

Clase 1

Análisis Inteligente de Datos - 2023

Prof: PhD. Débora Chan - debiechan@gmail.com



INGENIERÍA

Posgrados

Organización

- 1 Estadística Descriptiva
 - Organización de los Datos
- 2 Medidas descriptivas Univariadas
- 3 Representación gráfica
- 4 Información Multivariada
- 5 Medidas de posición y dispersión en datos multivariados
- 6 Análisis Multivariado

Estadística Descriptiva

El análisis descriptivo

es el primer paso para aproximarnos a la comprensión de la estructura visible y subyacentes de nuestros datos, así como para extraer información relevante para nuestros análisis.

Este análisis

comprende la organización de los datos, la visualización de los mismos y la presentación de resúmenes adecuados que orienten y faciliten análisis posteriores.

Niveles de Medición de Variables

- **Categóricas o cualitativas:** Las modalidades de estas variables sólo se distinguen por ser distintas, no se puede establecer un ordenamiento entre ellos. Son ejemplos de estas variables: barrio de residencia, modelo de auto, sexo.
 - **Cuasicuantitativas u ordinales:** Las modalidades de estas variables, se puede ordenar sin embargo no se puede establecer una distancia entre ellas. Por ejemplo: calificación de examen (A, B, C, D y E), estadío de una enfermedad (I, II, III o IV).

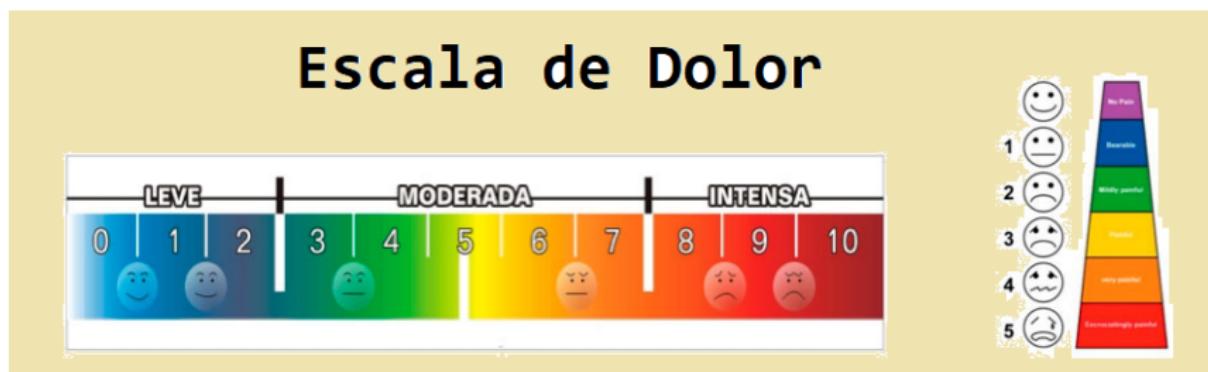


Niveles de Medición de Variables

- **Cuantitativas discretas:** Estas variables toman valores numéricos con la particularidad de que no existen valores intermedios entre dos consecutivos. Se vinculan al proceso de contar.
Ejemplos: cantidad de hijos, materias cursadas, dinero en una billetera.
 - **Cuantitativas continuas:** Estas variables también toman valores numéricos reales, asociándose generalmente al proceso de medir.
Ejemplos: distancia, edad, duración de un llamado.

Otras Escalas de Medición

Podemos mencionar las **escalas analógicas o visuales** que se utilizan en muchas ocasiones para que el paciente indique el grado de alguna variable “de nivel subjetivo” como dolor, bienestar, agrado, acuerdo-desacuerdo o sensaciones en general y controlar con la misma percepción en otro momento o respecto a otro tema.



Presentación de los Datos

Conviene organizar toda la información disponible para a fin de facilitar su comprensión e interpretación. El análisis sobre los datos crudos, puede resultar inabordable.

El tema es

¿Cómo convendría entonces organizar la información?

Cuando se analiza una única variable, es recomendable confeccionar una tabla denominada

Distribución de Frecuencias

Veamos qué aspecto tienen para cada tipo de variable.

Organización de los Datos

Distribución de Frecuencias: Variables Cualitativas

Para **datos cualitativos**: Las clases se definen según el interés de la investigación y las frecuencias absolutas se corresponden con la cantidad de observaciones registradas para esa clase. Registramos la venta de modelos Chevrolet en una concesionaria de Palermo.

Modelo	Frecuencia
Tracker	6
Spin	10
Onix	7
Cruze	12
TradeBlazer	17

Tabla: Ejemplo de distribución de frecuencias



Distribución de Frecuencias: Variables Cuantitativas

Las filas de la distribución de frecuencias:

- 🕒 En el caso de variables **discretas**, las modalidades quedan definidas por los valores del recorrido de la variable.
 - 🕒 En el caso de variables **continuas**, el recorrido de la variable se parte en 'intervalos de clase'.
 - 🕒 En **ambos casos**, se registra la frecuencia absoluta de cada modalidad (cantidad de observaciones en ella) o de cada intervalo (cantidad de observaciones dentro del rango del intervalo definido).

Distribución de Frecuencias: Variables Discretas

Estudiamos ahora la evolución de las ventas en una determinada sucursal, durante los últimos 50 meses.

Cantidad de Autos	Meses
12	2
14	3
15	7
16	4
17	8

Tabla: Ejemplo de variable discreta

Distribución de Frecuencias: Variables Continuas

Estamos interesados en estudiar el tiempo (en horas) dedicado a la lectura de textos de los alumnos de la diplomatura.

Intervalo de clase	f_i (frec. absoluta)
[0, 2)	7
[2, 4)	5
[4, 6)	6
[6, 8)	10
[8, 10)	4

Tabla: Ejemplo de frecuencias absolutas



Organización de los Datos

Frecuencias relativas y porcentuales

Si nos dicen 8 son los premiados de un curso...nos surge inmediatamente 8 de 10? 8 de 40? 8 de 120?...no es lo mismo!!

En cambio una frecuencia porcentual nos dice el 20% fueron premiados y nos da otra información!!.

Frecuencias Relativas

Si las frecuencias observadas en las m modalidades de la variable las denotamos con f_i , siendo $1 \leq i \leq m$ y $f_1 + f_2 + \dots + f_m = N$. entonces

$$f_r(i) = \frac{f_i}{N}$$

Frecuencias Porcentuales

$$f_P(i) = 100 * f_r(i)$$

Organización de los Datos

Frecuencias Relativas y Porcentuales (Ejemplo)

Int. de clase	$f\%$ (frec. relativa)	$f\%$ (frec. porcentual)
[0, 2)	7/32	21.88
[2, 4)	5/32	15.62
[4, 6)	6/32	18.75
[6, 8)	10/32	31.25
[8, 10)	4/32	12.5

Tabla: Ejemplo de frecuencias porcentuales

Organización

- 1 Estadística Descriptiva
- 2 Medidas descriptivas Univariadas
- 3 Representación gráfica
- 4 Información Multivariada
- 5 Medidas de posición y dispersión en datos multivariados
- 6 Análisis Multivariado

Medidas de Tendencia Central

Media Aritmética o Promedio muestral

Sintetizan un conjunto de valores con un solo valor que es el punto en torno al cual se localiza este conjunto de datos.

Es el promedio de las observaciones registradas. Dado un conjunto de datos $\{x_1, x_2, \dots, x_n\}$, su expresión es:

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

- ☒ Se puede calcular sólo para escalas de medición cuantitativas. Su cálculo es sencillo.
- ☒ Preserva la dependencia lineal; si $y = ax + b \Rightarrow \bar{y} = a\bar{x} + b$.
- ☒ No puede aplicarse a datos censurados.
- ☒ Es muy sensible a la presencia de valores extremos (muy alejados del conjunto de datos), vale decir que no es una medida robusta.

Medidas de Tendencia Central

Mediana o Q_2

Mediana: es el valor que divide a la distribución ordenada en dos partes iguales, cada una de las cuales contiene el 50% de las observaciones. Si la muestra ordenada es: $x^{(1)} \leq x^{(2)} \leq \dots \leq x^{(n)}$, entonces la mediana es

$$\tilde{x} = \begin{cases} x\left(\frac{n+1}{2}\right) & \text{si } n \text{ es impar} \\ \frac{x\left(\frac{n}{2}\right) + x\left(\frac{n}{2}+1\right)}{2} & \text{si } n \text{ es par} \end{cases}$$

Propiedades

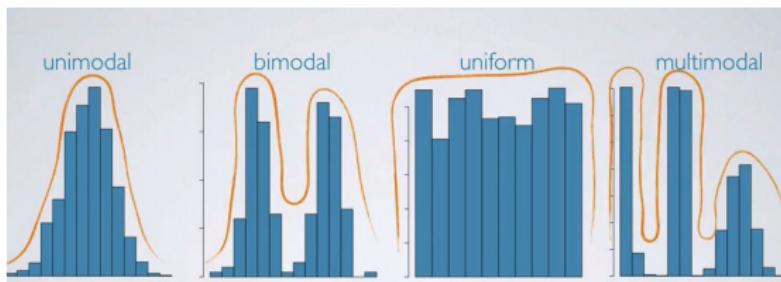
- 👉 Es de cálculo sencillo.
- 👉 Se puede calcular para escalas de medición al menos ordinales.
- 👉 Preserva la dependencia lineal; es decir, si $y = ax + b$ entonces $\tilde{y} = a\tilde{x} + b$.
- 👉 No es sensible a la presencia de valores extremos, por lo que es una medida robusta.

Medidas de Tendencia Central

Modo o Moda

Moda: es la observación de mayor frecuencia se denota Mo . No es una medida muy estable, dado que una sola observación puede cambiar el valor de la moda.

