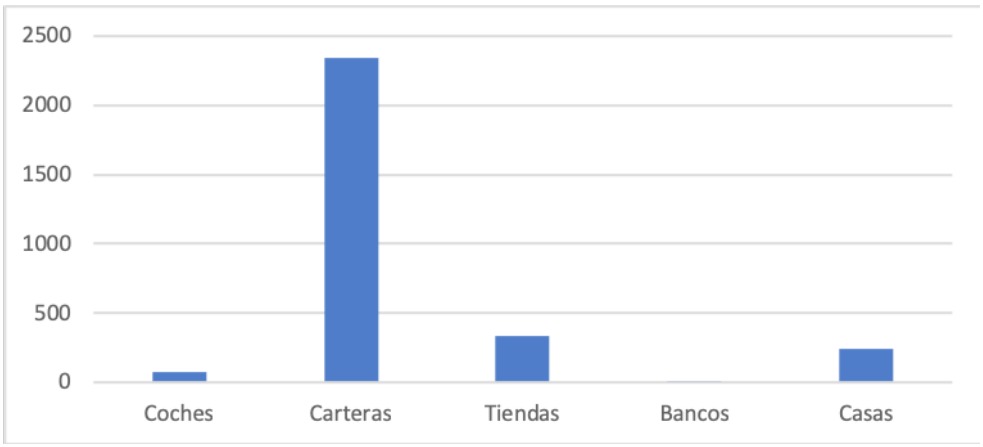


Gráfico de barras - categorías de ladrones.

GRÁFICO CIRCULAR

En los gráficos circulares cada sección es una categoría, y el área de estas es proporcional a la frecuencia relativa de esa categoría.

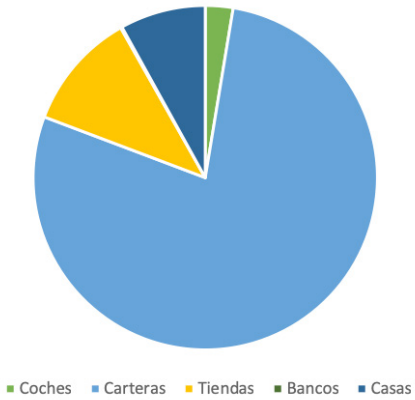


Gráfico circular - categorías de ladrones.

2.2. VARIABLES CUANTITATIVAS

Una variable cuantitativa es una variable numérica que puede tomar cualquier valor (normalmente entre un intervalo). Para analizar este tipo de variables, lo que haremos será **buscar su distribución**. Para ello, existen varios **estadísticos** que nos serán útiles:

- Media.
- Mediana.
- Moda.
- Percentiles y cuartiles.
- Rango intercuartil.
- Mínimo y máximo.
- Rango.
- Derivación estándar y variancia.

Antes de analizarlos con detalle, vamos a ver unos datos que usaremos de ejemplo para calcular los diferentes estadísticos.

Supongamos que queremos estudiar la temperatura de una ciudad. Para ello, durante un año, calculamos cada mes su temperatura media (en grados centígrados).

ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SET	OCT	NOV	DIC
1	2	6	13	19	21	29	30	17	10	5	1

Temperatura ciudad 1 (media por mes).

MEDIA

Sean $x_1, x_2, \ldots, x_n$ las n observaciones, calculamos la media $\overline{x}$ con la siguiente fórmula:

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

En nuestro ejemplo:

$$\overline{x} = \frac{1 + 2 + \ldots + 5 + 1}{12} = 12{,}8\ ^{o}C$$

MEDIANA

Tal y como su nombre indica, la mediana es **el valor del medio**. Si ordenamos los datos de mayor a menor, es el valor que hace que la mitad de los datos queden por debajo y la otra mitad por arriba.

El cálculo depende de si n es un número par o impar:

- n impar, una vez ordenados los datos de menor a mayor, la mediana es la observación que está en la posición $(n+1)/2$.
- n par, una vez ordenados los datos de menor a mayor, la mediana es la media entre las observaciones $n/2$ y $n/2 + 1$.

En nuestro ejemplo, primero ordenamos de menor a mayor:

1, 1, 2, 5, 6, 10, 13, 17, 19, 21, 29, 30

Como tenemos $n = 12$ observaciones, buscamos los valores en las posiciones $12/2 = 6$ y $12/2 + 1 = 7$. Esto corresponde a los valores 10 y 13.

La mediana es la media de estos dos, por lo tanto, 11,5 ºC.

MODA

La moda es el **valor** que ocurre **con más frecuencia**. En caso de haber dos valores que ocurren con la misma frecuencia, entonces habría dos modas.

En nuestro ejemplo, la moda sería 1 ºC, ya que es el único valor que ocurre más de una vez.

PERCENTILES Y CUARTILES

El percentil p de una distribución es el valor tal que el p % de las observaciones quedan por debajo de este valor. Fíjate que la mediana es el percentil 50, ya que el 50 % de las observaciones quedan por debajo.

Los ***cuartiles*** son los percentiles **25, 50 y 75**. Se denominan así puesto que dividen los datos en cuatro subconjuntos del mismo tamaño. Se denominan de la siguiente manera:

Q_1 = percentil 25 = 1.er cuartil
M = percentil 50 = 2.º cuartil o mediana
Q_3 = percentil 75 = 3.er cuartil

La manera de calcularlos es similar a la mediana, pero en lugar de buscar el valor de en medio, **buscamos el percentil** en cuestión.

Para calcular el Q_1 y el Q_3, una opción es **calcular la mediana**, usarla para dividir los datos en **dos subconjuntos**, y luego volver a **calcular la mediana** para cada uno de los subconjuntos resultantes.

En nuestro ejemplo, primero ordenamos de menor a mayor:

1, 1, 2, 5, 6, 10, 13, 17, 19, 21, 29, 30

Antes hemos calculado que la mediana era de 11,5 ºC. Efectivamente, esta divide nuestros datos en dos subconjuntos del mismo tamaño:

1, 1, 2, 5, 6, 10 | 13, 17, 19, 21, 29, 30

Como tenemos dos subconjuntos de $n = 6$ para ambos cuartiles buscaremos el valor entre las posiciones 3 y 4 y haremos la media. Para el Q_1 eso corresponde a los valores 2 y 5, mientras que para el Q_3 eso corresponde a los valores 19 y 21.

Así, en nuestro ejemplo, los cuartiles son Q_1 = 3,5 ºC, M = 11,5 ºC, Q_3 = 20 ºC.

RANGO INTERCUARTIL

Se denota como **IQR** (*interquartile rage*) y es la diferencia entre el Q_1 y el Q_3:

$$IQR = Q_3 - Q_1$$

En nuestro ejemplo, el rango intercuartil es de 16,5 ºC.

MÍNIMO Y MÁXIMO

Como su nombre indica, este estadístico nos indica el **valor mínimo (Mín.)** y el **valor máximo (Máx.)** de nuestros datos.

En nuestro ejemplo, el mínimo es 1 ºC y el máximo, 30 ºC.

RANGO

Usando los dos estadísticos anteriores podemos calcular el rango:

$$Rango = Máx. - Mín.$$

En nuestro ejemplo, el rango es de 29 ºC.

DESVIACIÓN ESTÁNDAR Y VARIANCIA

Conceptualmente, la desviación estándar es **cuán separados** están los **datos de la media**. Cuanto más separados, mayor será la desviación estándar.

Sean $x_1, x_2, \ldots, x_n$ las n observaciones, y sea $\overline{x}$ la media, calculamos la desviación estándar con la siguiente fórmula:

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

La **variancia** es la desviación estándar al cuadrado:

$$var = s^2$$

$$s = \sqrt{var}$$

En nuestro ejemplo:

$$s = \sqrt{\frac{(1 - 12{,}8)^2 + (2 - 12{,}8)^2 + \ldots + (5 - 12{,}8)^2 + (1 - 12{,}8)^2}{12 - 1}} = 10{,}4^{\circ}C$$

RECUERDA

Tanto la desviación estándar, como el rango, como el rango intercuartil son medidas de dispersión; nos indican cuán "separados" o dispersos son los datos.

OUTLIERS

Un *outlier* es una **observación anormal** en una muestra estadística, una observación que es numéricamente distante del resto de los datos.

Algunos de los estadísticos que acabamos de comentar pueden estar afectados por outliers; decimos que son *sensibles a los outliers*. Estos son: media, mínimo y máximo, desviación estándar y variancia.

2.2.1. RESUMEN DE LOS 5 NÚMEROS

Conociendo todos los estadísticos anteriores podemos calcular lo que se denomina el *resumen de los 5 números*. Este es:

$$(\text{Mín.}, Q_1, M, Q_3, \text{Máx.})$$

IMPORTANTE

El resumen de los 5 números no resulta muy útil para comparar distribuciones.

Supongamos que tenemos las temperaturas de una segunda ciudad.

ENE	FEB	MAR	ABR	MAY	JUN	JUL	AGO	SET	OCT	NOV	DIC
5	23	25	27	28	30	35	40	30	27	26	25

Temperatura ciudad 2 (media por mes).

Ahora podemos comparar el resumen de los 5 números de ambas.

	MÍN.	Q_1	M	Q_3	MÁX.
CIUDAD 1	1	3,5	11	20	30
CIUDAD 2	5	25	27	30	40

Resumen de los 5 números–Temperaturas Ciudad 1 y Ciudad 2.

Claramente, se observa que la ciudad 2 tiene un clima más cálido, puesto que todos los estadísticos muestran temperaturas superiores a la ciudad 1.

También parece que la ciudad 1 tiene un clima más constante, ya que tiene un rango de 29 ºC de diferencia, mientras que la ciudad 2 tiene un rango de 35 ºC de diferencia.

Observa que hemos comentado que "parece" que la ciudad 1 tiene un clima más constante, ya que podría ser que el rango de la ciudad 2 no fuera fiable debido a los outliers.

Para evitar confusiones, usaremos el **rango intercuartiles**, ya que este no es sensible a los outliers.

En nuestro ejemplo, precisamente, el rango intercuartil nos dice lo opuesto al rango; la ciudad 2 tiene un IQR de 5 ºC mientras que la ciudad 1 tiene un IQR de 16,5 ºC. Este sí que es un indicador fiable de que la temperatura de la ciudad 2 es más constante que la de la ciudad 1, hay menos dispersión en las observaciones de la ciudad 2.

2.2.2. REPRESENTACIÓN GRÁFICA

Las representaciones gráficas más comunes para una variable numérica son los **histogramas** y los **diagramas de cajas**.

HISTOGRAMA

Para crear un histograma los datos son reducidos a **intervalos**. Para cada intervalo, se calcula la **frecuencia** (o la frecuencia relativa) y se realiza el gráfico de barras correspondiente.

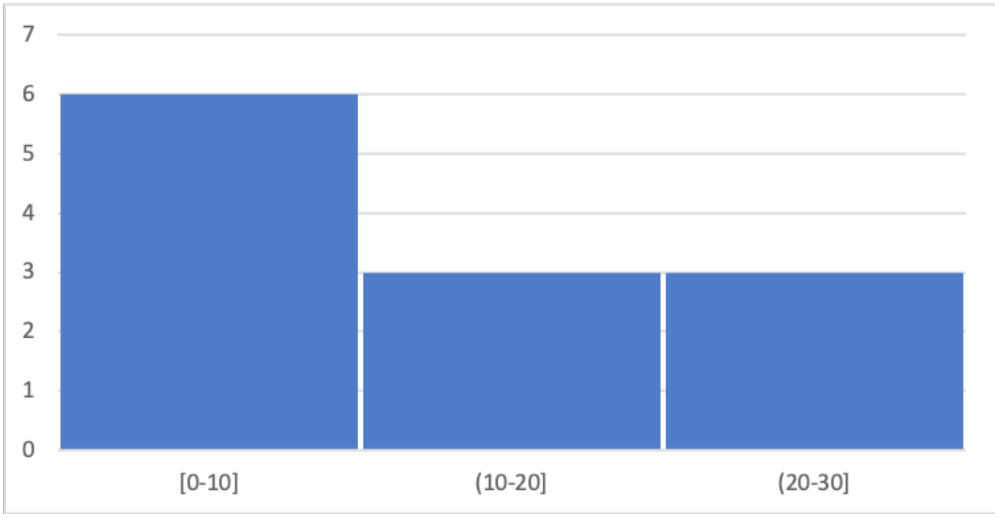
SABÍAS QUE...

El número de intervalos dependerá del nivel de detalle que se necesite.

Siguiendo nuestro ejemplo, vamos a hacer 3 intervalos: 0-10, 10-20, 20-30. Así, si partimos de la tabla "Temperatura ciudad 1 (media por mes)" (página 25), tenemos las siguientes frecuencias:

INTERVALO	FRECUENCIA	FRECUENCIA RELATIVA
[0-10]	6	0,50
[10-20]	3	0,25
[20-30]	3	0,25
TOTAL	12	1

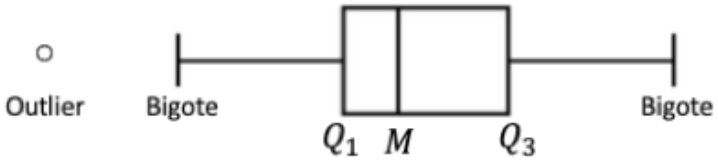
Tabla de frecuencias y frecuencias relativas – intervalos de temperatura ciudad.



Histograma – temperatura ciudad.

DIAGRAMA DE CAJAS

El diagrama de cajas es la representación gráfica del **resumen de los 5 números**.



Prototipo diagrama de cajas.

La caja está delimitada por el Q_1 y el Q_3, identificando la **mediana** con una raya que divide la caja en dos.

Luego están los **bigotes**. Sus límites se calculan con las siguientes fórmulas:

$$\text{Límite bigote inferior} = Q_1 - 1{,}5 \times \text{IQR}$$

$$\text{Límite bigote superior} = Q_3 + 1{,}5 \times \text{IQR}$$

- El **bigote inferior** será el mínimo valor de nuestros datos comprendido en el intervalo (límite bigote inferior, límite bigote superior).
- El **bigote superior** será el máximo valor de nuestros datos comprendido en el intervalo (límite bigote inferior, límite bigote superior).

IMPORTANTE

Todos los valores fuera de este intervalo son outliers, y todos se representarán en el diagrama de cajas como puntos sueltos.