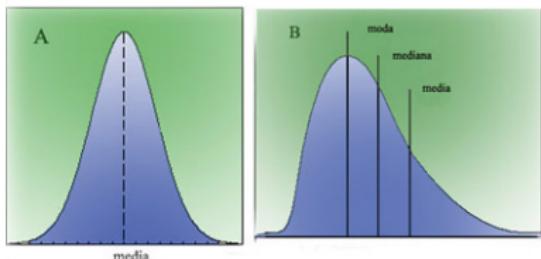
Puede no ser única, de hecho existen distribuciones bimodales o multimodales, en cuyo caso no resulta una medida de tendencia central muy informativa.



Medidas de Tendencia Central

Ejemplos

Presentamos varios casos para el cálculo de las medidas previamente



definidas.

-  Para los datos $\{12, 12, 15, 18, 23\}$, se tiene que $\tilde{x} = 15$, $\bar{x} = 16$ y $Mo = 12$.
-  Si los datos son $\{12, 12, 15, 17, 25, 25\}$, entonces $\tilde{x} = \frac{15 + 17}{2} = 16$, $\bar{x} = 17.67$ y existen dos modas $Mo = 12$ y $Mo = 25$.

Medidas de Tendencia Central

Media α -podada

Media α -podada: se define como el promedio de los datos centrales recortando el $\alpha\%$ de los valores más grandes y el $\alpha\%$ de los valores más chicos. Se denota como \bar{x}_α . Esta medida tiene como posiciones extremas a la media aritmética y a la mediana que se corresponden con $\alpha\% = 0$ y $\alpha\% = 50$ respectivamente. Calculemos la media podada al 10% para los siguientes datos:

2 – 4 – 5 – 6 – 7 – 7 – 8 – 8 – 8 – 9 – 9 – 10 – 13 – 14 – 14 – 14 – 15 – 15 – **15** – 25

Sin considerar los números en negrita,

$$\bar{x}_{0.10} = \frac{5 + 6 + 7 \cdot 2 + 8 \cdot 3 + 9 \cdot 2 + 10 + 13 + 14 \cdot 3 + 15 \cdot 2}{16} = 10.125$$

Medidas de Posición

Estadísticos de Orden

Los datos ordenados de menor a mayor se denotan como

$$x^{(1)} \leq x^{(2)} \leq \cdots \leq x^{(n)}$$

Entonces son estadísticos de orden

$x^{(1)}$ es el **valor mínimo** observado

$x^{(n)}$ es el **valor máximo** observado

La mediana puede pensarse como medida de tendencia central, o bien como un estadístico de orden.

Los cuantiles

sudividen al conjunto de datos en partes iguales(todas con igual cantidad de observaciones.) Pueden o no corresponder a valores observados.

Algunos Cuantiles

Los más usados son los **cuartiles** Q que dividen las observaciones en cuatro partes iguales, los **deciles** D que lo hacen en diez partes iguales y los **percentiles** P que lo hacen en 100 partes iguales.



Los cuartiles se denotan Q_1 , Q_2 y Q_3 y se denominan primer, segundo y tercer cuartil.

La mediana

coincide con el Q_2 , el D_5 y el P_{50}

El rango

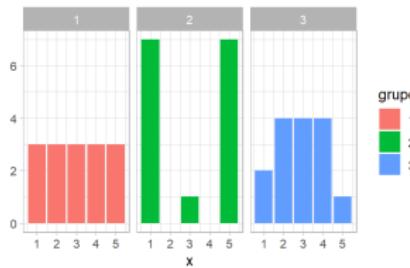
Estas medidas nos indican la variabilidad de los datos. Cuantifican la concentración de los datos alrededor de una medida de posición central.

Rango Muestral

Se define como la diferencia entre el valor máximo y el valor mínimo de la muestra, es decir:

$$rg(x) = x^{(n)} - x^{(1)}$$

Si bien su cálculo es sencillo, no resulta en general muy informativa.
Mismo rango es igual a misma variabilidad?



Medidas de Dispersion

Varianza

La Varianza Muestral

se define como el promedio de los cuadrados de las distancias de las observaciones a la media muestral; simbólicamente,

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

En el caso de la varianza poblacional el denominador es n .

Fundamento Teórico

La razón para este cambio de denominador es que la varianza muestral calculada de esta forma es un estimación mas precisa de la varianza poblacional, especialmente cuando n es pequeño.

Medidas de Dispersion

La Varianza: propiedades

Propiedades

-  Es de cálculo sencillo.
-  Sólo se puede calcular sólo para variables cuantitativas y con datos no censurados.
-  Si $y = ax + b$, entonces $s_y^2 = a^2 s_x^2$.
-  Las unidades de medición de la varianza son el cuadrado de las unidades de los datos originales.
-  Es muy sensible a la presencia de valores extremos. No es una medida *robusta*.
-  En los casos en que la media no resulta adecuada como medida de tendencia central, tampoco la varianza lo es como medida de dispersión.

Desviación Estándar

Desvío estandar muestral

Se define como la raíz cuadrada de la varianza y permite retornar a las unidades de medición originales. Simbólicamente:

$$s_x = \sqrt{s_x^2}$$

Coeficiente de variación (CV)

es una medida de dispersión relativa porque mide la proporción que representa el desvío estándar de la media aritmética. Es usual que se exprese en porcentaje.

Para comparar la dispersión de dos conjuntos de datos en una misma variable o dos variables de un mismo conjunto de datos medidas en distintas unidades es ideal.

Medidas de Dispersion

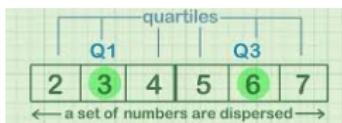
Alternativas Robustas: RI

Cuando la media no resulta representativa, tampoco serán adecuadas la varianza o la desviación estandar. Otras alternativas son RI y MAD.

Rango intercuartílico (RI)

Este valor informa el rango del 50% de los valores centrales del conjunto de datos. Se define como la diferencia entre el tercer cuartil y el primero. Simbólicamente: $RI = Q_3 - Q_1$

Veremos luego que corresponde al largo del diagrama de caja.



$$RI = Q_3 - Q_1$$

$$RI = 6 - 3 = 3$$

Medidas de Dispersion

Alternativas Robustas: MAD

MAD (Median Absolute Deviation)

es la mediana de los desvíos absolutos respecto de la mediana.

Consideremos el siguiente conjunto de observaciones $\{2, 3, 5, 7, 13, 50\}$

$$\text{La mediana es } \tilde{x} = \frac{5 + 7}{2} = 6.$$

Los desvíos respecto de la mediana son: $-4, -3, -1, 1, 7, 44$.

Los valores absolutos de los desvíos ordenados son: $1, 1, 3, 4, 7, 44$.

$$\text{La mediana de los valores absolutos de los desvíos es } MAD = \frac{3 + 4}{2} = 3.5.$$

Para hacerlo comparable con la DS, se propone la siguiente normalización

$$MADN(X) = \frac{MAD(X)}{0.6745}$$

En caso de normalidad coinciden **MADN y desvío standard coinciden**.

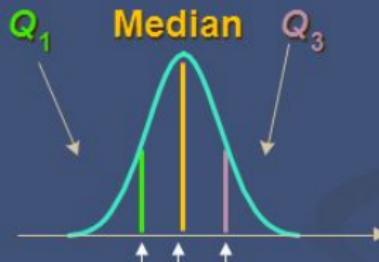
Otras medidas para caracterizar la distribución

Coeficiente de Asimetría

Coeficiente de asimetría muestral de Fisher

es una medida que describe la asimetría de la distribución de los datos con respecto a la media muestral. Su expresión analítica es

$$sk_F(x) = \frac{\sqrt{n} \sum_{j=1}^n (x_j - \bar{x})^3}{\left[\sum_{j=1}^n (x_j - \bar{x})^2 \right]^{\frac{3}{2}}}$$



Otras medidas para caracterizar la distribución

Coeficiente de Asimetría

Coeficiente de asimetría de Pearson

mide la asimetría cuantificando la separación entre la moda respecto de la desviación estándar. Este coeficiente es menos usual dado que requiere que la distribución sea unimodal. Su expresión es:

$$sk_P(x) = \frac{\bar{x} - Mo(x)}{s_x}$$

Coeficiente de asimetría de Bowley

se apoya en los cuartiles, focalizando en el 50% de los valores centrales. Se utiliza cuando la media y el ds no son representativos. Su expresión es

$$sk_B(x) = \frac{(q_3 - q_2) + (q_1 - q_2)}{q_3 - q_1} = \frac{q_3 + q_1 - 2\tilde{x}}{q_3 - q_1}$$

Otras medidas para caracterizar la distribución

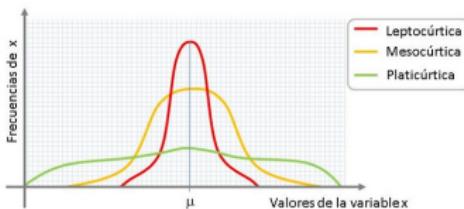
Coeficiente de Curtosis

Coeficiente de curtosis muestral

describe el grado de apuntamiento de una distribución. Describe el comportamiento de las colas de la distribución. Su expresión es:

$$k(x) = \frac{n \sum_{j=1}^n (x_j - \bar{x})^4}{\left[\sum_{j=1}^n (x_j - \bar{x})^2 \right]^2}$$

Las distribuciones leptocúrticas tienen coeficientes superiores a 3 y las platicúrticas coeficientes menores a 3.