En nuestro ejemplo, este sería el diagrama de cajas:

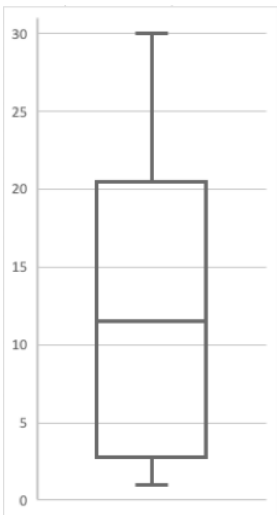


Diagrama de cajas – temperatura ciudad.

Si te fijas, al no haber outliers, cuadra perfectamente con el resumen de los 5 números.

SABÍAS QUE...

Los diagramas de cajas son muy útiles para comparar dos distribuciones.

Por ejemplo, podemos comparar las dos ciudades de antes.

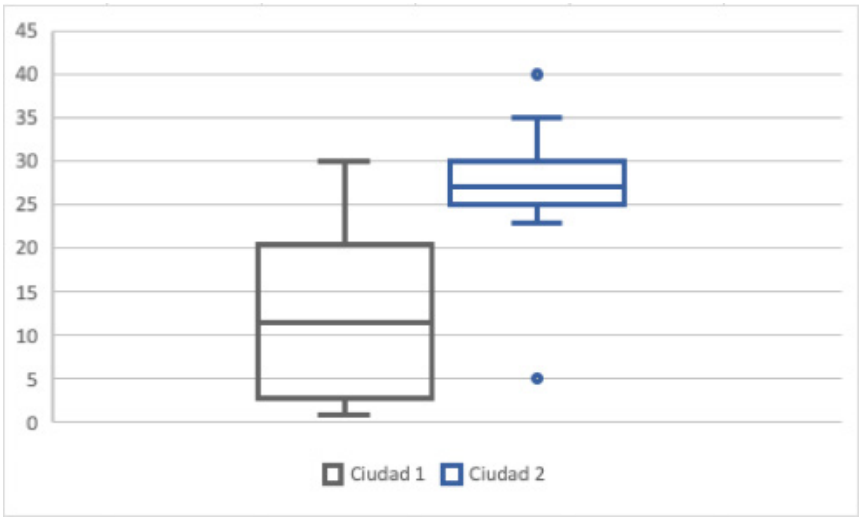
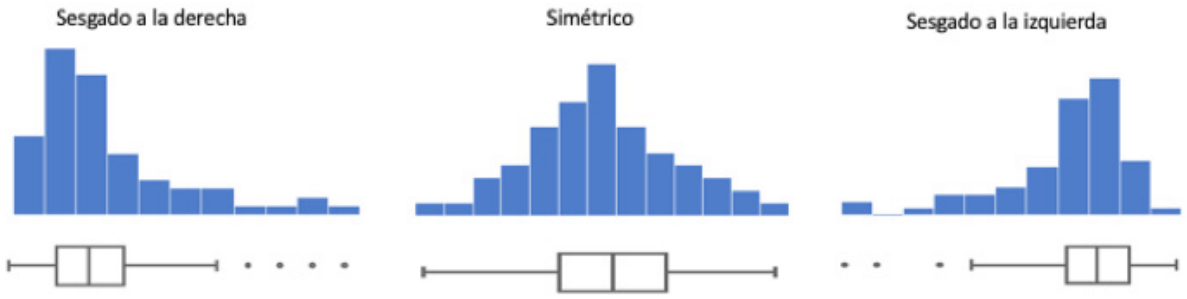


Diagrama de cajas – temperaturas ciudad 1 y ciudad 2.

Aquí vemos que, efectivamente, la ciudad 2 tiene un clima más cálido. Asimismo, que su clima es más constante, ya que la caja y los bigotes son más pequeños. También queda claro que las temperaturas mínima y máxima registradas son outliers.

Observa que un histograma y un diagrama de cajas están bastante relacionados; por lo general con ambos podemos llegar a las mismas conclusiones respecto a la distribución de los datos.



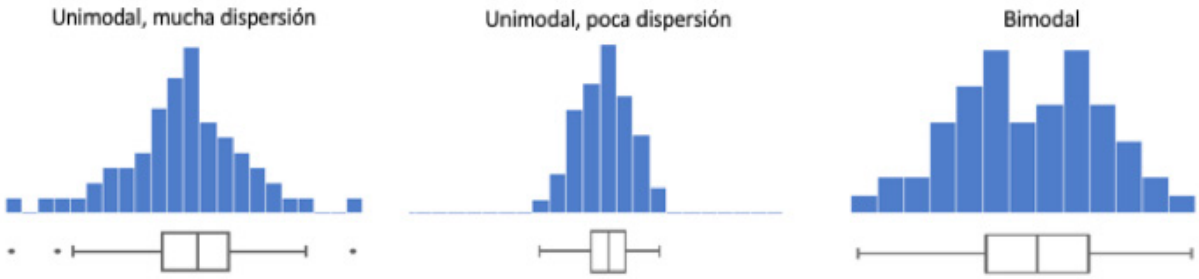
Ejemplo 1 de histogramas y sus respectivos diagramas de cajas.

En la figura anterior, el primer caso del histograma está sesgado a la derecha, es decir $M < \overline{x}$. Con el diagrama de cajas vemos que efectivamente la mayor parte de los datos está a la derecha.

En el segundo caso, tanto el histograma como el diagrama de cajas son simétricos, es decir, $M = \overline{x}$.

En el tercer caso, el histograma está sesgado a la izquierda, es decir, $M > \overline{x}$. Con el diagrama de cajas también vemos que las observaciones de la izquierda tienen menos dispersión que las de la derecha.

Veamos más ejemplos.



Ejemplo 2 de histogramas y sus respectivos diagramas de cajas.

En la figura anterior los tres casos son simétricos.

En el primer caso hay poca dispersión; esto lo podemos ver porque el histograma no tiene colas largas y el diagrama de cajas tiene los bigotes cortos.

En el segundo caso hay mucha dispersión; contrariamente al primero, el histograma tiene colas largas y el diagrama de cajas, bigotes largos e incluso outliers.

En el tercer caso, gracias al histograma, podemos ver que tenemos una distribución bimodal (con dos modas). Esto es algo que un diagrama de cajas no puede reflejar.

IMPORTANTE

Por este motivo (que un diagrama de cajas no puede reflejar una distribución bimodal) es útil siempre hacer un análisis previo de los datos (calcular los estadísticos que hemos visto) y **usar más de una visualización**, ya que dependiendo de la distribución de los datos podemos ver aspectos distintos según la visualización elegida.

3. ESTIMADORES BIVARIABLES

En este tema vamos a tratar cómo analizar la **relación entre dos variables**. Igual que en el tema anterior, diferenciaremos entre variables categóricas y variables numéricas.

3.1. VARIABLES CATEGÓRICAS

Para representar la relación entre dos variables categóricas se usan **tablas de contingencia**. Sería el equivalente a una tabla de frecuencias, donde las filas representan la variable 1 y las columnas, la variable 2.

Supongamos que queremos estudiar la relación entre el peso de la gente y su nivel de actividad física. Para ello, seleccionamos una muestra representativa de la población y recogemos los siguientes datos:

- Peso, clasificando entre: magro, ideal o sobrepeso.
- Nivel de actividad física: baja, moderada o intensa.

		PESO			
		MAGRO	IDEAL	SOBREPESO	**TOTAL**
ACTIVIDAD FÍSICA	BAJA	153	284	1.021	1.458
	MODERADA	238	644	445	1.327
	INTENSA	67	1.462	32	1.561
	TOTAL	458	2.390	1.498	4.346

Tabla de contingencia – peso vs. actividad física.

Las filas del total son las tablas de frecuencias de cada variable. También, en este caso, podemos representar los datos como porcentajes respecto al total.

En el ejemplo anterior la tabla quedaría así:

		PESO			
		MAGRO	IDEAL	SOBREPESO	**TOTAL**
ACTIVIDAD FÍSICA	BAJA	0,04	0,07	0,23	0,34
	MODERADA	0,05	0,15	0,10	0,30
	INTENSA	0,02	0,34	0,01	0,36
	TOTAL	0,11	0,55	0,34	1

Distribuciones marginales – peso vs. actividad física.

Las filas del total corresponderían a la tabla de frecuencias relativas de cada variable, y se denominan ***distribución marginal***. El término *marginal* proviene de que solamente examinamos una variable.

En el ejemplo, la distribución marginal de la actividad física sería 0,34, 0,30, 0,36. Es decir, si yo pregunto a una persona al azar, hay un 34 % de probabilidad de que esta tenga una actividad física baja.

Esto está muy bien para analizar las dos variables por separado, pero no resulta útil para analizarlas conjuntamente. Para ello usamos la **distribución condicional**, donde calculamos los porcentajes respecto a una variable solo.

		PESO		
		MAGRO	IDEAL	SOBREPESO
ACTIVIDAD FÍSICA	BAJA	0,33	0,12	0,68
	MODERADA	0,52	0,27	0,30
	INTENSA	0,15	0,61	0,02
	TOTAL	1	1	1

Distribución condicional de la actividad física respecto a los diferentes niveles de peso.

Esto es muy útil para conocer las probabilidades de pertenecer a cada grupo. Por ejemplo, si tomo a un ciudadano al azar y veo que tiene sobrepeso, sé que hay un 68 % de probabilidad de que este tenga una actividad física baja.

Si la distribución de una variable es la misma para todas las categorías de la segunda variable, entonces estas dos variables son **independientes**.

Por ejemplo, si la probabilidad de tener una actividad física baja fuera siempre del 30 % independientemente de si el peso es magro, ideal o sobrepeso, entonces significaría que el peso no tiene nada que ver con la actividad física, ambas serían variables independientes.

3.1.1. REPRESENTACIÓN GRÁFICA

La representación gráfica más común es mediante **barras agrupadas o apiladas**.

BARRAS AGRUPADAS

Este gráfico es parecido al gráfico de barras usado para representar una sola variable. Cada barra es una categoría, y la longitud de estas representa la frecuencia.

La diferencia reside en que para cada variable del eje X (variable 1) hay tantas barras como categorías de la variable 2.

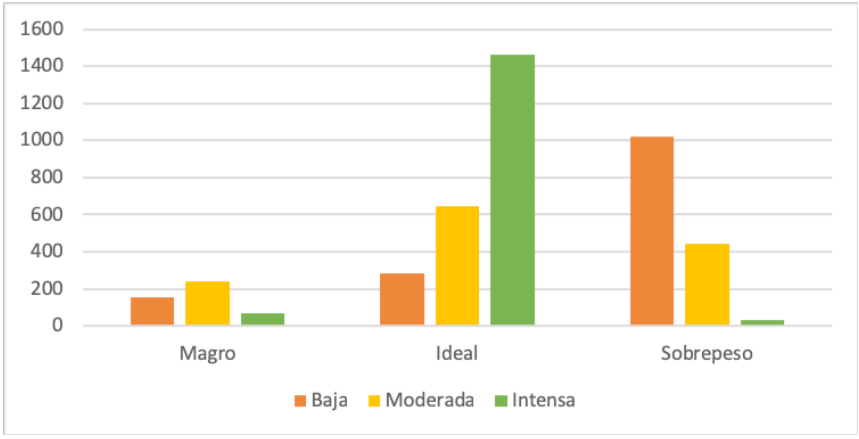


Gráfico de barras agrupadas peso vs. actividad física.

Este gráfico nos resulta muy útil para comparar la distribución de la actividad física entre los diferentes pesos.

Vemos claramente que la mayoría de la gente con un peso ideal tiene una actividad física intensa, mientras que la mayoría de gente con sobrepeso tiene una actividad física baja.

BARRAS APILADAS

Este es similar al gráfico de barras agrupadas, pero en vez de tener una barra para cada categoría de la variable 2, hay una sola barra y los valores se sobreponen uno encima de otro.

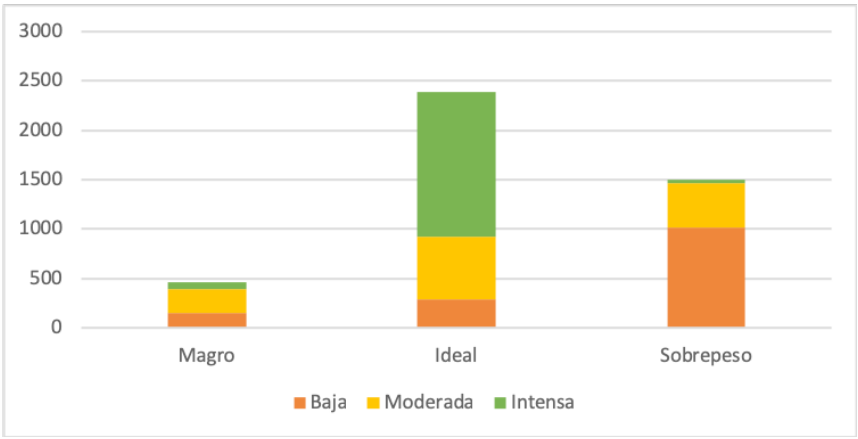


Gráfico de barras apiladas – peso vs. actividad física.

Esta visualización es muy adecuada para comparar los diferentes pesos entre ellos.

Podemos observar que la mayoría de la población tiene un peso ideal, seguido por un grupo considerable de gente con sobrepeso y, finalmente, un conjunto muy reducido de gente por debajo del peso recomendado.

SABÍAS QUE...

Este tipo de gráfico también se puede realizar sobre la distribución condicional en cuyo caso todas las columnas sumarán el 100 %.

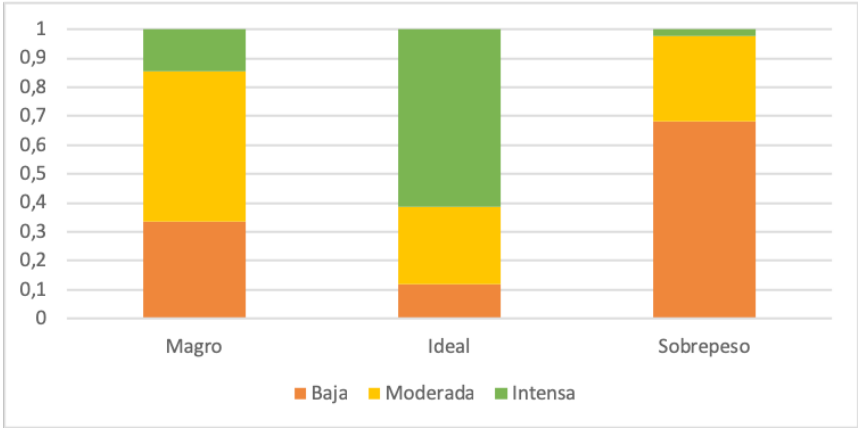


Gráfico de barras apiladas sobre la distribución condicional – peso vs. actividad física.