Organización

1 Estadística Descriptiva

2 Medidas descriptivas Univariadas

3 Representación gráfica

- Alternativas al Boxplot (recientes)

4 Información Multivariada

5 Medidas de posición y dispersión en datos multivariados

6 Análisis Multivariado

Diagrama circular

Es adecuado para representar la distribución de variables cualitativas y cuasicuantitativas. Permite visualizar la proporción captada por cada categoría de la variable. Algunos de estos diagramas permiten visualizar dos niveles de la variable.

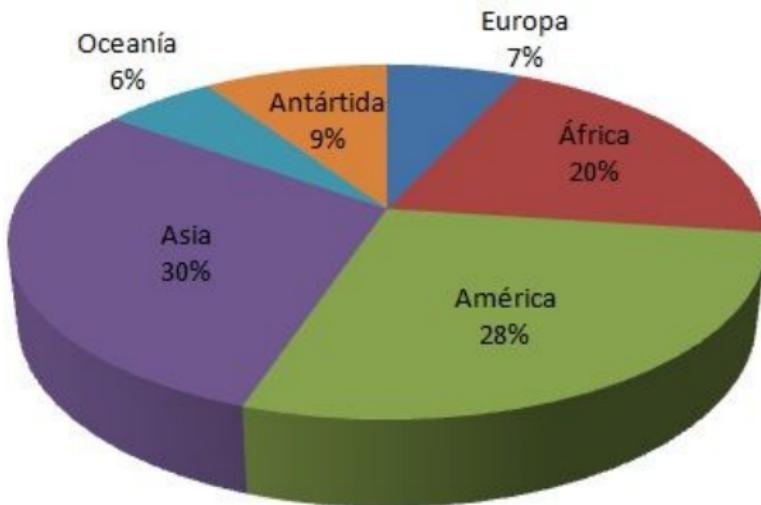
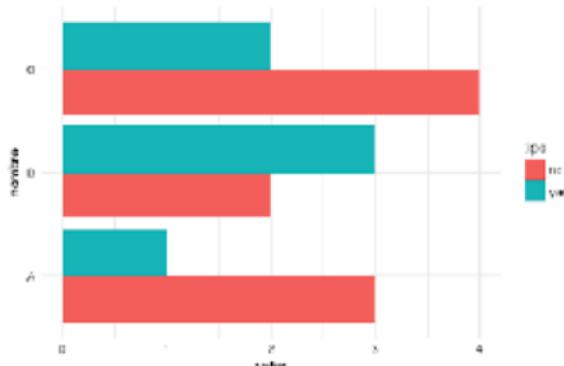


Gráfico de Barras / Superpuestas / Adyacentes

Es adecuado para representar variables cualitativas y también permite apreciar la distribución conjunta de más de una variable. En uno de los ejes se colocan los valores de la variable y la longitud de las barras es proporcional a la frecuencia de cada valor.



Pirámides poblacionales

Pirámide Poblacional Isla Santiago

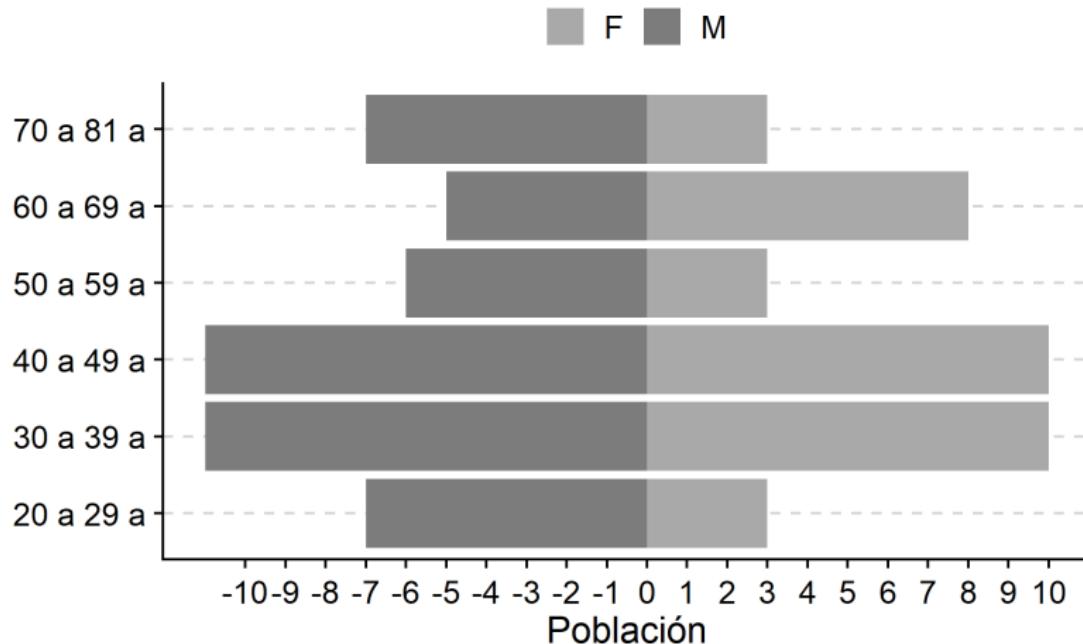
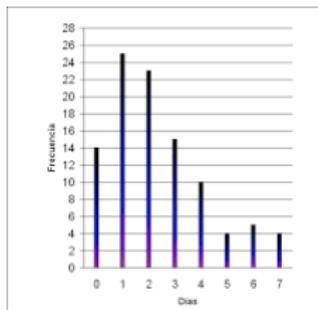
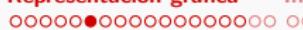
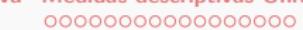


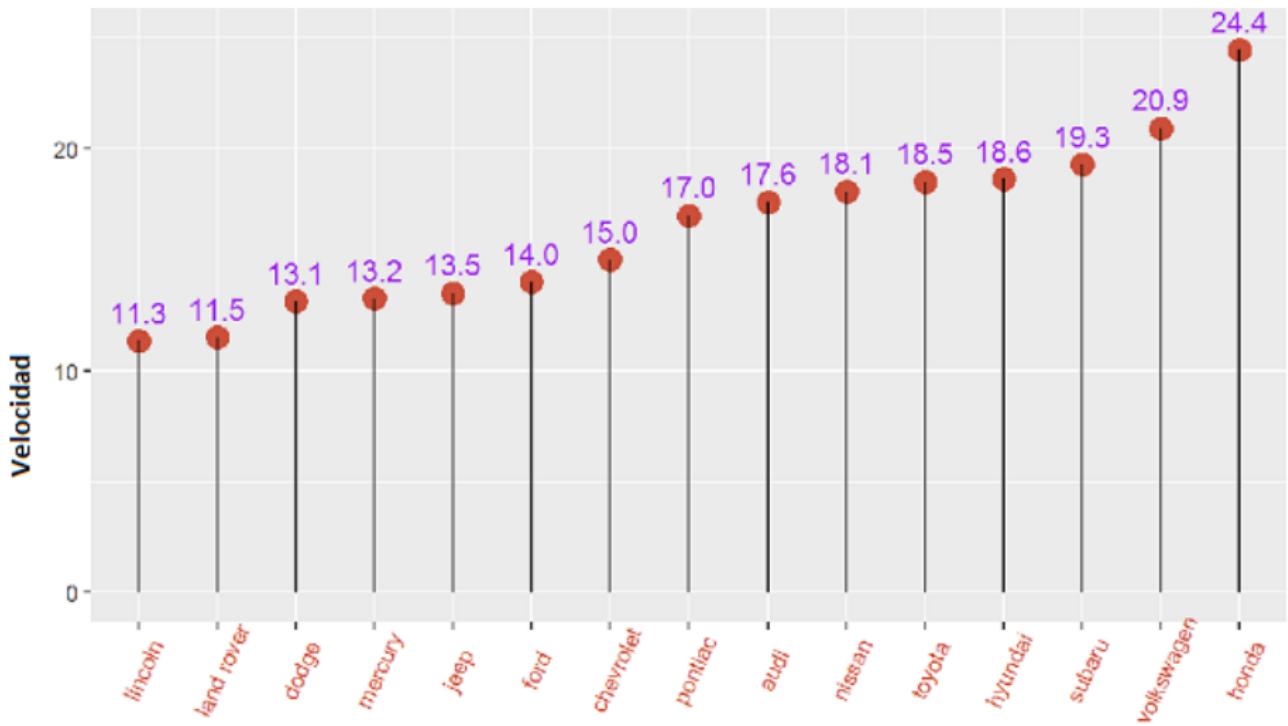
Gráfico de bastones

Es adecuado para representar la distribución de frecuencias de una variable discreta. Nuevamente sobre el eje horizontal se representan los valores de la variable y sobre el vertical bastones con longitud proporcional a la frecuencia de ese valor.



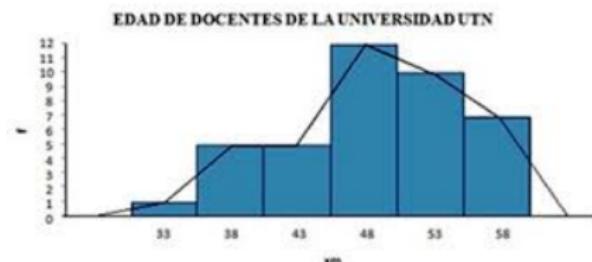


Una versión más moderna: Piruleta (lollipop)



Histograma y polígono de frecuencias

Se usa para representar distribuciones de variables continuas. Permite visualizar la forma de la distribución. El rango de la variable se subdivide en intervalos de clase y la altura es proporcional a la frecuencia sólo en el caso en que todos los intervalos tengan la misma longitud.



Uniendo los puntos medios de las bases superiores de los rectángulos del histograma se construye un polígono de frecuencias. Si la longitud de las bases de los rectángulos se redujera indefinidamente, el polígono de frecuencias tendería a la curva de densidad de la distribución.

Histogramas

Cómo elegir el ancho de Clase?

Varios autores propusieron respuestas alternativas a esta pregunta.

El **número de intervalos**, k , sugerido por las siguientes tres reglas depende de la cantidad n de datos. Las reglas proponen tomar la parte entera y son

- $k = \lfloor 10 \log(n) \rfloor$, • **Dixon y Kronmal (1965)**- $k = r2\sqrt{n}$, **Velleman (1976)**.
- $k = \lfloor 1 + \log_2(n) \rfloor$, • **Sturges (1926)**- $h_n = 3.49sn^{-1/3}$, **Scott (1979)**.
- $h_n = 2Rn^{-1/3}$, **Freedman y Diaconis (1981)**.

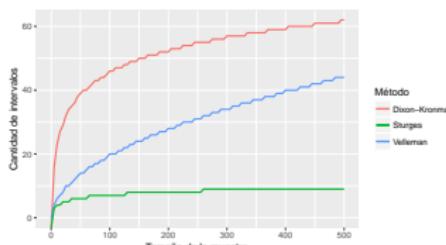
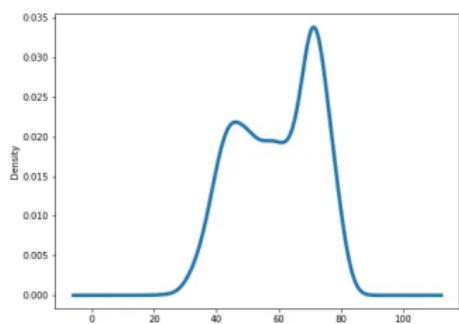
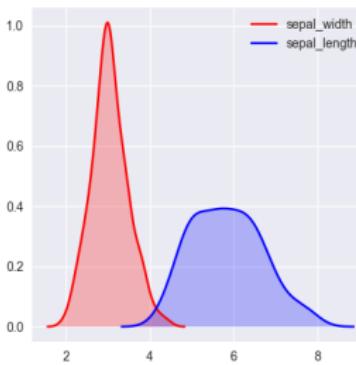


Gráfico de Densidad

Una sola distribución



Dos distribuciones



Qué se puede observar en esta gráfica?

Tiene ventajas sobre el histograma?

Boxplot o Diagrama de Caja

John Wilder Tukey (1915-2000) propuso este gráfico sencillo, de fácil lectura para presentar datos numéricos.

- ✿ Se dibuja un rectángulo o caja (*box*) cuyos extremos son los cuartiles primero y tercero. Dentro de ella, se dibuja un segmento que corresponde a la mediana o segundo cuartil.
- ✿ A partir de cada extremo, se dibuja un segmento o bigote (*whisker*), hasta el dato más alejado que está, a lo sumo, a 1.5 veces RI del extremo de la caja.
- ✿ Se denominan *outliers* moderados a los datos cuya distancia a uno de los extremos de la caja es mayor que 1.5 veces el RI y menor que 3 veces el RI. Mientras que los *outliers* severos son los datos que están a una distancia mayor a 3 veces el RI de uno de los extremos de la caja.

Boxplot o Diagrama de Caja

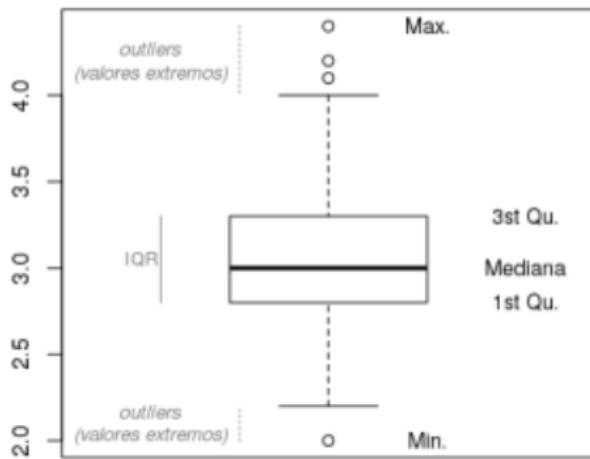
En un *boxplot* se aprecian distintos aspectos de la distribución de los datos:



posición- dispersión



asimetría -puntos anómalos o *outliers*



Detección de Outliers

Datos Salvajes Atípicos o Ouliers

Son datos alejados de alguna forma del patrón general del conjunto. Su detección puede determinar o influenciar fuertemente los resultados de un análisis estadístico clásico.

Esto ocurre porque muchas de las técnicas habitualmente usadas son muy sensibles a la presencia de este tipo de observaciones, especialmente en el caso de datos multivariados.

En presencia de estos datos salvajes, es conveniente recurrir a medidas robustas.

Los *outliers* deben ser **cuidadosamente inspeccionados**. Si no hay evidencia de error y su valor es posible **no deben ser eliminados**. Pueden estar alertando de anomalías de un tratamiento o patología, conjuntos especiales de clientes, etc.

Detección de Ouliers

Ejemplo

Para la siguiente muestra con $n = 13$, tenemos los siguientes datos:

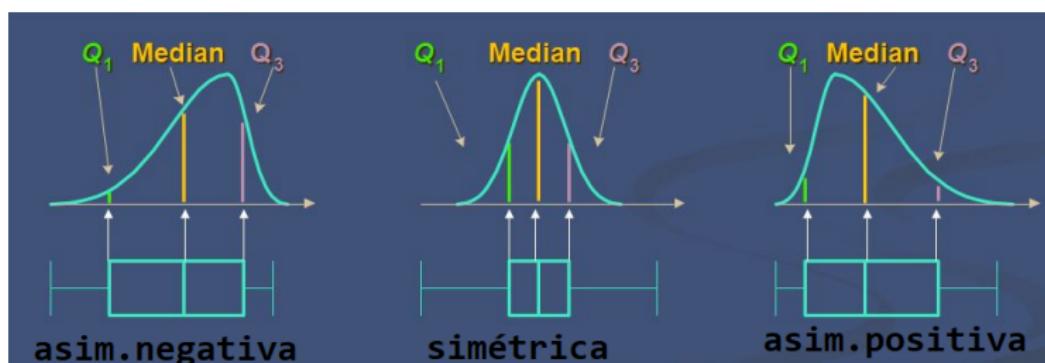
$$\{14, 18, 24, 26, 35, 39, 44, 45, 56, 62, 68, 87, 201\}$$

- ▽ $\tilde{x} = 44 \quad Q_1 = 25 \quad Q_3 = 65 \quad R.I. = 65 - 25 = 40$
- ▽ $Q_3 + 1.5 \cdot R.I. = 65 + 60 = 125 \quad Q_1 - 1.5 \cdot R.I. = 25 - 60 = -35$
- ▽ $VAS = 92$ (valor adyacente superior: mayor valor observado inferior a 125, es el extremo superior del segundo bigote.)
- ▽ $VAI = 14$ (valor adyacente inferior: menor valor observado superior a -35, es el extremo inferior del primer bigote.)
- ▽ $Q_3 + 3 * R.I. = 65 + 120 = 185$ y $Q_1 - 3 * R.I. = 25 - 120 = -95$
- ▽ $201 > Q_3 + 3 * R.I.$ por lo tanto es un *outlier severo*.

Cómo se aprecia la simetría en el Boxplot?

Aspecto

En la Figura podemos apreciar el aspecto del *boxplot* para distribuciones simétricas y asimétricas.



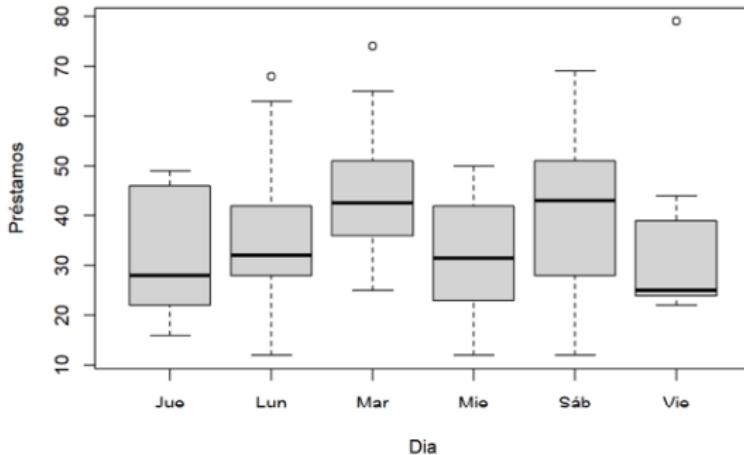
Cómo se aprecia la simetría en el Boxplot?

Observaciones

- ♣ En las distribuciones simétricas, la Mediana está cerca del centro de la caja y los bigotes tienen longitudes similares.
 - ♣ En distribuciones con asimetría positiva (o hacia la derecha), la Mediana se ubica más cerca del Q1, y/o el bigote inferior es de menor tamaño que el bigote superior. Suelen aparecer valores atípicos altos.
 - ♣ En distribuciones con asimetría negativa (o hacia la izquierda), la situación es inversa a la anterior.
 - ♣ Cuando hay varios *outliers* puede que la influencia de ellos se enmascare, es decir que para ciertas medidas se compense el efecto de unos con el efecto de otros.