Igual que con las barras agrupadas, visualizar las probabilidades condicionales nos permite comparar la distribución de la actividad física entre los diferentes pesos fácilmente. No solo vemos que la mayoría de la gente con un peso ideal tiene una actividad física intensa y que la mayoría de gente con sobrepeso tiene una actividad física baja, sino que también comprobamos claramente que la gente por debajo de su peso ideal suele tener una actividad física moderada.

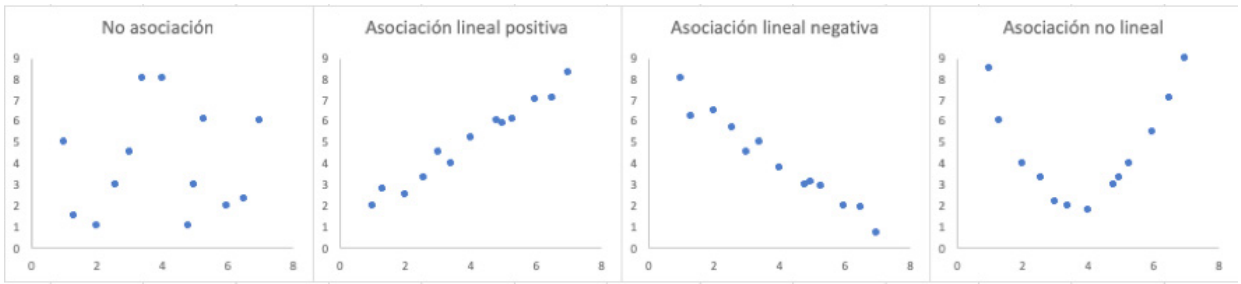
3.2. VARIABLES CUANTITATIVAS

Para estudiar la relación entre dos variables numéricas, miraremos lo correlacionadas que están. La **correlación** mide la asociación entre dos variables aleatorias.

Veremos en detalle las asociaciones entre variables en el siguiente apartado (representación gráfica) pero, a grandes rasgos, dos variables pueden:

- No estar asociadas.
- Estar asociadas linealmente (asociación lineal positiva, asociación lineal negativa).
- Estar asociadas no linealmente (asociación exponencial, asociación cuadrática, etc.).

Veamos un ejemplo de cada.



Ejemplos de asociación entre variables numéricas.

IMPORTANTE

Es fundamental no confundir correlación con causalidad.

Cuando hay causalidad, una variable implica la otra.

EJEMPLO

El fumar y el tener cáncer de pulmón, cuanto más fumas, más aumentan las probabilidades de tener cáncer de pulmón. Fumar implica que las probabilidades de tener cáncer de pulmón aumenten.

Cuando existe correlación simplemente es que estas dos variables tienen algún tipo de asociación.

EJEMPLO

La venta de helados está correlacionada con el uso de ropa veraniega. Cuando aumenta la venta de helados, también lo hace el número de gente que viste con ropa veraniega. Pero obviamente comprarte un helado no implica que lleves ropa de verano ni que llevar ropa de verano conlleva que vayas a comprarte un helado. Es obvio que ambas variables están relacionadas con el calor; más calor implica más helados y más ropa veraniega, pero en otros casos esto no es tan claro.

Existen diversas maneras de calcular la correlación; la más común es mediante el **coeficiente de correlación de Pearson** que mide la asociación lineal entre dos variables numéricas. Aunque sea el coeficiente de correlación más usado, el coeficiente de Pearson es sensible a outliers, por este motivo se deben **limpiar los datos** antes de calcularlo.

Hay otros coeficientes de correlación que no son sensibles a los outliers que, aunque no los tratemos aquí, conviene saber cuáles son: el coeficiente de Kendall y el de Spearman.

COEFICIENTE DE CORRELACIÓN DE PEARSON

Este coeficiente mide lo fuerte que es una relación lineal. Puede tomar valores entre [-1,1]:

- r = -1 indica una perfecta asociación lineal negativa.
- r = 1, una perfecta asociación lineal positiva.
- r = 0, que no hay ninguna asociación lineal.

IMPORTANTE

Que no exista asociación lineal no significa que no haya ningún tipo de correlación entre las variables. Podemos tener un coeficiente $r = 0$ y que las dos variables tengan una relación no lineal.

Siendo $(x_1,y_1),\ldots,(x_n,y_n)$ las observaciones, calculamos el coeficiente de correlación mediante la siguiente fórmula:

$$r = \frac{\sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{s_x}\right)\left(\frac{y_i - \overline{y}}{s_y}\right)}{n - 1}$$

EJEMPLO

Imaginemos que queremos estudiar la relación entre las precipitaciones de una región (x) y el número de incendios (y) que se producen en esta. Para ello recogemos datos de 10 regiones.

REGIÓN	PRECIPITACIONES (mm)	NÚMERO DE INCENDIOS
R1	30	4
R2	105	0
R3	78	1
R4	42	3
R5	21	5
R6	60	2
R7	102	0
R8	55	2
R9	81	0
R10	83	1

Precipitaciones y número de incendios por región.

Si calculamos la media y la desviación estándar para las dos variables tenemos:

$$\overline{x} = 65{,}7;\ s_x = 28{,}9;\ \overline{y} = 1{,}8;\ s_y = 1{,}8$$

Ahora ya podemos calcular el coeficiente de correlación de Pearson:

$$r = \frac{\left(\frac{30 - 65,7}{28,9}\right)\left(\frac{4 - 1,8}{1,8}\right) + \ldots + \left(\frac{83 - 65,7}{28,9}\right)\left(\frac{1 - 1,8}{1,8}\right)}{10 - 1} = -0,97$$

Si analizamos el resultado, como $r \cong -1$ hay una relación lineal muy fuerte. Por otra parte, como $r < 0$ significa que la asociación lineal es negativa.

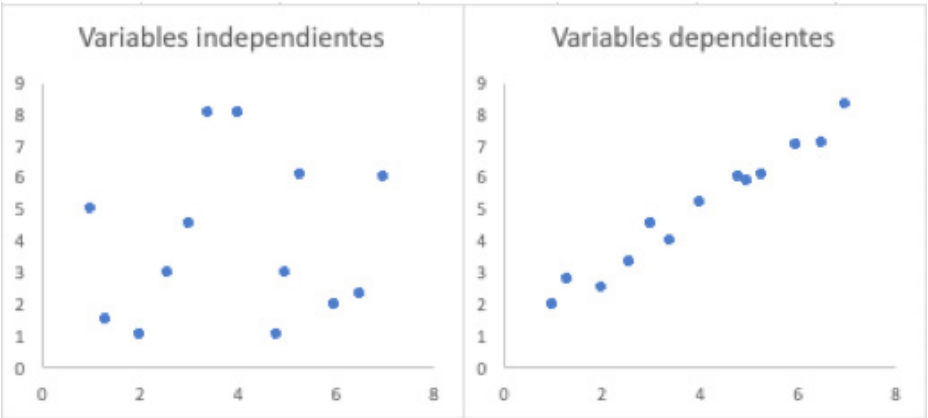
3.2.1. REPRESENTACIÓN GRÁFICA

Para representar dos variables cuantitativas gráficamente se usan los **diagramas de dispersión**.

- El **eje X** representa una de las dos variables. Por convenio suele representar la variable explicativa (variable independiente), la que se usa para predecir los resultados de la otra. En el ejemplo anterior, la variable explicativa serían las precipitaciones en una región.
- El **eje Y** representa la otra, la variable de respuesta (variable dependiente). En el ejemplo anterior, la variable respuesta serían el número de incendios.

Para cada observación se dibuja un punto.

Este tipo de gráfico nos permite saber de forma visual si dos variables están asociadas. Si los puntos tienen una **tendencia**, significa que existe una correlación entre las dos variables, mientras que, si los puntos están **dispersos** de manera aleatoria, probablemente las variables son independientes.



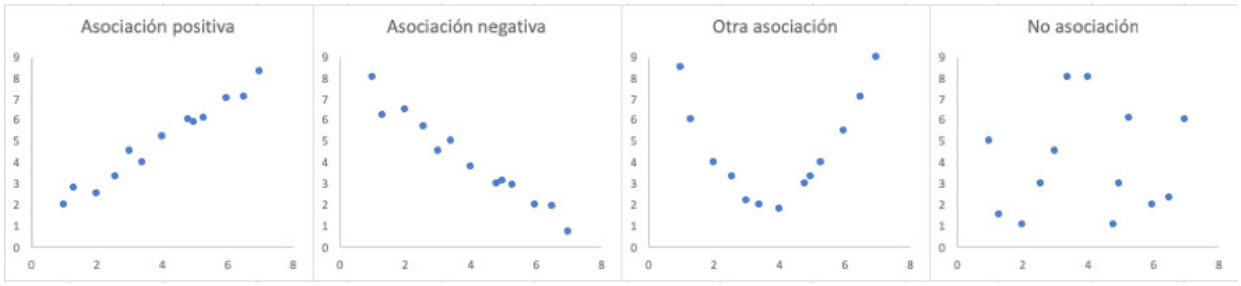
Ejemplo de dependencia entre variables en un diagrama de dispersión.

Para describir la asociación entre dos variables mediante un diagrama de dispersión **cuatro propiedades** nos ayudan:

- Dirección de la asociación.
- Forma
- Fuerza.
- Puntos que destacar.

DIRECCIÓN DE LA ASOCIACIÓN

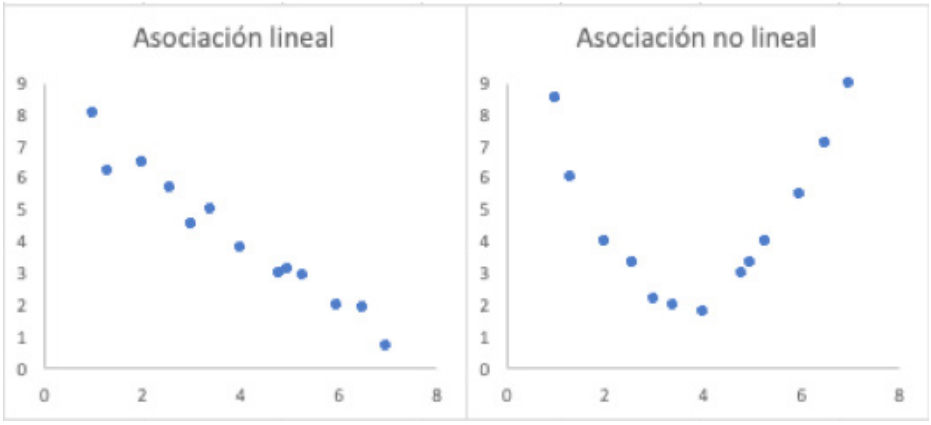
- **Asociación positiva**: diremos que existe una asociación positiva cuando, si incrementa el valor de una variable, también incrementa el de la otra.
- **Asociación negativa**: diremos que hay una asociación negativa cuando, si aumenta el valor de una variable, disminuye el de la otra.
- **Otro tipo de asociación**: ocurre cuando las dos variables están asociadas positivamente y negativamente.
- **No asociación**: en este caso, no podemos apreciar ningún tipo de relación en el gráfico.



Ejemplo de direcciones de asociación en un diagrama de dispersión.

FORMA

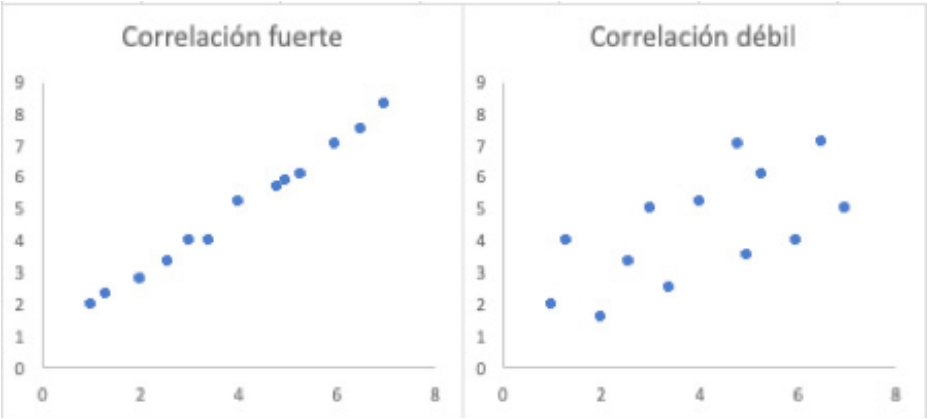
- **Lineal**: consideraremos que dos variables tienen una asociación lineal cuando los puntos del diagrama de dispersión siguen una línea recta.
- **No lineal**: consideraremos que dos variables tienen una asociación no lineal cuando los puntos del diagrama de dispersión no siguen una línea recta. Las distribuciones no lineales más comunes son la **exponencial** y la **cuadrática**.



Ejemplo de formas en un diagrama de dispersión.

FUERZA

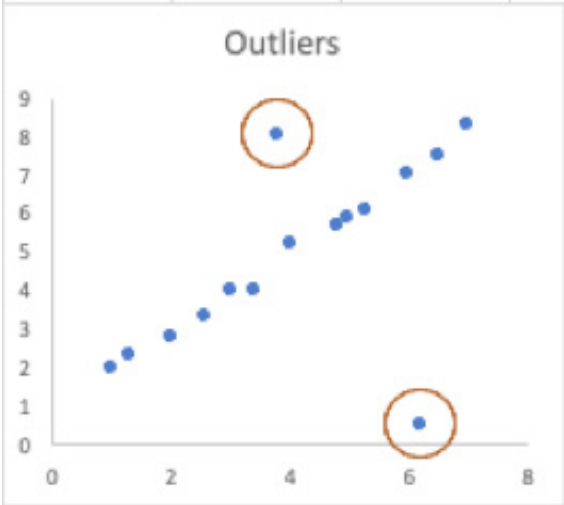
La fuerza de correlación entre dos variables está relacionada con la dispersión de estas. Cuanto **más fuerte** sea la relación, **más clara será la tendencia** en el gráfico, es decir, más definida estará la forma. Si la fuerza es débil, los puntos estarán más dispersos, formando una nube poco definida.