Boxplots comparativos

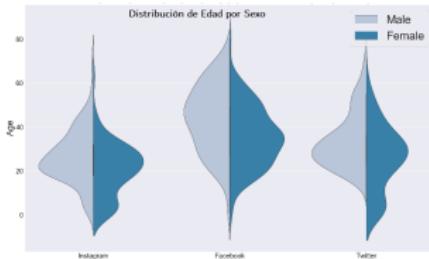
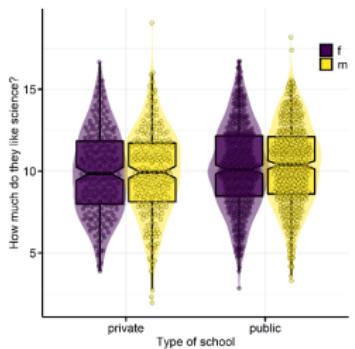
Los *boxplots* comparativos, permiten comparar el comportamiento de una variable en distintas poblaciones o varias variables en una población.



Realicen un mini-informe acerca de la gestión de préstamos y sus montos según el día de la semana en esta empresa financiera.

Alternativas al Boxplot (recientes)

Gráfico de Violín

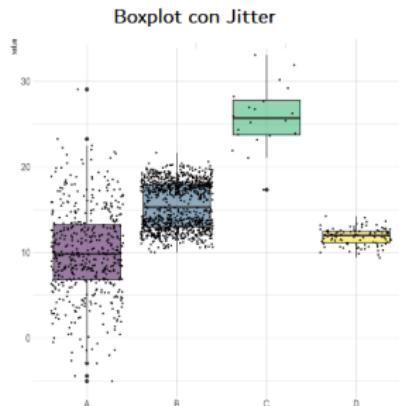
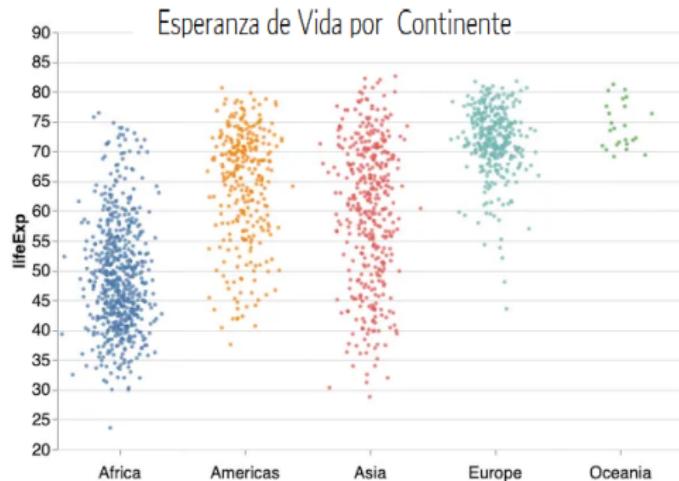


Qué se puede observar en esta gráfica?

Tiene ventajas sobre el histograma?

Alternativas al Boxplot (recientes)

Gráfico de Puntos (jitter.plot)



Qué se puede observar en esta gráfica?

Tiene ventajas sobre el boxplot?

Organización

- 1 Estadística Descriptiva
- 2 Medidas descriptivas Univariadas
- 3 Representación gráfica
- 4 Información Multivariada
- 5 Medidas de posición y dispersión en datos multivariados
- 6 Análisis Multivariado

Forma de Presentación de los Datos

La Base

La forma más usual en la que se presenta un conjunto de datos multivariados es una tabla donde se listan los valores de p variables observadas sobre n elementos.

	Variable ₁	...	Variable _j	...	Variable _p
Individuo ₁	$X_{1,1}$...	$X_{1,j}$...	$X_{1,p}$
⋮	⋮		⋮		⋮
Individuo _i	$X_{i,1}$...	$X_{i,j}$...	$X_{i,p}$
⋮	⋮		⋮		⋮
Individuo _n	$X_{n,1}$...	$X_{n,j}$...	$X_{n,p}$

Forma de Presentación de los Datos

Características

- ❖ Las **variables** aparecen en las columnas y son características o atributos que toman modalidades diferentes en los individuos de la población. Interesa estudiar el comportamiento de este conjunto de variables en este conjunto de observaciones.
- ❖ Los **individuos** aparecen en las filas. Son los ejemplares o elementos sobre los cuales se miden los atributos.
- ❖ Las tablas tendrán entonces n **filas** y p **columnas**; siendo n el número de instancias de observación también denominadas unidades de análisis y p la cantidad de variables observadas (o seleccionadas) sobre las cuales basaremos nuestro análisis.

Expresión Matricial de los Datos

Los datos pueden ser acomodados en una matriz de la siguiente manera

$$X = \begin{pmatrix} & \text{Variables en columnas} & \\ x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} \quad \begin{matrix} & \text{Individuos en filas} & \end{matrix}$$

Denotamos a cada elemento genérico de esta matriz como x_{ij} , que representa el valor de la variable j observado sobre el individuo i (fila i , columna j).

Ejemplo

Los datos de las galletitas se exhiben en la Tabla siguiente:

marca	valor.energ	carboh	prot	grasas	sodio
cerealitas	439	65	11	15	574
fajitas	466	57	10	22	828
express s/sal	445	69	11	14	12
oreo	478	67	5.6	21	363
melba	464	70	6.3	18	263
pepitos	463	66	7.1	19	136
criollitas	438	69	11	13	431
merengadas	418	69	6.3	13	201
sonrisas	423	70	6.8	13	241
maná	444	73	9	13	375
guinditas	407	70	6	12	106.7
pepas	437	60	6.7	18	76.67
polvorón	410	56.7	6.3	18	66.7

Ejemplo(continuación)

El análisis de datos multivariantes tiene por objeto el **estudio estadístico de varias variables** medidas en un subconjunto de elementos de una población.

La descripción de los datos multivariantes **comprende el estudio de cada variable aisladamente y también de las relaciones que quedan definidas entre ellas.**



En este ejemplo, con respecto a la matriz de datos, $p = 5$ y $n = 13$. El valor $x_{23} = 10$ representa la cantidad en gramos de proteínas cada 100 g de la segunda de las Fajitas; es decir, la tercera variable para las galletitas de la segunda fila.

Complejidad del Problema Multivariado

En los casos univariados, basta con estimar dos parámetros para la variable:

- ★ uno de centralidad (por ejemplo la media),
- ★ uno de dispersión (por ejemplo la varianza).

En el caso multivariado p -características sobre cada individuo, se deberán estimar:

$$p \text{ medias, } p \text{ varianzas y } \frac{p(p - 1)}{2} \text{ covarianzas.}$$

Vale decir que, tendremos que aproximar el valor de:

$$2p + \frac{p(p - 1)}{2} = \frac{p^2 + 3p}{2}$$

parámetros.

cuando p crece este número ya es difícil de manejar!!

Parámetros a Estimar

En un caso multivariado con p características medidas sobre cada individuo.

Variables	Parámetros a estimar
2	5
3	9
4	14
5	20
6	27
7	35
8	44
9	54
10	65

En el ejemplo de las galletitas se tiene que $p = 5$, lo que implica estimar 20 parámetros. Notemos que no depende de la cantidad de observaciones!!

Objetivos del Análisis Exploratorio

Algunos de los objetivos que se fijan en el análisis exploratorio son los siguientes.

- ✳ Familiarizarnos con los datos.
- ✳ Descubrir regularidades.
- ✳ Hallar asociaciones de variables.
- ✳ Analizar la existencia de estructuras ocultas.
- ✳ Detectar anomalías.
- ✳ Comprender los patrones descubiertos.
- ✳ Sintetizar la información.

Con estos propósitos resultará de utilidad disponer de los datos de forma tal que podamos observar y describir estos patrones.

Dos Variables Categóricas

Tabla de clasificación cruzada

En la Tabla mostramos las Consideraciones de los clientes respecto del consumo y respecto de la garantía al comprar aire acondicionado.

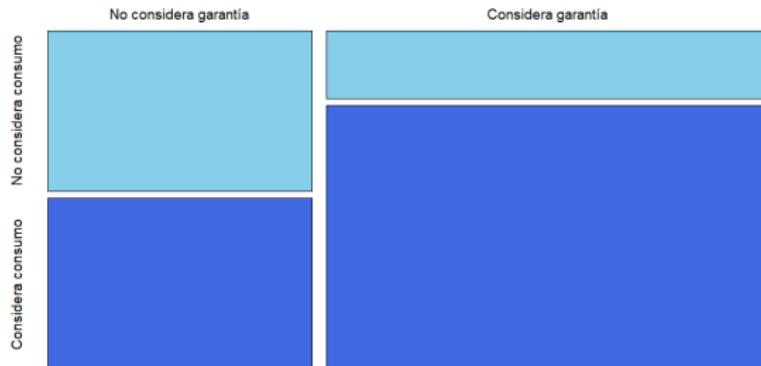
		Consumo		
		NO	SI	TOTAL
Garantía	NO	258	280	538
	SI	184	719	903
	TOTAL	442	999	1441

Cada una de las variables tiene dos niveles, por lo cual la tabla tiene dos filas y dos columnas, sin considerar la fila y la columna de totales.

Dos Variables categóricas

Gráfico de Mosaico

Cuando las dos variables consideradas son categóricas, una representación adecuada es el gráfico de mosaicos.

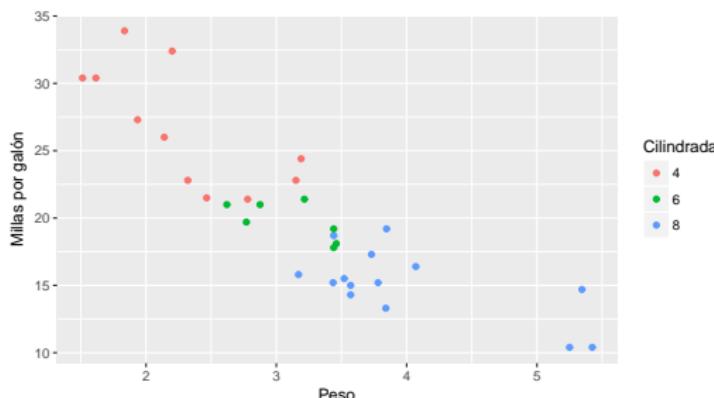


Se aprecia que es menor la proporción de consideraciones del consumo entre los que consideraron la garantía que entre los que la consideraron.