Ejemplo de fuerza de correlación en un diagrama de dispersión.

PUNTOS QUE DESTACAR

Es decir, **outliers**, puntos que se salen del patrón. Por ejemplo, en un diagrama de dispersión con una forma lineal fuerte, un punto que queda fuera de la línea sería un punto por destacar.



Ejemplo de puntos por destacar en un diagrama de dispersión.

EJEMPLO

Estudiemos el diagrama de dispersión de los datos que hemos visto antes: precipitaciones vs. número de incendios.

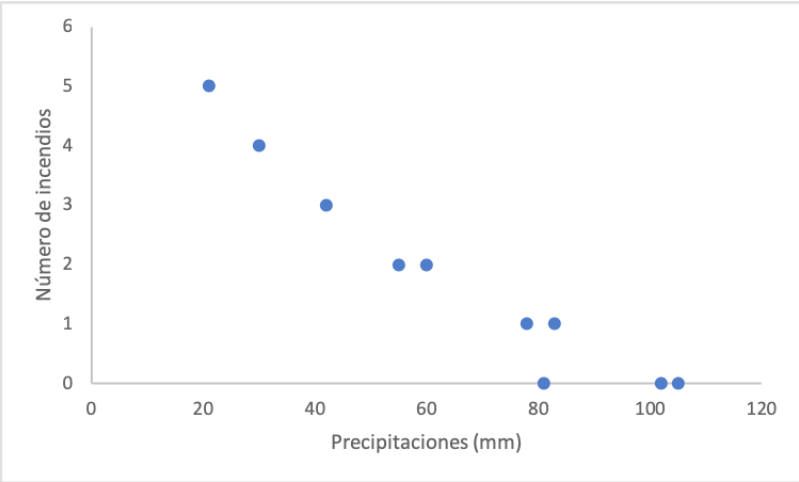


Diagrama de dispersión – precipitaciones vs. número de incendios.

Y, a continuación, analicemos las cuatro propiedades vistas anteriormente:

- Dirección de asociación: negativa.
- Forma: lineal.
- Fuerza: correlación fuerte.
- Puntos que destacar: no.

Por lo tanto, estas dos variables tienen una fuerte relación lineal negativa. Es la misma conclusión a la que habíamos llegado con el coeficiente de correlación de Pearson.

3.3. REGRESIÓN

IMPORTANTE

La **regresión** es un modelo entre dos variables que nos permite, dado un valor de la variable explicativa *x*, predecir el valor de la variable respuesta *y*.

Con una muestra de observaciones de dos variables es muy probable que no tengamos observaciones para todos los puntos de la variable explicativa *x*.

Si cogemos el ejemplo anterior con las precipitaciones y el número de incendios, solo tenemos observaciones para algunas cantidades de precipitaciones. Esto no siempre es suficiente; ¿qué ocurriría si el año siguiente en una región hay 70 mm de precipitaciones? ¿Cómo sabré cuántos incendios habrá si hasta la fecha nunca ha habido esa cantidad exacta de precipitaciones?

Aquí es donde entra el concepto de regresión, que nos permite **estimar la relación** entre dos variables, sean cualitativas o cuantitativas.

3.3.1. REGRESIÓN LOGÍSTICA

Es la regresión usada cuando se quiere ver la asociación entre **una variable respuesta categórica** y **una o más variables explicativas** (pueden ser categóricas o cuantitativas). Analizaremos solo el caso de las variables categóricas.

Lo que haremos es representar la variable categórica mediante una variable cuantitativa que toma un número diferente para cada uno de sus valores. Para simplificar nos vamos a centrar en variables respuesta dicotómicas.

EJEMPLO

Supongamos que tenemos las notas finales de varios alumnos que han terminado la educación secundaria obligatoria (ESO). Concretamente sabemos si un alumno ha sacado una nota alta (notable o excelente) o una nota baja, y con esto queremos predecir si el alumno hará bachillerato o no.

	NOTA ALTA ($X = 1$)	**NOTA BAJA ($X = 0$)**	
BACHILLERATO ($Y = 1$)	63	25	88
NO BACHILLERATO ($Y = 0$)	52	94	146
	115	**119**	**234**

Tabla de contingencia – estudios vs. notas.

En este caso tenemos dos alternativas de variable respuesta: hacer bachillerato, no hacer bachillerato. La variable respuesta puede tomar dos valores $Y = \{0,1\}$.

Sean $X_1,...,X_n$ las variables explicativas, el modelo es:

$$Y = b_0 + b_1 X_1 + ... + b_n X_n$$

Donde b_0 es el término independiente o constante y $b_0,...,b_n$ son los coeficientes de regresión asociados a las variables independientes.

En el ejemplo anterior tenemos una sola variable explicativa $X_1 = \{0,1\}$. Explicaremos el modelo siguiendo el ejemplo. Hemos elegido un ejemplo con una sola variable explicativa para simplificar los cálculos.

Nuestro objetivo es, dada una nota alta o baja, estimar la probabilidad de que el alumno haga bachillerato, es decir *P(Y* = 1*)*. Esta probabilidad viene dada por la siguiente fórmula:

$$P(Y = 1) = \frac{1}{1 + e^{-(b_0 + b_1 X_1)}}$$

Fíjate que esta fórmula garantiza que el valor obtenido se encuentre entre 0 y 1.

Aplicando un poco de álgebra vemos que la ecuación anterior es equivalente a:

$$\ln\left(\frac{P(Y = 1)}{1 - P(Y = 1)}\right) = b_0 + b_1 X_1$$

Para calcular los coeficientes b_i se usa el **método de máxima verosimilitud**. La idea fundamental de este método es tomar, como estimación del parámetro estudiado, el valor que haga máxima la probabilidad de obtener la muestra observada.

Las ecuaciones para calcular los coeficientes b_i usando el método de máxima verosimilitud son tediosas y complicadas, por este motivo no las veremos aquí. Pero ello no impide que podamos calcular los coeficientes de nuestro ejemplo.

Si usamos la fórmula anterior, podemos encontrar b_0, ya que por $X_1 = 0$ tenemos:

$$b_0 = \ln\left(\frac{P(Y = 1 \mid X_1 = 0)}{1 - P(Y = 1 \mid X_1 = 0)}\right) = \ln\left(\frac{25/119}{1 - 25/119}\right) = \ln(0{,}27) = -1{,}3$$

Observa que:

$$\frac{P(Y = 1)}{1 - P(Y = 1)} = \frac{P(Y = 1)}{P(Y = 0)}$$

La probabilidad de un suceso dividida entre la probabilidad de que no ocurra dicho suceso se conoce como ***odds***.

En nuestro ejemplo:

$$\frac{P(Y = 1)}{P(Y = 0)} = \frac{P(\text{bachillerato})}{P(\text{no bachillerato})} = \frac{88/234}{146/234} = 0{,}6$$

También podemos calcular el odds dado que la nota sea alta:

$$\frac{P(Y = 1 \mid X_1 = 1)}{P(Y = 0 \mid X_1 = 1)} = \frac{63/115}{52/115} = 1{,}21$$

Y el odds dado que la nota sea baja:

$$\frac{P(Y = 1 \mid X_1 = 0)}{P(Y = 0 \mid X_1 = 0)} = \frac{25/119}{94/119} = 0{,}27$$

Si dividimos estos dos últimos tenemos el **odds ratio** (**OR**), que cuantifica cuánto más probable es cursar bachillerato si la nota de la ESO es alta respecto a cuando la nota de la ESO es baja.

$$OR = \frac{1{,}21}{0{,}27} = 4{,}5$$

Por tanto, es 4,5 veces más probable que un alumno curse bachillerato si su nota de la ESO es alta.

IMPORTANTE

En términos generales, el OR es el cociente de la probabilidad de que se produzca un suceso (hacer bachillerato) cuando está presente el factor (nota alta) respecto a cuando no lo está.

En variables explicativas dicotómicas, como es nuestro caso: $X_1 = \{0,1\}$, el odds ratio nos permite calcular el coeficiente asociado b_1.

$$b_1 = \ln (OR_{X_1}) = \ln (4{,}5) = 1{,}5$$

De este modo ya tenemos nuestro modelo:

$$\ln \left(\frac{P(\text{bachillerato})}{P(\text{no bachillerato})} \right) = -1{,}3 + 1{,}5\, X_1$$

O, equivalentemente:

$$P(Y = 1) = \frac{1}{1 + e^{-(-1{,}3+1{,}5X_1)}}$$

Fórmula que nos permite calcular nuestro objetivo: la probabilidad de que un alumno haga bachillerato en el caso de que haya sacado una nota alta ($X_1 = 1$):

$$P(Y = 1) = \frac{1}{1 + e^{-(-1{,}3+1{,}5 \cdot 1)}} = 0{,}55$$

Y, en el caso de que haya sacado una nota baja ($X_1 = 0$):

$$P(Y = 1) = \frac{1}{1 + e^{-(-1,3+1,5\cdot 0)}} = 0,21$$

Fíjate que en este caso las probabilidades coindicen con las de la distribución condicional. Esto ocurre porque tenemos una sola variable explicativa categórica.

	NOTA ALTA ($X = 1$)	NOTA BAJA ($X = 0$)
BACHILLERATO ($Y = 1$)	0,55	0,21
NO BACHILLERATO ($Y = 0$)	0,45	0,79
	1	**1**

Distribución condicional de los estudios respecto a las notas.

3.3.2. REGRESIÓN LINEAL

Es la regresión usada cuando se tienen **dos variables cuantitativas asociadas linealmente**.

La ecuación de una recta es $\hat{y} = b_0 + b_1 x$, donde b_0 es la intersección con el eje y y b_1 es el pendiente de la recta ($b_1 > 0 \rightarrow$ asociación lineal positiva, $b_1 < 0 \rightarrow$ asociación lineal negativa).

En una regresión lineal buscaremos los parámetros b_0 y b_1 tales que la recta obtenida sea la que se ajuste mejor a nuestros datos.

Una vez tengamos estos parámetros, para cada observación x obtendremos un valor $\hat{y}$. Lo ideal es que $\hat{y} = y$, esto significaría que el punto predicho coincide con nuestra observación (x,y).

Para cada observación (x,y) definimos:

- y = variable respuesta real.
- $\hat{y}$ = variable respuesta predicha.
- $e = y - \hat{y}$ = error residual.

Para encontrar la ecuación de la recta, buscaremos b_0 y b_1 tales que minimizan la suma de los errores residuales al cuadrado ($\sum e^2$). El método para ello se denomina ***mínimos cuadrados*** y las fórmulas para tal son:

$$b_1 = r \frac{S_y}{S_x}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

EJEMPLO

Vamos a retomar los datos de precipitaciones vs. el número de incendios.

REGIÓN	PRECIPITACIONES (mm)	NÚMERO DE INCENDIOS
R1	30	4
R2	105	0
R3	78	1
R4	42	3
R5	21	5
R6	60	2
R7	102	0
R8	55	2
R9	81	0
R10	83	1

Precipitaciones y número de incendios por región.

Teníamos los siguientes estadísticos:

$$\bar{x} = 65{,}7;\ s_x = 28{,}9;\ \bar{y} = 1{,}8;\ s_y = 1{,}8;\ r = -0{,}97$$

Con ellos podemos calcular b_0 y b_1:

$$b_1 = -0{,}97\ \frac{1{,}8}{28{,}9} = -0{,}06$$

$$b_0 = 1{,}8 - (-0{,}06) \cdot 65{,}7 = 5{,}7$$

La recta de la regresión lineal según el método de mínimos cuadrados es:

$$\hat{y} = 5{,}7 - 0{,}06x$$

Como el pendiente (b_1 = -0,06) es negativo significa que la recta es decreciente. La manera de interpretarlo es que, por cada aumento de precipitaciones de 1 mm, el número de incendios disminuye en 0,06.

El coeficiente de intersección (b_0 = 5,7) nos indica que en caso de no tener precipitaciones (0 mm) el número de incendios estimado sería $5{,}7 \cong 6$.

Esta ecuación nos permite predecir puntos que no tenemos. Por ejemplo, si una región tiene 70 mm de precipitaciones, ¿cuántos incendios se estima que va a tener?

$$\hat{y} = 5{,}7 - 0{,}06 \cdot 70 = 1{,}5$$

La respuesta es que una región con 70 mm de precipitaciones se estima que va a tener entre 1 y 2 incendios.

IMPORTANTE

Es muy importante tener en cuenta que una regresión lineal solo se puede usar para predecir puntos dentro del intervalo de observaciones, nunca se debe extrapolar.

En nuestro caso, podemos predecir los incendios para regiones con entre 21 mm y 105 mm de precipitaciones.

Vamos a graficar la recta que hemos calculado.

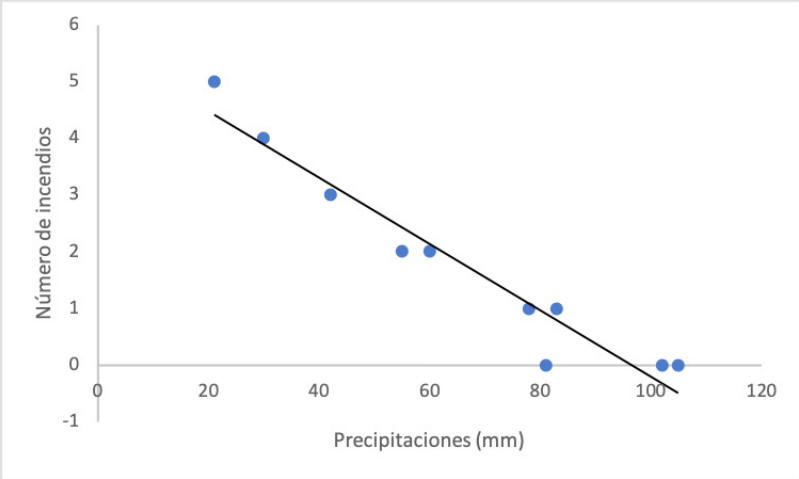


Diagrama de dispersión con regresión lineal.

Con el gráfico queda claro que la recta obtenida se ajusta a los datos de forma óptima.