Dos Variables Cuantitativas

Diagrama de Dispersión

Con el conjunto de datos `mtcars` de R, donde se han tomado características de consumo, cilindradas, peso, número de carburadores y trasmisión en diferentes modelos de autos. Generamos el siguiente diagrama de dispersión.



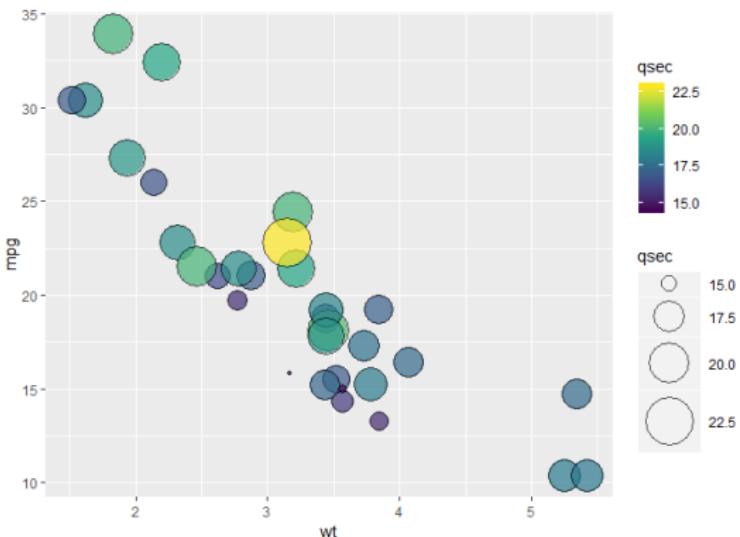
Dos Variables Cuantitativas

Diagrama de Dispersion (observaciones)

Esta Figura tiene representadas tres variables y podemos apreciar al mismo tiempo:

- Características individuales de la variable 'Peso'.
- Características individuales de la variable 'Millas por galón'.
- Relación entre variables cuantitativas por grupo definido por las cilindradas y en general.
- Posicionamiento de los grupos definidos por las cilindradas respecto de ambas.
- Posicionamiento relativo de los grupos en la relación establecida.
- En este caso tres variables dos cuantitativas y una categórica.

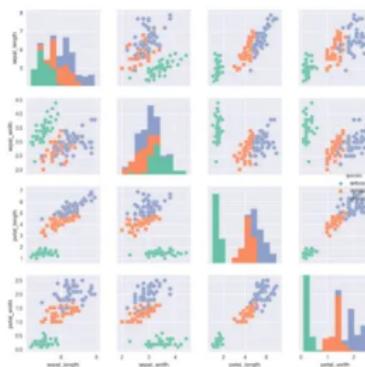
Una Versión más moderna (con tamaño por frecuencia)



Varias Variables cuantitativas

Dispersograma

Utilizando los datos de iris construimos un dispersograma.



En el dispersograma se aprecia la variación conjunta de cada par de variables, las diferentes especies se distinguen por color.

Varias Variables cuantitativas

Coordenadas Paralelas

Los gráficos de coordenadas paralelas constituyen una alternativa para la visualización datos multidimensionales.

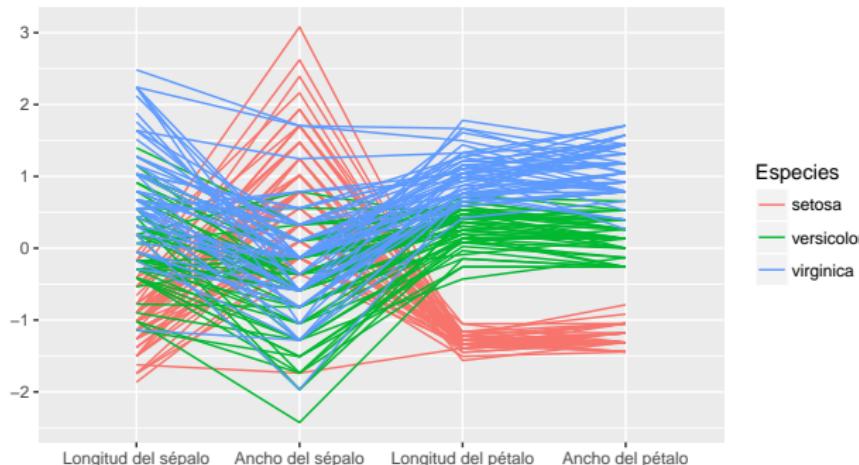


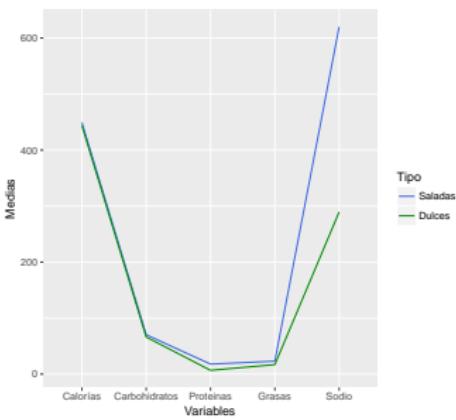
Figure: Gráfico de coordenadas paralelas(Datos Iris de R)

Varias Variables cuantitativas

Coordenadas Paralelas (observaciones)

- En lugar de usar ejes perpendiculares (x , y , z) se utilizan ejes paralelos.
- Cada atributo es representado en uno de estos ejes paralelos con sus respectivos valores.
- Se escalan los valores de los distintos atributos para que la representación de los mismos tenga la misma altura.
- Cada individuo se representa mediante una línea que une los puntos que le corresponden en los distintos ejes.
- De esta forma, se puede apreciar la similitud de las observaciones.
- También puede compararse la forma de distintos subgrupos o definir patrones, realizando el gráfico con diferentes colores para cada subgrupo.

Gráfico de Perfiles Multivariados



Se representan los valores medios o medianos de variable observada en distintos grupos de individuos. Esto permite comparar la posición central de estas variables en los distintos individuos o grupos definidos. Se aprecia en la misma que la composición nutricional media de las galletitas dulces y saladas es similar en todas las variables estudiadas, excepto en el contenido de sodio.

Curvas de nivel de la Densidad Normal Bivariada

Las curvas de nivel unen puntos de igual cantidad de observaciones (frecuencia) en nuestra base. De este modo, los distintos colores ayudan a identificar regiones de mayor o menor densidad de observaciones.

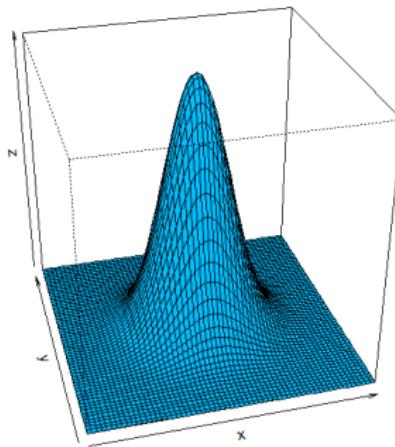


Figure: Gráfico de la distribución Normal Bivariada

Curvas de Nivel (observaciones)

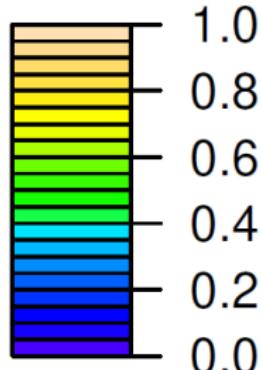
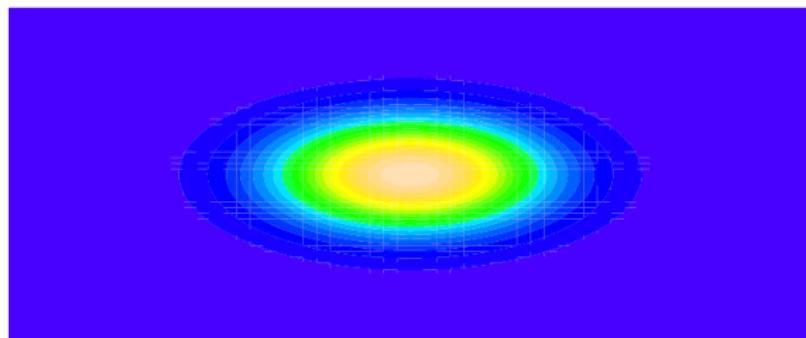
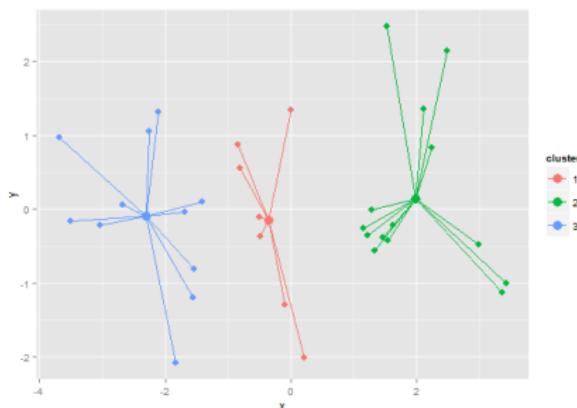


Figure: Gráfico de las curvas de nivel de la distribución Normal Bivariada

En este caso la gama elegida recorre la escala de azul a amarillo pasando por niveles medios de verde (mezcla de los colores extremos).