Una manera de analizar lo buena que es la aproximación es mediante el **gráfico de residuos**. En este gráfico, para cada observación x se calcula $e = y - \hat{y}$. Para interpretarlo hay que tener en cuenta:

- Cuanto más próximos a 0 estén los errores mejor será nuestra regresión.
- Los puntos no deben tener ninguna asociación. Si los puntos del gráfico de residuos muestran algún tipo de asociación (una curva, por ejemplo), entonces el modelo de regresión lineal no es apropiado para representar nuestros datos.

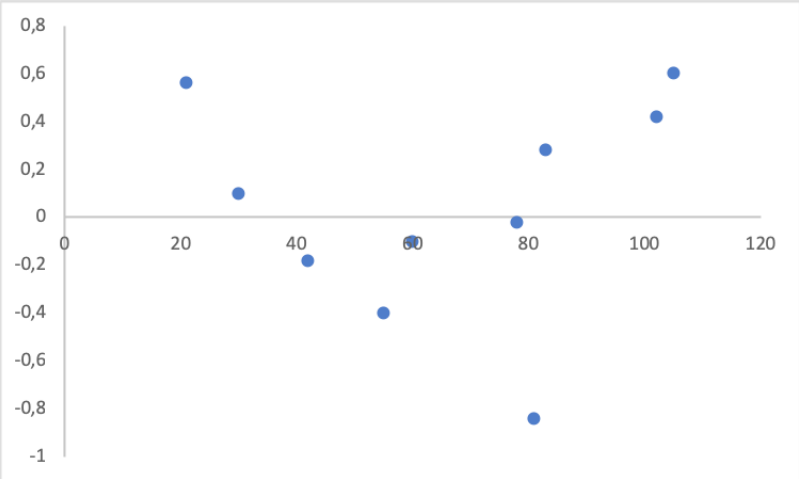


Gráfico de residuos.

Podemos ver que los residuos no muestran ningún tipo de asociación y todos son próximos a 0, lo que nos indica que la regresión lineal es apropiada para representar nuestros datos.

3.3.3. REGRESIÓN NO LINEAL

Es la regresión utilizada cuando se tienen **dos variables cuantitativas no asociadas linealmente**.

Aquí no veremos ejemplos de regresiones no lineales, pero el procedimiento es el mismo que en una regresión lineal, solo que en este caso buscaremos los parámetros $b_0,\ldots,b_n$ tales que la curva obtenida sea la que se ajuste mejor a nuestros datos:

$$\hat{y} = b_0 + b_1 x + \ldots + b_n x^n$$

IMPORTANTE

Antes de optar por una regresión no lineal se recomienda ver si se pueden transformar los datos de las observaciones para tener una asociación lineal entre ellos, ya que eso facilita mucho los cálculos y la comprensión.

Por ejemplo, si tenemos una asociación exponencial, podemos hacer la transformación $(x,y) \sim (x, \log(y))$ y tendremos una asociación lineal.

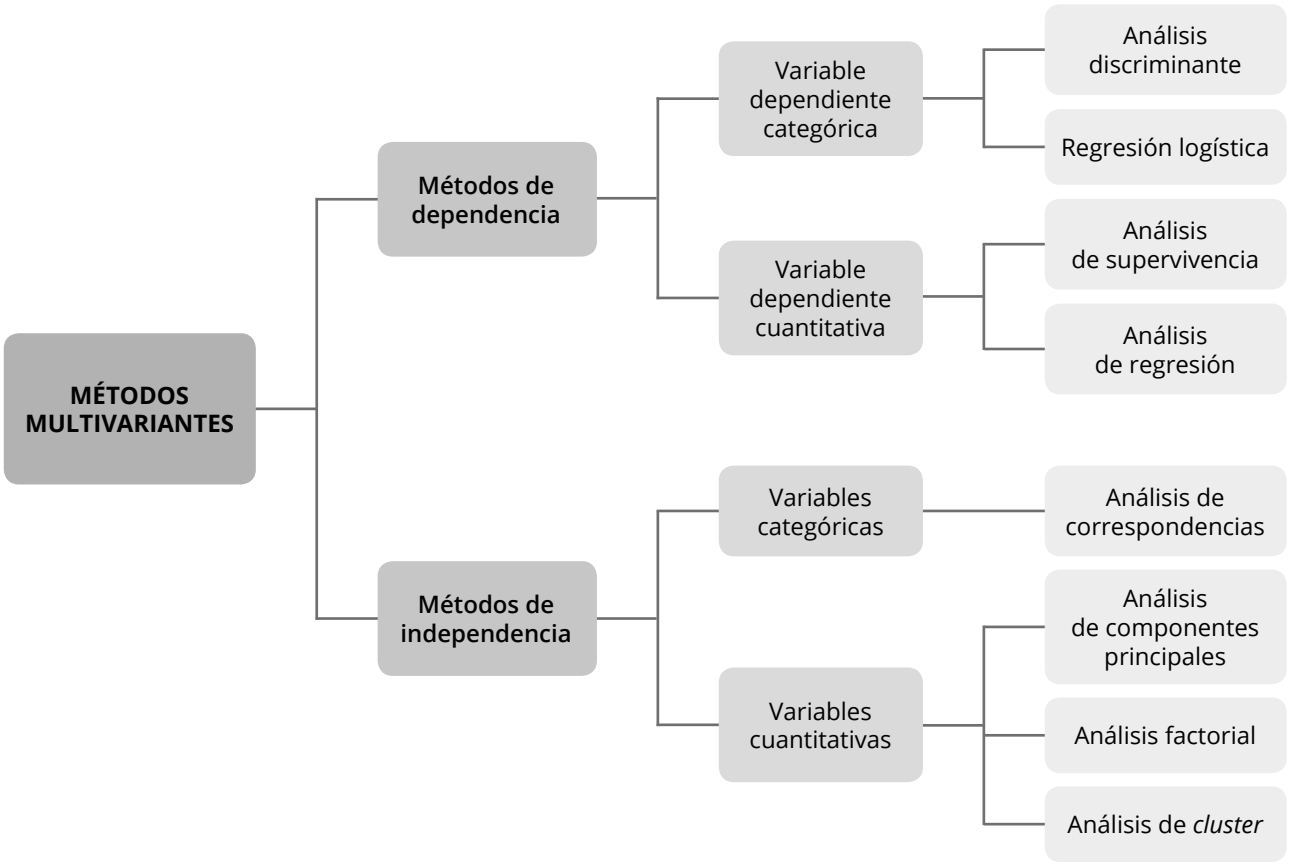
4. ESTIMADORES MULTIVARIABLES

4.1. INTRODUCCIÓN

Hemos visto cómo analizar una sola variable y relaciones entre dos variables. En este tema veremos qué hacer cuando tenemos más de dos variables, que suele ser el caso de cualquier estudio. La idea es encontrar la forma de analizar todos los datos conjuntamente. Para ello hay multitud de técnicas, y cuál usar dependerá de los datos que se tengan que analizar.

Vamos a clasificar los métodos multivariantes en dos:

- **Métodos de dependencia (análisis supervisado)**: disponemos de una (o más) variable dependiente, y el resto son variables independientes que queremos usar para explicar la variable dependiente.
- **Métodos de independencia (análisis no supervisado)**: no distinguimos entre variables dependientes e independientes, el objetivo es estudiar las relaciones entre todas las variables.



4.1.1. MÉTODOS DE DEPENDENCIA

Estos dependerán de si la variable dependiente es **categórica o cuantitativa**. Vamos a ver algún ejemplo de método para usar en cada caso.

VARIABLE DEPENDIENTE CATEGÓRICA

Análisis discriminante

Dados unos datos clasificados en varios grupos, este análisis busca encontrar **relaciones lineales** entre variables tales que discriminen o separen mejor los diferentes grupos. Este análisis forma parte de los denominados *métodos de estimación de relación*.

EJEMPLO

Tenemos datos sobre varias tiendas, estas están clasificadas entre rentables y no rentables. Queremos estudiar la relación entre las diversas variables que nos permita diferenciar entre una tienda rentable y otra que no.

Regresión logística

La hemos visto anteriormente en el "Tema 3. Estimadores bivariables". Se usa cuando queremos ver la **asociación** entre una **variable respuesta categórica** y una o más variables **explicativas**. Este análisis también forma parte de los "métodos de estimación de relación".

EJEMPLO

Queremos estudiar la relación entre tener cáncer de pulmón y una serie de variables categóricas sobre la persona: el sexo, la población donde vive, si es fumador o no, etc.

VARIABLE DEPENDIENTE CUANTITATIVA

Análisis de regresión

El concepto es el mismo que el de la regresión lineal visto anteriormente. Tenemos una (o más) **variable dependiente** que, en este caso, depende **de varias variables independientes**. Es otro caso de "método de estimación de relación".

EJEMPLO

Queremos estudiar el gasto mensual de una persona en restaurantes en función de su nivel de ingresos, sexo, edad, nivel educativo, población donde vive, etc.

Análisis de supervivencia

Similar al análisis de regresión, pero en este caso la variable dependiente es el **tiempo** de supervivencia (tiempo que dura una situación).

EJEMPLO

Queremos estudiar el tiempo que una persona permanece desempleada en función de su sexo, edad, nivel educativo, población donde vive, etc.

4.1.2. MÉTODOS DE INDEPENDENCIA

VARIABLES CATEGÓRICAS

Análisis de correspondencias

Son tablas de contingencia multidimensionales. Sirven para visualizar las observaciones en **dos dimensiones**. Estas tienen que ser lo más representativas posible, se trata de que cada eje aporte la máxima información posible.

EJEMPLO

Dada una lista de ciudades, queremos ver cuáles son más parecidas en función del número de habitantes, empresas, servicios, etc. Queremos encontrar dos ejes que representen las variables explicativas y visualizar las diferentes ciudades en el plano para saber cuáles son más cercanas entre ellas.

VARIABLES CUANTITATIVAS

Análisis de componentes principales

Este análisis forma parte de los denominados *métodos de reducción de dimensionalidad*. Se utiliza para estudiar las relaciones **reduciendo el número de variables**.

EJEMPLO

Supongamos que un economista quiere estudiar el estado financiero de una empresa y dispone para ello de una multitud de ratios financieras. El problema se resolvería mediante un análisis de componentes principales, construyendo a partir de las ratios varios índices numéricos que definan el estado financiero de la empresa.

Análisis factorial

Igual que el análisis de componentes principales, se usa para estudiar las relaciones reduciendo el número de variables. La diferencia es que en este caso las variables nuevas (factores) **no son observables**. También forma parte de los "métodos de reducción de dimensionalidad".

EJEMPLO

Supongamos que un psicólogo quiere determinar los factores que caracterizan la inteligencia de un individuo y dispone para ello de las respuestas a un test de inteligencia. El problema se resolvería mediante un análisis factorial.

Análisis de cluster

Este análisis **agrupa** la muestra en **grupos similares** (clusters) según las diferentes variables. Forma parte de los denominados *métodos de segmentación*.

EJEMPLO

Supongamos que queremos agrupar a los clientes de una tienda en función de su frecuencia de compra, su gasto, los productos que compran, etc.

Hemos visto brevemente que, aparte de si las variables son cuantitativas o cualitativas, estos análisis se pueden clasificar en métodos de estimadores de relación, de reducción de dimensionalidad, de segmentación, etc. Vamos a ver en detalle algunos de estos análisis.

4.2 ESTIMADORES DE RELACIÓN

Como su nombre indica, se trata de encontrar relaciones entre variables. El análisis más común de esta categoría es el **discriminante**.

4.2.1. ANÁLISIS DISCRIMINANTE

Supongamos que tenemos una variable dependiente categórica (mínimo dos categorías, $q \geq 2$) y diversas variables explicativas numéricas ($X_1,...,X_n$). La idea del análisis discriminante es **encontrar relaciones lineales** entre las variables explicativas tales que discriminen mejor entre los grupos de la variable dependiente. Estas relaciones lineales se denominan *funciones discriminantes* ($Y_1,...,Y_m$).

El **número máximo** de funciones discriminantes es $m = \min(q - 1, n)$.

Por lo tanto, queremos encontrar las m funciones lineales:

$$Y_i = \lambda_{i1} X_1 + ... + \lambda_{in} X_n$$

Los coeficientes λ_{ij} se calculan de manera que maximicen la varianza entre los grupos. Es decir, deben separar el máximo posible los q grupos.

SABÍAS QUE...

Varios programas tienen paquetes que calculan las funciones discriminantes, por ejemplo, en R se puede usar la libraría MASS.

Y_1 es la combinación lineal que proporciona una mayor discriminación e Y_m es la combinación lineal que proporciona una menor discriminación.

Las funciones discriminantes se pueden utilizar para asignar a qué categoría corresponderían observaciones nuevas. También se puede mirar a qué categoría correspondería una observación existente según las funciones discriminantes, así se puede estimar la probabilidad de error.

Lo que se hace es calcular la probabilidad de que la observación pertenezca a cada uno de los q grupos. Esta probabilidad condicionada se calcula mediante la **regla de Bayes**:

$$P(\text{Grupo i}|D) = \frac{P(D|\text{Grupo i})}{\sum_{i=1}^{q} P(D|\text{Grupo i})\, P(\text{Grupo i})}$$

Es decir, queremos calcular la probabilidad de que un objeto pertenezca al Grupo i (con $i = 1,\ldots,q$) dada una puntuación discriminante $D = (y_{j1},\ldots,y_{jm})$. Esta se puede calcular ya que disponemos del resto de elementos de la fórmula:

- $P(D|\text{Grupo } i)$ es la probabilidad de obtener la puntuación discriminante D estando en el Grupo i. Esta se puede calcular a partir de las funciones discriminantes.
- $P(\text{Grupo } i)$ es la probabilidad de pertenecer al Grupo i. Por lo general, se usa la frecuencia relativa de la categoría i en la muestra.

EJEMPLO

Supongamos que queremos diferenciar entre dos especies de árboles muy parecidas que se encuentran en un bosque ($q = 2$). Para ello, recogemos datos de 32 árboles, 16 de cada categoría; la información que recopilamos es:

- X_1 = Diámetro del tronco.
- X_2 = Altura del árbol.
- X_3 = Altura hasta la primera rama.
- X_4 = Longitud de las hojas.
- X_5 = Anchura de las hojas.

Supongamos que tenemos la siguiente función discriminante:

	Y_1
DIÁMETRO	0,05
ALTURA ÁRBOL	-0,08
ALTURA RAMA	0
LONGITUD	0,09
ANCHURA	0,09

Coeficientes discriminantes lineales.

En este caso solo tenemos una función discriminante porque $m = \min(q - 1, n) = \min(1, 5) = 1$. Si lo ponemos como función:

$$Y_1 = 0{,}05X_1 + \ldots + 0{,}09X_5$$

Supongamos ahora que se nos proporciona el diámetro, altura, etc., de dos árboles nuevos, y queremos estimar a cuál de las dos especies pertenecen.

Para ello, calculamos las probabilidades mediante la regla de Bayes y obtenemos:

	ESPECIE 1	ESPECIE 2
ÁRBOL 1	0,75	0,25
ÁRBOL 2	0,17	0,83

Probabilidades de pertenecer a la especie 1 y 2.

Por lo tanto, estimamos que el árbol 1 pertenece a la especie 1, mientras que el árbol 2 pertenece a la especie 2.

También podemos estimar la especie con los 32 árboles de la muestra inicial, así nos haremos una idea de lo buenas que son las predicciones. Supongamos que estos son los resultados:

		REAL	
		ESPECIE 1	ESPECIE 2
PREDICCIÓN	ESPECIE 1	14	3
	ESPECIE 2	3	12

Probabilidades de pertenecer a la especie 1 y 2.

La predicción es bastante buena, acierta en 26/32 casos.

4.3. REDUCCIÓN DE DIMENSIONALIDAD

Es bastante común en un estudio recoger mucha información, aunque luego sea muy difícil analizar los datos y nos encontremos con que alguna variable no aporta información. Es por esta razón que muchas veces puede ser útil reducir la dimensionalidad, es decir, **reducir el número de variables**. Para ello, se usan las relaciones entre variables para crear variables nuevas. Estas se denominan ***factores*** (si son inobservables) o ***componentes principales*** (si son observables).

4.3.1. ANÁLISIS DE COMPONENTES PRINCIPALES

Como acabamos de comentar, cuando recogemos una muestra de datos para un estudio es muy común recoger tantas variables como sea posible. Esto hace muy complicado poder estudiar todas las relaciones entre ellas.

Sin embargo, es muy probable que parte de las variables recogidas estén **correlacionadas**, es decir, que de manera indirecta aporten la misma información.

El análisis de componentes principales lo que hace es coger todas las n variables correlacionadas y **transformarlas** en otro conjunto de m **nuevas variables incorreladas** ($m < n$), es decir, que no tengan repetición o redundancia entre sí, estas nuevas variables recogen la mayor parte de la información o variabilidad de los datos.

Las nuevas variables son combinaciones lineales de las anteriores. Supongamos que tenemos n variables $X_1,\ldots,X_n$, queremos crear un conjunto nuevo $Y_1,\ldots,Y_m$ con $m < n$ tal que:

$$Y_j = \lambda_{j1} X_1 + \ldots + \lambda_{jn} X_n$$

Estas se van construyendo según el orden de importancia en cuanto a la variabilidad total que recogen de la muestra. Es decir, queremos **maximizar la varianza**. Así, Y_1 será el componente que aporte mayor varianza e Y_m el que menos varianza aporte.

Para calcular los λ_{ij} necesitamos la **matriz de correlaciones**. Denominamos esta matriz:

$$A = \begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \ldots & a_{nn} \end{pmatrix}$$

Esta matriz representa la correlación entre todos los pares de variables. Es decir, a_{ij} es la correlación entre la variable X_i y la variable X_j. Se cumple que:

- Los valores $|a_{ij}|$ están comprendidos entre [0,1], siendo 0 correlación nula y 1 correlación total.
- Es una matriz simétrica, es decir, $a_{ij} = a_{ji}$, ya que la correlación entre la variable i y la variable j es la misma que la correlación entre la variable j y la variable i.
- Su diagonal son unos ($a_{ii} = 1$, $i = 1,\ldots,n$), ya que la correlación de una variable consigo misma es 1.

Con esta matriz de correlaciones podemos calcular los componentes principales. Los componentes de la matriz de componentes principales (λ_{ij}) son los **autovectores** de la matriz de correlaciones.

De cada componente se puede calcular la proporción de varianza que aporta (respecto a las variables originales). Esto nos sirve para ordenar los componentes de mayor a menor varianza y elegir cuántos vamos a usar para el análisis.

EJEMPLO

Supongamos que estamos estudiando la esperanza de vida de los árboles en una ciudad. Para ello recogemos información sobre la ciudad:

- X_1 = temperatura anual.
- X_2 = nivel de contaminación del aire.
- X_3 = población.
- X_4 = velocidad media del viento.
- X_5 = precipitación anual media.
- X_6 = días lluviosos al año.

Supongamos también que tenemos la siguiente matriz de correlaciones entre las variables:

	TEMPERATURA	CONTAMINACIÓN	POBLACIÓN	VIENTO	PRECIPITACIONES	DÍAS
TEMPERATURA	1	-0,19	-0,06	-0,35	0,39	-0,43
CONTAMINACIÓN	-0,19	1	0,96	0,24	-0,03	0,13
POBLACIÓN	-0,06	0,96	1	0,21	-0,03	0,04
VIENTO	-0,35	0,24	0,21	1	-0,01	0,16
PRECIPITACIONES	0,39	-0,03	-0,03	-0,01	1	0,5
DÍAS	-0,43	0,13	0,04	0,16	0,5	1

Matriz de correlaciones.

Podemos ver que las variables más correlacionadas entre ellas son la población y la contaminación, y las menos correlacionadas el viento y las precipitaciones.

Si calculamos los autovectores de esta matriz, obtenemos los componentes principales:

	COMP. 1 (Y_1)	COMP. 2 (Y_2)	COMP. 3 (Y_3)	COMP. 4 (Y_4)	COMP. 5 (Y_5)	COMP. 6 (Y_6)
TEMPERATURA	0,33	-0,13	0,67	-0,31	-0,56	-0,14
CONTAMINACIÓN	-0,61	-0,17	0,27	0,14	0,1	-0,7
POBLACIÓN	-0,58	-0,22	0,35	0	0	0,7
VIENTO	-0,35	0,13	-0,3	-0,87	-0,11	0
PRECIPITACIONES	0	0,62	0,51	-0,17	0,57	0
DÍAS	-0,24	0,71	0	0,31	-0,58	0

Matriz de componentes principales.

Así, por ejemplo:

$$Y_1 = 0{,}33\ X_1 - 0{,}61\ X_2 - 0{,}58\ X_3 - 0{,}35\ X_4 - 0{,}24\ X_6$$

Este primer componente y_1 se podría etiquetar como "calidad del aire", ya que está muy relacionado negativamente con la contaminación (-0,61) y el tamaño de la población (-0,58).

Podríamos calificar el segundo componente y_2 como "tiempo húmedo", ya que está muy relacionado positivamente con el nivel de precipitaciones (0,62) y el número de días lluviosos (0,71).

Y, como último ejemplo, el tercer componente y_3 se podría etiquetar como "clima", ya que está muy relacionado positivamente con la temperatura (0,67) y las precipitaciones (0,51).

No siempre se puede etiquetar un componente, pero sigue siendo interesante ver si existe una fuerte relación entre variables.

Vamos a ver ahora la proporción de varianza que aporta cada componente, así decidiremos cuántos son necesarios.

	COMP. 1 (Y_1)	COMP. 2 (Y_2)	COMP. 3 (Y_3)	COMP. 4 (Y_4)	COMP. 5 (Y_5)	COMP. 6 (Y_6)
PROPORCIÓN DE VARIANZA	0,37	0,25	0,23	0,13	0,02	0,01
PROPORCIÓN ACUMULADA	0,37	0,62	0,85	0,98	0,99	1

Proporción de varianza de los componentes principales.

Los tres primeros componentes aportan el 85 % de la varianza de las variables originales y a partir de aquí el incremento es mínimo, así que nos quedaremos con tres componentes.

Esto significa que podemos reducir la dimensionalidad de $n = 6$ a $m = 3$ variables manteniendo el 85 % de la variabilidad que aportaban las n variables iniciales.

4.3.2. ANÁLISIS FACTORIAL

Cuando estamos estudiando las relaciones entre variables, no siempre disponemos de todas las variables explicativas de las que estas dependen.

El análisis factorial es un modelo que nos permite **relacionar** las **variables observadas** (variables explicativas que sí que tenemos) con los factores o **variables no observadas** (variables explicativas que no tenemos). Las variables iniciales tienen que estar correlacionadas, sino el análisis no tiene ningún sentido.

SABÍAS QUE...

El análisis factorial tiene muchos puntos en común con el de componentes principales, y busca esencialmente nuevas variables o factores que expliquen los datos.

El análisis factorial puede ser:

- **Exploratorio**: no se conoce *a priori* el número de factores y es en la aplicación empírica donde se determina este número.
- **Confirmatorio**: los factores están fijados a priori, utilizándose contrastes de hipótesis para su corroboración.

Sean $X_1,...,X_n$ las variables iniciales observadas y $F_1,...,F_m$ los factores o variables no observadas (con $m < n$). Estas se relacionan mediante:

$$X_i = \lambda_{i1} F_1 + ... + \lambda_{im} F_m$$

Donde λ_{ij} es el peso del factor F_j.

Nuestro objetivo es encontrar λ_{ij} y, en el caso del análisis factorial exploratorio, el número de factores m.

En el análisis factorial hay varios métodos para calcular λ_{ij}. Los más comunes son el **método de las componentes principales**, que ya hemos visto en el apartado análisis de componentes principales, y el **método de máxima verosimilitud**, que está basado en el anterior (cálculo de la matriz de correlaciones, etc.). Dado que anteriormente ya hemos estudiado este método, no entraremos en detalle en este apartado.

Otros métodos que se pueden usar son: método de los ejes principales, método de mínimos cuadrados, etc.

Vamos a ver cómo conocer el número de factores en el caso del análisis factorial exploratorio. La forma más común es mediante un **test de hipótesis**. Este criterio se puede calcular si el método utilizado para estimar los factores es el de máxima verosimilitud.

Se comienza con un valor pequeño (normalmente $m = 1$) y calculamos el estadístico del test (*p-value*):

- Si **no es significativo**, aceptamos m. Consideraremos que el estadístico no es significativo cuando es $p \geq 0{,}05$.

- Si **es significativo** ($p < 0{,}05$), aumentamos m en una unidad y repetimos el proceso hasta que el estadístico no sea significativo.

EJEMPLO

Queremos estudiar los años de vida estimados de diferentes países según el sexo y la edad de la persona. Es decir, dada una persona aleatoria de un país, si conocemos su sexo y su edad, queremos ser capaces de saber cuántos años vivirá. Por ello recogemos datos sobre la esperanza de vida dependiendo del sexo (hombre, mujer) y de la edad (0-25, 25-50, 50-75, > 75). Así tenemos $n = 8$ variables observadas:

- X_1 = h0
- X_2 = h25
- X_3 = h50
- X_4 = h75
- X_5 = m0
- X_6 = m25
- X_7 = m50
- X_8 = m75

Queremos encontrar el número de factores $m < n = 8$ tales que nos aporten la misma información que las variables observadas y nos permitan entender mejor los años de vida estimados.

Hacemos el test de hipótesis con un solo factor ($m = 1$).

	FACTOR 1 (F_1)
h0	0,87
h25	0,73
h50	0,78
h75	0,55
m0	0,89
m25	0,99
m50	0,94
m75	0,68

p-value
1,88e-24

Análisis factorial con 1 factor.

El test de hipótesis es significativo, con lo cual vamos a repetir el test con dos factores (m = 2).

	FACTOR 1 (F_1)	FACTOR 2 (F_2)
h0	0,97	0,18
h25	0,67	0,33
h50	0,48	0,65
h75	0,12	0,76
m0	0,97	0,19
m25	0,79	0,6
m50	0,57	0,82
m75	0,19	0,89

p-value
1,91e-5

Análisis factorial con 2 factores.

El test de hipótesis es significativo, con lo cual vamos a repetir el test con tres factores (m = 3).

	FACTOR 1 (F_1)	FACTOR 2 (F_2)	FACTOR 3 (F_3)
h0	0,96	0,12	0,23
h25	0,65	0,17	0,44
h50	0,43	0,35	0,79
h75	0	0,53	0,66
m0	0,97	0,22	0
m25	0,76	0,56	0,31
m50	0,54	0,73	0,4
m75	0,16	0,87	0,28

p-value
0,46

Análisis factorial con 3 factores.

El test de hipótesis no es significativo, así que aceptamos m = 3 como el número de factores óptimos.

Si analizamos los resultados, podríamos etiquetar el primer factor como “niños”, ya que está muy relacionado con la esperanza de vida en el nacimiento (*h0* = 0,96, *m0* = 0,97).

El segundo factor se relaciona con la “vejez”, ya que los coeficientes más altos son los correspondientes a la edad de > 75 años (*h75* = 0,53, *m75* = 0,87).

Finalmente, el tercer factor tiene los pesos factoriales más altos en las esperanzas de vida de hombres entre 50 y 75 años (*h50* = 0,79, *h75* = 0,66), por lo que podríamos etiquetarlo como "hombres mayores".

De esto concluimos que los años de vida estimados se pueden explicar conociendo la esperanza de vida de los niños, la vejez y los hombres mayores. Hemos pasado de 8 variables observadas a solo 3 factores.

4.4. SEGMENTACIÓN

Cuando analizamos muchas variables, hay veces que lo que necesitamos es **agrupar** las observaciones según **similitudes entre estas variables**.

EJEMPLO

Supongamos que disponemos de una tabla con todos los coches que hay en el mercado a día de hoy, y disponemos de muchas variables con sus características: tamaño, color, acabados, motor, potencia, marca, etc. Supongamos que queremos saber qué coches son parecidos dadas todas estas variables. Para este tipo de problemas se usa la segmentación.

Son muchos los tipos de segmentación; la más común es la de **clusters**. También son populares los árboles de decisión y las segmentaciones de clientes (muy usadas en *marketing* para saber a qué clientes impactar con una campaña publicitaria).

SABÍAS QUE...

A día de hoy se usa la segmentación para infinidad de cosas:

- En astronomía se hacen clusters para definir galaxias.
- En *retail* se segmentan los clientes en *premium*, potenciales, etc.
- En ciencias ambientales, para clasificar ríos según la calidad de las aguas.
- Etc.