Gráficos de estrellas

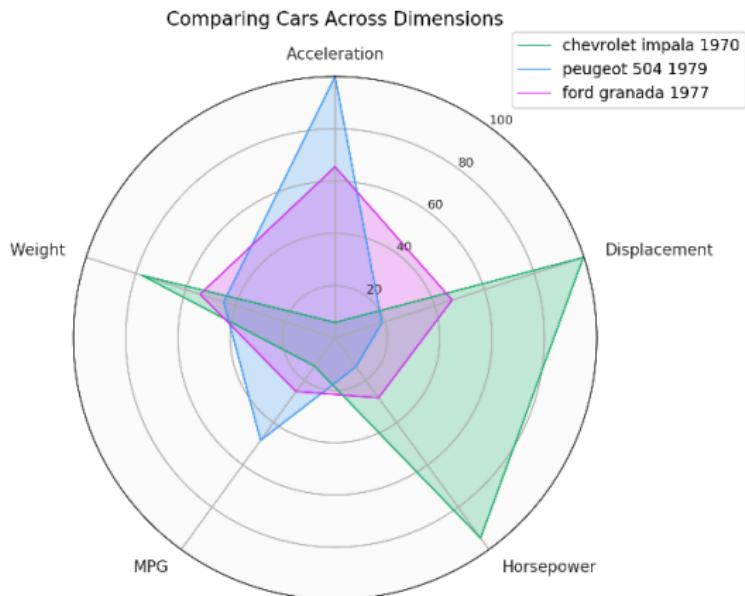
Cuando todas las variables consideradas son cuantitativas para poder detectar estructuras similares, resulta adecuado el gráfico de estrellas. Usamos nuevamente un conjunto subdividido en tres categorías.



Gráficos de Estrella (observaciones)

- ▶ Cada variable es representada con un radio de una estrella, la longitud del radio está dada por el valor de la variable en un individuo o bien por el promedio de observaciones de esa variable en el grupo.
- ▶ Por ejemplo podríamos representar en una estrella los autos familiares y en otra los utilitarios.
- ▶ En la Figura se aprecian similitudes y diferencias entre los grupos definidos por los colores.

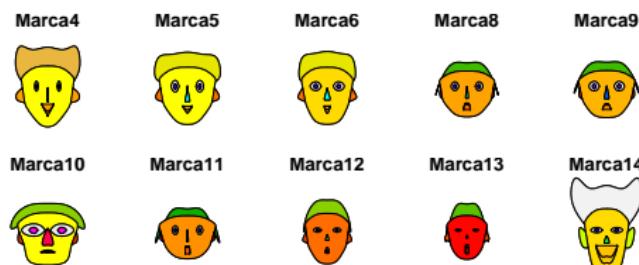
Gráficos de radar



Qué se puede apreciar en este gráfico?

Gráficos de caras de Chernoff

Las caritas de Chernoff son un método gráfico mediante el cual ciertas características cuantitativas de un grupo de observaciones se asocian con datos físicos de la cara de una persona. Esto permite realizar un dibujo que representa dichas características, y visualizar fácilmente similitudes y diferencias entre individuos, dado que estamos habituados a hacerlos con personas.



Organización

- 1 Estadística Descriptiva
- 2 Medidas descriptivas Univariadas
- 3 Representación gráfica
- 4 Información Multivariada
- 5 **Medidas de posición y dispersión en datos multivariados**
- 6 Análisis Multivariado

Vector Medio (vector de medias muestrales)

Un conjunto de p variables observadas sobre n individuos puede representarse mediante una matriz $X \in \mathbb{R}^{n \times p}$.

vector de medias muestral

se define

$$\bar{x} = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p) \in \mathbb{R}^p$$

donde \bar{x}_i se refiere al promedio de la i -ésima variable (columna) observada; es un vector formado por la media de cada una de las variables observadas.

Propiedades del vector de medias

Sean $X, Y \in \mathbb{R}^{n \times p}$ matrices que guardan los datos observados, y sean $A, B \in \mathbb{R}^{p \times k}$ y $C \in \mathbb{R}^{n \times k}$ matrices escalares, entonces:

$\overline{XA + C} = \bar{X}A + C.$

$\overline{XA + YB} = \bar{X}A + \bar{Y}B.$

Matriz de Varianzas y Covarianzas

matriz de varianzas y covarianzas muestral

Se define como:

$$\widehat{\Sigma} = \frac{1}{n}(X - \bar{X})^t(X - \bar{X})$$

donde \bar{X} es una matriz que en cada una de sus columnas tiene el promedio muestral de la variable respectiva repetido tantas veces como individuos tiene el conjunto de observaciones.

La matriz $\widehat{\Sigma}$, de tamaño $p \times p$, resulta ser simétrica y su diagonal principal está formada por las varianzas muestrales de cada una de las variables observadas; mientras que fuera de su diagonal, se encuentran las covarianzas muestrales de cada par de variables.

Matriz de varianzas y covarianzas

Propiedades

- La matriz de covarianzas muestral es simétrica: $\widehat{\Sigma}^t = \widehat{\Sigma}$, es decir que para todo i, j se cumple que $\widehat{\Sigma}_{ij} = \widehat{\Sigma}_{ji}$.
- $\widehat{\Sigma}$ estima a la matriz de varianzas y covarianzas poblacional $\Sigma = E[(X - 1_n\mu)^t(X - 1_n\mu)]$ que también es simétrica.
- La matriz de covarianzas (poblacional o muestral) es semidefinida positiva; es decir, que todos sus autovalores son mayores o iguales a cero.
- Si $Y = XA + B$, $\Sigma_Y = A^t \Sigma_X A$, siendo $A \in \mathbb{R}^{p \times k}$ y $B \in \mathbb{R}^{n \times k}$ matrices de escalares.
- $\widehat{\Sigma} = \frac{1}{n}(X - 1_n\bar{x})^t(X - 1_n\bar{x})$, siendo 1_n el vector columna de n unos.

Ejemplo

Vamos a buscar la matriz de covarianza muestral correspondiente al

conjunto de observaciones dado por $X = \begin{pmatrix} 10 & 4 \\ 15 & 1 \\ 20 & 7 \end{pmatrix}$.

Tenemos que

$$\bar{x} = \begin{pmatrix} 15 & 4 \end{pmatrix}, \quad \bar{X} = 1_3 \bar{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} 15 & 4 \end{pmatrix} = \begin{pmatrix} 15 & 4 \\ 15 & 4 \\ 15 & 4 \end{pmatrix} \quad \text{y}$$

$$\widehat{\Sigma} = \frac{1}{3} \left[\begin{pmatrix} 10 & 4 \\ 15 & 1 \\ 20 & 7 \end{pmatrix} - \begin{pmatrix} 15 & 4 \\ 15 & 4 \\ 15 & 4 \end{pmatrix} \right]^t \left[\begin{pmatrix} 10 & 4 \\ 15 & 1 \\ 20 & 7 \end{pmatrix} - \begin{pmatrix} 15 & 4 \\ 15 & 4 \\ 15 & 4 \end{pmatrix} \right]$$

$$= \frac{1}{3} \begin{pmatrix} 50 & 15 \\ 15 & 18 \end{pmatrix} = \begin{pmatrix} 16.6 & 5 \\ 5 & 6 \end{pmatrix}$$

Transformaciones del Conjunto de Datos

En algunas ocasiones, para optimizar el análisis de la información disponible, es conveniente realizar transformaciones a los datos. Las transformaciones pueden ser por filas o por columnas, o sea por individuos o por variables, dependiendo de los objetivos de las mismas.
Los objetivos más usuales de estas transformaciones son:

- Hacer comparables las magnitudes.
- Modificar la escala de medición.
- Satisfacer alguna propiedad estadística.

Transformaciones por filas

Las transformaciones por filas se aplican con el objeto de hacer comparables los valores asignados a los distintos individuos u objetos de análisis.

Por ejemplo, cuando un grupo de jueces deben evaluar un conjunto de individuos o productos, suele ocurrir que algunos de ellos tengan tendencia a poner puntuaciones muy altas o muy bajas de manera subjetiva, lo cual sesga el estudio. Para neutralizar estas diferencias se utilizan transformaciones por filas tales como las que veremos a continuación.



Transformaciones por individuo

Un juez podría tener una tendencia a puntuaciones muy altas o muy bajas lo cual sesgaría el estudio. Para neutralizar la influencia de esta tendencia, se realizan transformaciones por fila. Por ejemplo, la siguiente

$$T(x) = \begin{cases} \frac{x - \bar{x}}{x_{\max} - \bar{x}} & \text{si } x > \bar{x} \\ \frac{x - \bar{x}}{\bar{x} - x_{\min}} & \text{si } x < \bar{x} \end{cases}$$

La transformación de las puntuaciones superiores a la media de cada juez resultarán positivas, mientras que las que resulten inferiores a la media resultarán negativas. A las puntuaciones superiores se las normaliza por la distancia entre la media y el máximo, mientras que a las inferiores por la distancia entre la media y el mínimo.

Transformaciones por Columnas

Variables aleatorias estandarizadas

Suele denominarse a la transformación de estandarizado como *z-scores* o puntuaciones *Z*, ya que tienen la característica de tener media 0 y varianza 1. Las mismas se realizan restando a las observaciones el valor medio muestral y dividiendo esta diferencia por el desvíos estándar muestral. Simbólicamente,

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{s_j^2}} \quad (1)$$

Estas transformaciones tienen sentido en el caso en que la media y el desvío resulten una buena representación de la centralidad y la dispersión respectivamente. En caso contrario, pueden considerarse en forma alternativa la mediana y la desviación intercuartil o la mediana y el MAD.