Dado que no podemos estudiar todos los tipos de segmentación existentes, nos centraremos en la más popular: los clusters.

4.4.1. CLUSTERS

Dadas unas observaciones, el análisis de clusters busca similitudes entre ellas y las agrupa en diversos clusters o grupos.

Siempre habrá diversos criterios de agrupación; cuál elegir dependerá de lo que nos interese. Supongamos que queremos clasificar los perros de una protectora. Podemos agruparlos por tamaño, por color, por edad, etc.

El análisis de clusters se divide en dos métodos: **jerárquicos** y **no jerárquicos**.

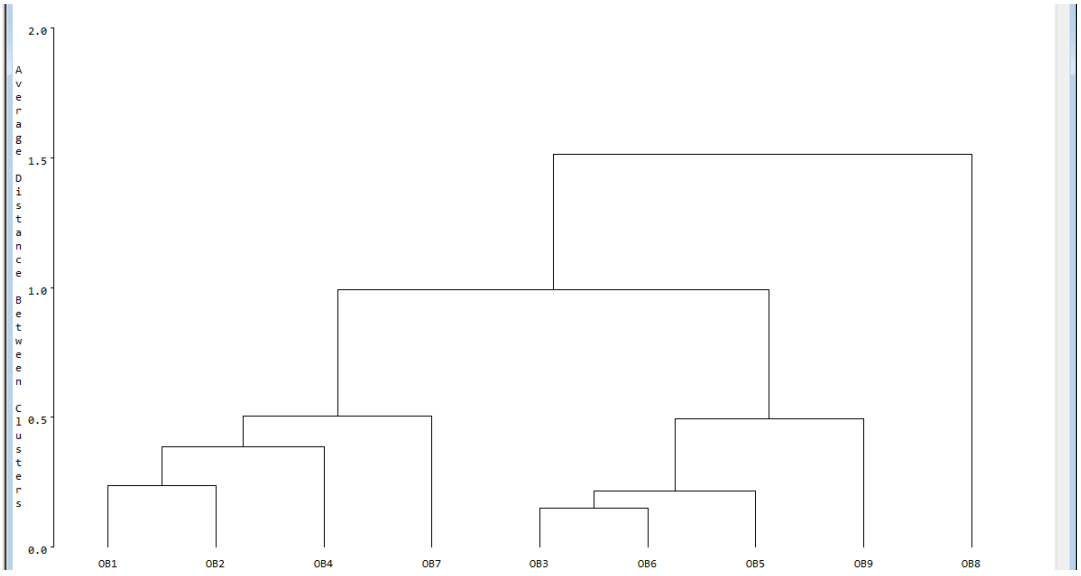
MÉTODOS JERÁRQUICOS

En cada paso del algoritmo solo se tiene en cuenta **una observación**. Una vez una observación ha sido asignada a un grupo, esta ya no cambia más de grupo.

Existen dos tipos de métodos jerárquicos:

- **Aglomerativos**: partimos de todas las observaciones, calculamos las dos más similares (según el criterio elegido) y consideramos la combinación de las dos como una sola observación. Repetimos el proceso hasta que acabemos con una sola observación.
- **Divisivos**: partimos de un único grupo con todas las observaciones, dividimos en dos grupos de tal manera que los dos grupos estén lo más alejados posible. Repetimos el proceso hasta que haya tantos grupos como observaciones.

Independientemente del método utilizado, en ambos se acaba generando un **dendograma**, que es un gráfico en forma de árbol donde se ve cómo se juntan las observaciones.



Ejemplo dendograma.

Para generar el dendograma primero tenemos que decidir qué **criterio de agrupación** usar, es decir, necesitamos poder medir cuándo dos observaciones son cercanas y cuándo son lejanas.

Ejemplo

Para este ejemplo usaremos un método jerárquico aglomerativo con la mínima distancia entre dos puntos.

Supongamos que tenemos 5 puntos y calculamos las distancias entre ellos:

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{ccccc} 0 & 2 & 3 & 4 & 5 \end{array} \\ \begin{pmatrix} 0 & & & & \\ 8 & 0 & & & \\ 2 & 9 & 0 & & \\ 6 & 5 & 8 & 0 & \\ 9 & 7 & \boxed{1} & 7 & 0 \end{pmatrix} \end{array}$$

Al ser un método jerárquico aglomerativo partimos de 5 clusters, e iremos juntando observaciones hasta tener un único cluster.

La matriz de distancias nos indica que las dos observaciones más próximas son la 3 y la 5. Así pues, creamos el primer cluster (3 + 5) y volvemos a calcular la matriz de distancias.

Las distancias entre los puntos 1, 2 y 4 no han cambiado, por lo que solo tenemos que calcular la distancia de estos con el punto "nuevo".

$$d_{35,1} = \min (d_{3,1}, d_{5,1}) = \min (2,9) = 2$$
$$d_{35,2} = \min (d_{3,2}, d_{5,2}) = \min (9,7) = 7$$
$$d_{35,4} = \min (d_{3,4}, d_{5,4}) = \min (8,7) = 7$$

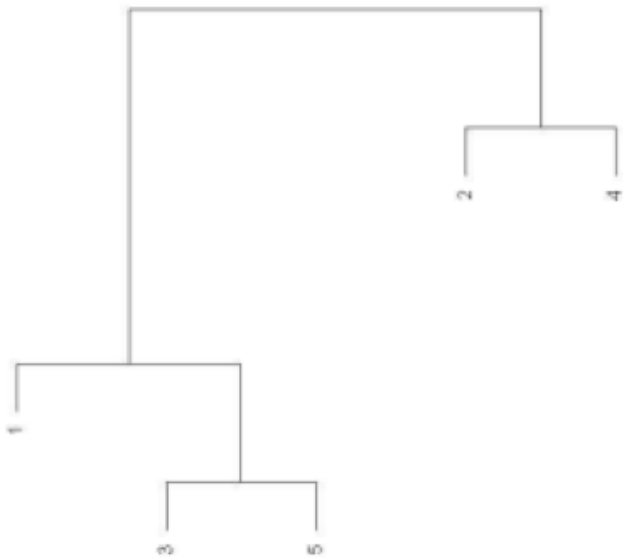
$$\begin{array}{c} \\ (35) \\ 1 \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (35) & 1 & 2 & 4 \end{array} \\ \begin{pmatrix} 0 & & & \\ \boxed{2} & 0 & & \\ 7 & 8 & 0 & \\ 7 & 6 & 5 & 0 \end{pmatrix} \end{array}$$

Las dos observaciones más próximas son la 3 + 5 y la 1. Usando el mismo procedimiento volvemos a calcular la matriz de distancias:

$$\begin{array}{c} \\ (351) \\ 2 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} (351) & 2 & 4 \end{array} \\ \begin{pmatrix} 0 & & \\ 7 & 0 & \\ 6 & \boxed{5} & 0 \end{pmatrix} \end{array}$$

El siguiente cluster es el que forman los puntos 2 y 4.

Ya nos hemos quedado con solo dos clusters, ahora ya podemos dibujar el dendograma:



Dendograma.

MÉTODOS NO JERÁRQUICOS

Comienzan con una solución inicial, un número *k* de grupos o clusters fijado de antemano y agrupa los objetos para obtener estos *k* clusters.

SABÍAS QUE...

La ventaja es que no se tiene que especificar la matriz de distancias ni almacenar las iteraciones, hecho que lo hace mucho más eficaz en cuanto a programación, ya que requiere mucha menos memoria que los métodos jerárquicos.

La elección inicial de *k* se hace al azar (a no ser que se disponga de información previa) y se va ajustando en función de los resultados obtenidos. Una vez fijado *k*, los elementos de cada cluster van cambiando a cada iteración.

El método no jerárquico más común es el de **k-medias** (*k-means*). Este método parte de *k centroides* (uno para cada cluster) y asigna cada observación a su centroide más cercano, normalmente según la distancia euclidiana. Luego se vuelven a recalcular los centroides con todos los puntos de cada cluster y se vuelven a reasignar todas las observaciones a su centroide más próximo.

Este proceso se repite hasta que los clusters se estabilizan. Veamos un ejemplo.

EJEMPLO

Supongamos que tenemos dos variables x_1, x_2 y cuatro observaciones. Y decidimos que queremos tener $k = 2$ clusters usando la distancia euclidiana.

	x_1	x_2
1	6	2
2	-2	0
3	3	-1
4	-3	-3

Empezamos de forma arbitraria haciendo los clusters (1-2) y (3-4). Ahora tenemos que calcular los centroides de estos clusters:

CLUSTER (1-2)	Centroide $x_1 = \frac{6 + (-2)}{2} = 2$
	Centroide $x_2 = \frac{2 + 0}{2} = 1$
CLUSTER (3-4)	Centroide $x_1 = \frac{3 + (-3)}{2} = 0$
	Centroide $x_2 = \frac{-1 + (-3)}{2} = -2$

Ahora calculamos la distancia euclidiana de cada observación a cada centroide:

$$d_{1,12} = (6 - 2)^2 + (2 - 1)^2 = 17$$
$$d_{1,34} = (6 - 0)^2 + (2 - (-2))^2 = 52$$

$$d_{2,12} = ((-2) - 2)^2 + (0 - 1)^2 = 17$$
$$d_{2,34} = ((-2) - 0)^2 + (0 - (-2))^2 = 8$$

$$d_{3,12} = (3 - 2)^2 + ((-1) - 1)^2 = 5$$
$$d_{3,34} = (3 - 0)^2 + ((-1) - (-2))^2 = 10$$

$$d_{4,12} = ((-3) - 2)^2 + ((-3) - 1)^2 = 41$$
$$d_{4,34} = ((-3) - 0)^2 + ((-3) - (-2))^2 = 10$$

Las observaciones 1 y 3 están más cerca del centroide (1-2), mientras que las observaciones 2 y 4 están más cerca del centroide (3-4).

Cambiamos pues los clusters a (1-3) y (2-4) y repetimos los cálculos. Volvemos a calcular los centroides:

CLUSTER (1-3)	Centroide $x_1 = \frac{6 + 3}{2} = 4,5$
	Centroide $x_2 = \frac{2 + (-1)}{2} = 0,5$
CLUSTER (2-4)	Centroide $x_1 = \frac{(-2) + (-3)}{2} = -2,5$
	Centroide $x_2 = \frac{0 + (-3)}{2} = -1,5$

Y la distancia de cada observación a cada centroide:

$$d_{1,13} = (6 - 4,5)^2 + (2 - 0,5)^2 = 4,5$$
$$d_{1,24} = (6 - (-2,5))^2 + (2 - (-1,5))^2 = 84,5$$

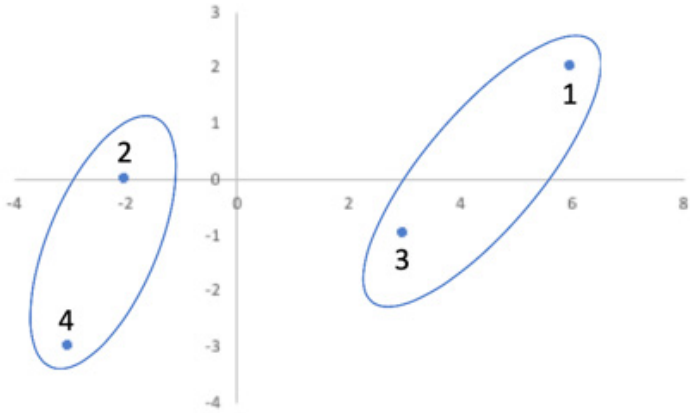
$$d_{2,13} = ((-2) - 4,5)^2 + (0 - 0,5)^2 = 42,5$$
$$d_{2,24} = ((-2) - (-2,5))^2 + (0 - (-1,5))^2 = 2,5$$

$$d_{3,13} = (3 - 4,5)^2 + ((-1) - 0,5)^2 = 4,5$$
$$d_{3,24} = (3 - (-2,5))^2 + ((-1) - (-1,5))^2 = 30,5$$

$$d_{4,13} = ((-3) - 4,5)^2 + ((-3) - 0,5)^2 = 68,5$$
$$d_{4,24} = ((-3) - (-2,5))^2 + ((-3) - (-1,5))^2 = 2,5$$

Las observaciones 1 y 3 están más cerca del centroide (1-3), mientras que las observaciones 2 y 4 están más cerca del centroide (2-4).

Ya hemos llegado a un equilibrio, por lo tanto, dadas las observaciones iniciales, con el método no jerárquico de k-medias con *k* = 2, los dos clusters resultantes son (1-3) y (2-4). Si lo graficamos, queda muy claro:



Clusters con k-medias.

4.5. PREVISIÓN

Cuando recogemos datos para una muestra podemos tener una **imagen fija del momento**, por ejemplo, si estudiamos la diversidad cultural que hay en los colegios a día de hoy, o podemos tener una **evolución temporal**, por ejemplo, el precio de las viviendas a lo largo de los años.

Es común encontrarnos con observaciones de evolución temporal. Cómo tratar este tipo de datos es un campo muy extenso, del cual solo vamos a dar algunas nociones básicas.

4.5.1. ANÁLISIS DE SERIES TEMPORALES

Una serie temporal es una secuencia de observaciones medidas en intervalos de tiempo (típicamente equidistantes).

EJEMPLO

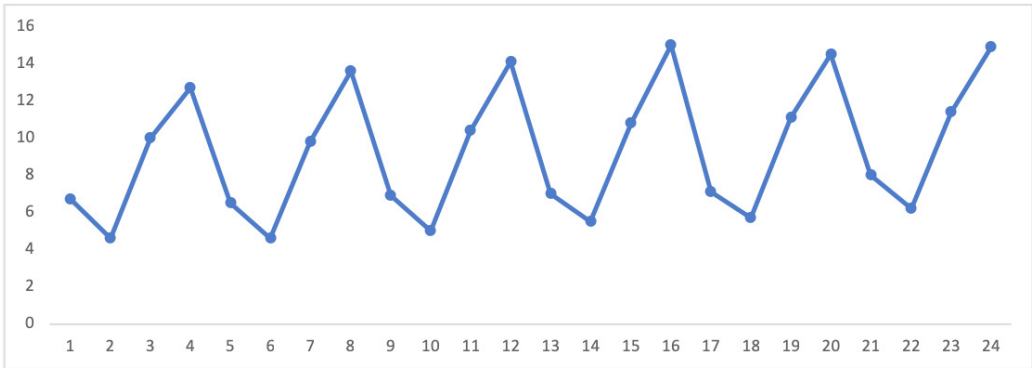
Las ventas de juguetes en una ciudad.

AÑO	ESTACIÓN	JUGUETES (EN MILLONES)
2013	Invierno	6,7
2013	Primavera	4,6
2013	Verano	10
2013	Otoño	12,7
2014	Invierno	6,5
2014	Primavera	4,6
2014	Verano	9,8
2014	Otoño	13,6
2015	Invierno	6,9
2015	Primavera	5
2015	Verano	10,4
2015	Otoño	14,1
2016	Invierno	7
2016	Primavera	5,5
2016	Verano	10,8
2016	Otoño	15
2017	Invierno	7,1

2017	Primavera	5,7
2017	Verano	11,1
2017	Otoño	14,5
2018	Invierno	8
2018	Primavera	6,2
2018	Verano	11,4
2018	Otoño	14,9

Serie temporal: ventas de juguetes.

Si la graficamos, observamos la forma típica de una serie temporal:

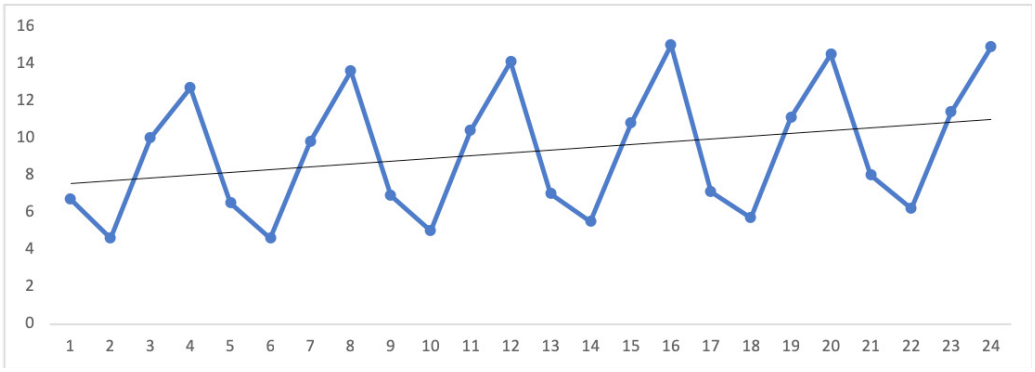


Serie temporal: ventas de juguetes.

En una serie temporal hay cuatro componentes que explican su progresión. Con estos cuatro elementos se puede recrear la serie temporal.

TENDENCIA

Es lo que hace la serie temporal a la larga. En nuestro ejemplo crece linealmente.



Tendencia de la serie temporal: ventas de juguetes.

Vamos a calcular la línea de la tendencia. Si no fuera lineal, primero transformaríamos los datos (típicamente con el logaritmo) para que fuera lineal.

Queremos encontrar la recta $y = ax + b$ tal que la distancia a los puntos sea mínima. Podemos pensar x como el instante de tiempo, así la observación 1 será el instante de tiempo 1: $t_1 = x_1 = 1$. Las fórmulas para encontrar a y b son:

$$a = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$

En nuestro ejemplo:

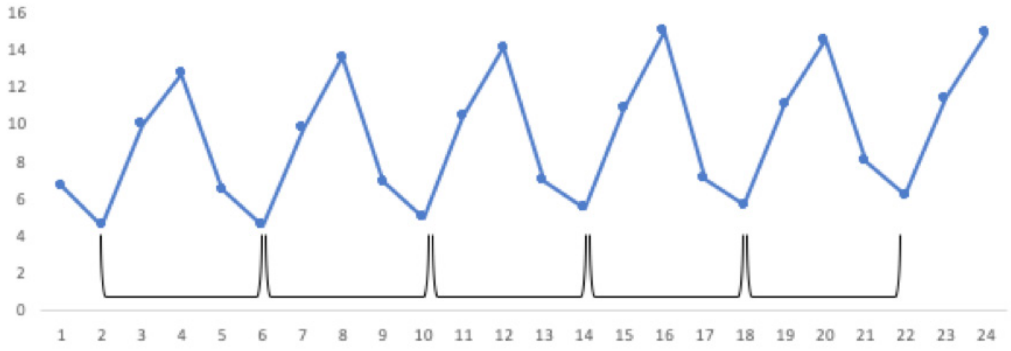
$$a = \frac{(1 - 12{,}5)(6{,}7 - 9{,}3) + \ldots + (24 - 12{,}5)(14{,}9 - 9{,}3)}{(1 - 12{,}5)^2 + \ldots + (24 - 12{,}5)^2} = 0{,}15$$

$$b = 9{,}3 - 0{,}15 \cdot 12{,}5 = 7{,}4$$

Así que la recta que representa la tendencia es $y = 0{,}15x + 7{,}4$. Si te fijas, esta es la recta de la imagen "Tendencia de la serie temporal: ventas de juguetes".

ESTACIONALIDAD

Son las fluctuaciones periódicas. En nuestros ejemplos anuales, cada año se repite el mismo patrón.



Estacionalidad de la serie temporal: ventas de juguetes.

Para capturar la estacionalidad usaremos una **media móvil** (*moving average*). Esta consiste en transformar los datos de tal manera que las observaciones "nuevas" son la media de las antiguas. La media se hace con la longitud del periodo de fluctuación, este se denomina ***orden***. En nuestro caso, como la fluctuación es anual y tenemos cuatro observaciones por año, el orden será cuatro.

La media móvil puede ser centrada o no; si:

- **Es centrada**, en cada observación se usan observaciones anteriores y posteriores a partes iguales.
- **No es centrada**, en cada observación se usan solo observaciones anteriores.

Típicamente se utiliza la centrada, que se calcula usando la siguiente fórmula.

Vamos a ver la fórmula por una media móvil centrada de orden 4 en el instante *t* (MM_c^4), pero esta es extrapolable a cualquier orden.

$$MM_c^4 = \frac{1}{4}\sum_{i=-2}^{2} \alpha_i \, y_{t-i} \quad \alpha_2 = \alpha_{-2} = 0{,}5; \qquad \alpha_i = 1 \qquad i = \{-1,0,1\}$$

Observa que, como depende de dos observaciones anteriores, podemos empezar a calcularla a partir de la tercera observación. En nuestro caso, la primera media móvil sería:

$$MM_c^4 = \frac{1}{4}(0{,}5 \cdot 6{,}7 + 4{,}6 + 10 + 12{,}7 + 0{,}5 \cdot 6{,}5) = 8{,}5$$

Una vez tenemos las medias móviles, calculamos la diferencia con las observaciones originales:

AÑO	ESTACIÓN	JUGUETES (EN MILLONES)	MM_c^4	DIFERENCIA
2013	Invierno	6,7		
2013	Primavera	4,6		
2013	Verano	10	8,48	1,52
2013	Otoño	12,7	8,45	4,25
2014	Invierno	6,5	8,43	-1,92
2014	Primavera	4,6	8,51	-3,91
2014	Verano	9,8	8,68	1,12
2014	Otoño	13,6	8,78	4,82
2015	Invierno	6,9	8,9	-2
2015	Primavera	5	9,04	-4,04
2015	Verano	10,4	9,11	1,29
2015	Otoño	14,1	9,19	4,91
2016	Invierno	7	9,3	-2,3

2016	Primavera	5,5	9,46	-3,96
2016	Verano	10,8	9,59	1,21
2016	Otoño	15	9,63	5,37
2017	Invierno	7,1	9,69	-2,58
2017	Primavera	5,7	9,66	-3,96
2017	Verano	11,1	9,71	1,38
2017	Otoño	14,5	9,89	4,61
2018	Invierno	8	9,99	-1,99
2018	Primavera	6,2	10,08	-3,88
2018	Verano	11,4		
2018	Otoño	14,9		

Media móvil en serie temporal: ventas de juguetes.

Ahora ya podemos calcular la media para cada elemento del periodo. Es decir, la media en invierno, primavera, verano y otoño.

$$\text{invierno} = \frac{(-1{,}92) + \ldots + (-1{,}99)}{5} = -2{,}16$$

$$\text{primavera} = \frac{(-3{,}91) + \ldots + (-3{,}88)}{5} = -3{,}95$$

$$\text{verano} = \frac{1{,}52 + \ldots + 1{,}38}{5} = 1{,}31$$

$$\text{otoño} = \frac{4{,}25 + \ldots + 4{,}61}{5} = 4{,}8$$

Esta es la estacionalidad que buscábamos.

CICLOS

Fluctuaciones no periódicas.

En nuestro ejemplo no hay, pero, si tuviéramos los datos de muchos más años, seguramente veríamos ciclos debidos a la economía. Seguiría habiendo una tendencia creciente, pero durante unos años crecería menos o hasta decrecería, mientras que en otros crecería más rápido.

IRREGULARIDADES

Representa irregularidades *random* que no se pueden describir mediante los otros componentes. Son los **residuos**.

Como hemos comentado antes, una serie temporal se puede recrear mediante estos cuatro elementos. En nuestro ejemplo, como no tenemos un ciclo, se compondrá de:

$$y_t = \text{tendencia} + \text{estacionalidad} + \text{irregularidades}$$

$$y_t = (0{,}15t + 7{,}4) + S_t + I_t$$

Donde $S_t = \{-2{,}16; -3{,}95; 1{,}31; 4{,}8\}$, dependiendo de si la observación en el instante *t* corresponde a invierno, primavera, verano u otoño.

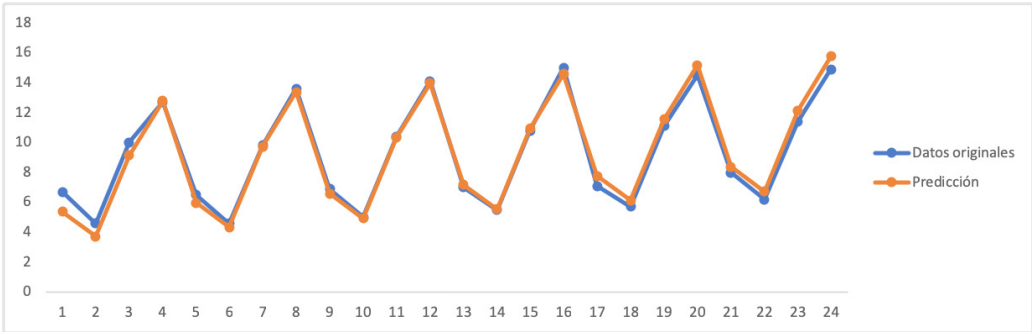
Retomando la tabla, tendremos:

AÑO	ESTACIÓN	JUGUETES (EN MILLONES)	ESTACIONALIDAD	TENDENCIA
2013	Invierno	6,7	-2,16	7,53
2013	Primavera	4,6	-3,95	7,68
2013	Verano	10	1,31	7,83
2013	Otoño	12,7	4,8	7,98
2014	Invierno	6,5	-2,16	8,13
2014	Primavera	4,6	-3,95	8,28
2014	Verano	9,8	1,31	8,43
2014	Otoño	13,6	4,8	8,58
2015	Invierno	6,9	-2,16	8,73
2015	Primavera	5	-3,95	8,88
2015	Verano	10,4	1,31	9,03
2015	Otoño	14,1	4,8	9,18
2016	Invierno	7	-2,16	9,33
2016	Primavera	5,5	-3,95	9,48
2016	Verano	10,8	1,31	9,63
2016	Otoño	15	4,8	9,78
2017	Invierno	7,1	-2,16	9,93
2017	Primavera	5,7	-3,95	10,08

2017	Verano	11,1	1,31	10,23
2017	Otoño	14,5	4,8	10,38
2018	Invierno	8	-2,16	10,53
2018	Primavera	6,2	-3,95	10,68
2018	Verano	11,4	1,31	10,83
2018	Otoño	14,9	4,8	10,97

Estacionalidad y tendencia en serie temporal: ventas de juguetes.

Si comparamos la serie original con la estimada (estacionalidad + tendencia), tenemos una aproximación muy buena:



Estimación de la serie temporal – ventas de juguetes.

RECUERDA

Las irregularidades son las diferencias que hacen que la predicción no sea exacta, es decir, los residuos.

ANEXO

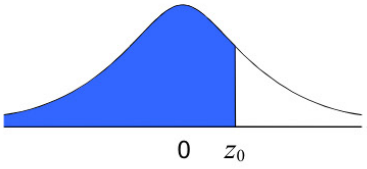
Probabilidad acumulada inferior para distribución normal N(0,1)

www.vaxasoftware.com

μ = Media

σ = Desviación típica

$$P(z \le z_0) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_0} e^{\frac{-z^2}{2}} dz$$



Tipificación: $z_0 = \frac{x - \mu}{\sigma}$

z_0	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	z_0
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359	**0,0**
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753	**0,1**
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141	**0,2**
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517	**0,3**
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879	**0,4**
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224	**0,5**
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549	**0,6**
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852	**0,7**
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133	**0,8**
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389	**0,9**
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621	**1,0**
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830	**1,1**
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015	**1,2**
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177	**1,3**
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319	**1,4**
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441	**1,5**
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545	**1,6**
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633	**1,7**
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706	**1,8**
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767	**1,9**
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817	**2,0**
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857	**2,1**
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890	**2,2**
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916	**2,3**
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936	**2,4**
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952	**2,5**
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964	**2,6**
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974	**2,7**
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981	**2,8**
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986	**2,9**
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900	**3,0**
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929	**3,1**
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950	**3,2**
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965	**3,3**
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976	**3,4**
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983	**3,5**
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989	**3,6**
3,7	0,99989	0,99990	0,99990	0,99990	0,99991	0,99991	0,99992	0,99992	0,99992	0,99992	**3,7**
3,8	0,99993	0,99993	0,99993	0,99994	0,99994	0,99994	0,99994	0,99995	0,99995	0,99995	**3,8**
3,9	0,99995	0,99995	0,99996	0,99996	0,99996	0,99996	0,99996	0,99996	0,99997	0,99997	**3,9**

BIBLIOGRAFÍA

JAMES, G. (2014) *An Introduction to Statistical Learning with applications*. Berlín: Springer.

JANERT, P. (2011) *Data Analytics with Open Source Tools*. Sebastopol: O'Reilly Media.

MCKINNEY, W. (2012) *Python for Data Analysis*. Sebastopol: O'Reilly Media.