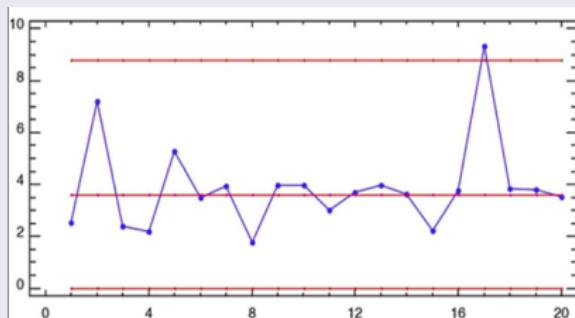
Organización

- 1 Estadística Descriptiva
- 2 Medidas descriptivas Univariadas
- 3 Representación gráfica
- 4 Información Multivariada
- 5 Medidas de posición y dispersión en datos multivariados
- 6 Análisis Multivariado

Análisis Multivariado vs Análisis Univariado

¿En qué nos beneficia realizar el análisis conjunto de todas las variables?

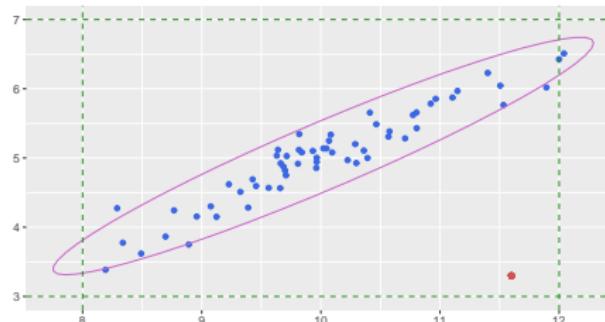
Consideremos un conjunto de cajas producidas por una máquina o un operador. Si observamos el comportamiento de una sola variable, podemos detectar si alguna observación está alejada de la mayor parte de los datos.



En la Figura podemos apreciar si el dato excede o está por debajo de las especificaciones, pero no podremos apreciar si la forma es la adecuada o no.

Análisis Multivariado vs Análisis Univariado

El *scatter plot* o gráfico de dispersión nos permite identificar variables que siguen el patrón general de interacción pero se alejan del centro de las variables. Asimismo permite identificar puntos que están dentro del rango de ambas variables pero la forma de su interacción no es la forma general del grupo.



Análisis Multivariado vs Análisis Univariado

Nos preguntamos ahora, ¿qué podemos observar en el dispersograma ???

- ✿ Cuáles variables parecen asociadas.
- ✿ Cuáles variables no parecen asociadas.
- ✿ Qué sentido se le encuentra a dichas asociaciones.
- ✿ Qué fuerza se le encuentra a dichas asociaciones.

Sin embargo, deberíamos encontrar un modo de cuantificar estas apreciaciones, siendo la covarianza muestral una forma posible.

Covarianza y Correlación

Covarianza muestral

Es una medida de asociación lineal entre dos variables. Se calcula sobre el conjunto de observaciones x_{ij} , mediante la siguiente fórmula:

$$s_{ik} = \frac{1}{n} \sum_{j=1}^n (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

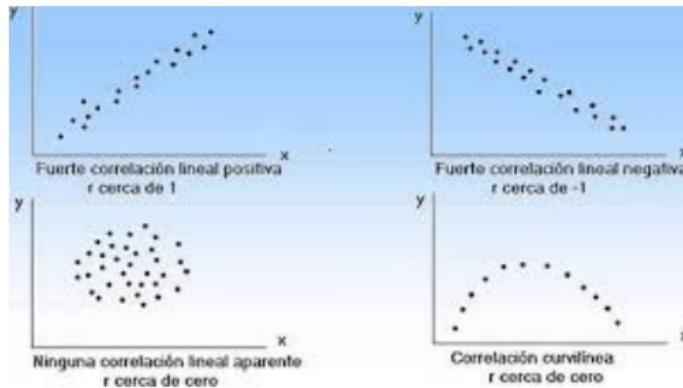
la matriz de varianzas y covarianzas es de la forma

$$\Sigma = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{pmatrix}$$

Covarianza y Correlación

Donde

- * $s_{ik} > 0$ indica una asociación lineal positiva entre los datos de las variables.
- * $s_{ik} < 0$ indica una asociación lineal negativa entre los datos de las variables.
- * $s_{ik} = 0$ indica que no hay una asociación lineal entre los datos de las variables.



Propiedades destacables de la covarianza

- * $\text{Cov}(X, X) = \text{Var}(X)$.
- * $\text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y)$.
- * Dados vectores aleatorios X e Y y matrices de constantes A y B , vale que $\text{Cov}(AX, BY) = A \text{Cov}(X, Y) B^t$.
- * La covarianza sólo detecta asociación lineal, mientras que otros tipos de asociación no son captadas por esta medida.

Ejemplo

Sean X e Y dos variables aleatorias tales que $\mu_X = 4$, $\sigma_X^2 = 2$, siendo $Y = -2X + 3$.

Utilizando propiedades de la varianza y de la esperanza matemática, tenemos que:

$$\mu_Y = -2 \cdot 4 + 3 = -5, \quad \sigma_Y^2 = 4 \cdot 2 = 8 \quad \text{y}$$

$$\text{Cov}(X, Y) = \text{Cov}(X, -2X+3) = -2\text{Cov}(X, X) = -2\text{Var}(X) = -2 \cdot 2 = -4$$

Luego, si consideramos el vector aleatorio (X, Y) , por lo visto, la matriz de covarianzas está dada por

$$\Sigma = \begin{pmatrix} 2 & -4 \\ -4 & 8 \end{pmatrix}$$

Es inmediato observar que el determinante de esta matriz es nulo; es decir, esta matriz es **singular**.

Ejemplo: ¿Por qué sucede esto?

Porque una de las variables es función lineal de la otra, el conjunto formado por ambas resulta **linealmente dependiente** y, por lo tanto, el determinante es nulo.

El valor (magnitud) de la covarianza depende las unidades en que se miden las variables. Este inconveniente puede salvarse realizando una estandarización. Así se obtiene una medida de la fuerza de la relación que no depende de las unidades de medición.

$$\text{Cov}(aX + b, cY + d) = ac \text{ Cov}(X, Y) \quad \forall a, b, c, d \in \mathbb{R}$$

Observación: La varianza muestral es la covarianza muestral entre los datos de la i -ésima variable con ella misma, algunas veces se denota como s_{ii}

Correlación muestral

El coeficiente de correlación lineal es una medida de asociación lineal para las variables, definida como la covarianza de los datos estandarizados.

Para los datos de la i -ésima y k -ésima variable se define como

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}} \sqrt{s_{kk}}}$$

La matriz de correlación muestral es de la forma

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1n} \\ r_{21} & 1 & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & 1 \end{pmatrix}$$

Entonces, r_{jk} es la correlación muestral entre Z_j y Z_k , columnas j y k de las variables estandarizadas. Tanto s_{ik} como r_{ik} son muy sensibles a la presencia de datos atípicos (*outliers*). En presencia de datos atípicos será recomendable utilizar otras medidas de asociación.

Propiedades de la correlación muestral

- ❖ $|r_{ik}| \leq 1$.
- ❖ Si $r_{ik} = 1$ significa que los datos yacen sobre una línea recta de pendiente positiva.
- ❖ Si $r_{ik} = -1$ significa que los datos yacen sobre una línea recta de pendiente negativa.
- ❖ Si $0 < r_{ik} < 1$ significa que los datos se ubican alrededor de una línea recta de pendiente positiva.
- ❖ Si $-1 < r_{ik} < 0$ significa que los datos se ubican alrededor de una línea recta de pendiente negativa.
- ❖ Si $r_{ik} = 0$ indica que no hay una asociación lineal entre las dos variables.

Traza de una matriz

Traza

de una matriz cuadrada a la suma de los elementos de la diagonal principal. Simbólicamente, si $A \in \mathbb{R}^{n \times n}$, $tr(A) = \sum_{i=1}^n a_{ii}$.

Siempre es posible calcular la traza de una matriz cuadrada. La traza es un número real, puede ser positivo, negativo o nulo. En el caso de las matrices de varianzas y covarianzas, como en el caso de las matrices de correlación, la traza es positiva.

Si $A = \begin{pmatrix} 2 & 3 \\ -4 & 8 \end{pmatrix}$, entonces $tr(A) = 2 + 8 = 10$.

Traza de la matriz de varianzas y covarianzas

Debido a que en una matriz de covarianzas, la diagonal principal

- está constituida por las varianzas de las variables que son valores mayores o iguales a cero, la traza de la misma es no negativa.
- la traza es entonces la suma de las varianzas de las variables consideradas en el conjunto de datos por lo cual indica de alguna forma la magnitud del problema.
- Como los elementos de la diagonal principal de la matriz de correlaciones son unos, su traza es la cantidad de variables p .

Trazas: Ejemplos

Siendo la matriz de covarianzas:

$$\Sigma = \begin{pmatrix} 2 & -4 \\ -4 & 8 \end{pmatrix}$$

y la matriz de correlaciones:

$$Corr = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}$$

Luego, $tr(\Sigma) = 2 + 8 = 10$ y $tr(Corr) = 1 + 1 = 2$.

Correlogramas

Nos permiten visualizar la fuerza y el sentido de la correlación entre un conjunto de variables.

- El color azul indica correlación positiva, el rojo negativa.
- Cuanto mayor es la intensidad del color más cercano a 1 en el caso positivo y a -1 en el caso negativo se encuentra el coeficiente de correlación.

