ANÁLISIS MULTIVARIANTE

Exploración e identificación de datos ausentes, outliers y supuestos del análisis multivariante

Mg. Víctor Guevara P

Universidad Nacional José Faustino Sánchez Carrión

Mayo 2023

Normas de participación.

- Ingresar puntual
- Realizar las actividades encomendadas.

Análisis previo de los datos

Análisis previo de los datos

La preparación de datos incluye los procesos de obtención de datos, limpieza, ingeniería de características y análisis exploratorio. Los analistas de datos informan que estos procesos suelen consumir del 60 al 90 por ciento del tiempo de un proyecto.

Caso 1: Nivel de obesidad

Conjunto de datos para la estimación de niveles de obesidad basados en hábitos alimentarios y condición física en individuos de Colombia, Perú y México.

Objetivo:

- El informe presenta datos para la estimación de los niveles de obesidad en individuos de los países de México, Perú y Colombia, con base en sus hábitos alimentarios y condición física.
- Los datos contienen 17 atributos y 2111 registros, los registros están etiquetados con la variable de clase NObesidad (Nivel de Obesidad), que permite clasificar los datos utilizando los valores de Peso Insuficiente, Peso Normal, Sobrepeso Nivel I, Sobrepeso Nivel II, Obesidad Tipo I , Obesidad Tipo II y Obesidad Tipo III.

Revisar: https://www.sciencedirect.com/science/article/pii/S2352340919306985

Caso 1: Nivel de obesidad

```r
# Importar el conjunto de datos
obesidad_df<-read.csv("https://raw.githubusercontent.com/sombragris1/SMedicas/main/Obesidad.csv",
                      sep = ";", stringsAsFactors = TRUE, encoding = "latin1")
```

```r
# Mostramos las primeras 6 observaciones
head(obesidad_df)
```

```
##      Genero Edad Estatura Peso Fam_sobrepeso FAVC FCVC NCP          CAEC Fuma
## 1  Femenino   21     1.62 64.0            Si   no    2   3 Algunas veces   no
## 2  Femenino   21     1.52 56.0            Si   no    3   3 Algunas veces   Si
## 3 Masculino   23     1.80 77.0            Si   no    2   3 Algunas veces   no
## 4 Masculino   27     1.80 87.0            no   no    3   3 Algunas veces   no
## 5 Masculino   22     1.78 89.8            no   no    2   1 Algunas veces   no
## 6 Masculino   29     1.62 53.0            no   Si    2   3 Algunas veces   no
##   CH2O SCC FAF TUE          CALC         Transporte            N_Obesidad
## 1    2  no   0   1         Nunca Transporte público           Peso normal
## 2    3  Si   3   0  Algunas veces Transporte público           Peso normal
## 3    2  no   2   1 Frecuentemente Transporte público           Peso normal
## 4    2  no   2   0 Frecuentemente            Caminar  Nivel de sobrepeso I
## 5    2  no   0   0  Algunas veces Transporte público Nivel de sobrepeso II
## 6    2  no   0   0  Algunas veces          Automóvil           Peso normal
```

Caso 2: Predicción del cáncer de mama

El conjunto de datos de datos para esta práctica corresponde a una muestra de distintas características de pacientes que se sometieron a la detección de cáncer de mama.

Las características se calculan a partir de una imagen digitalizada de una aspiración con aguja fina (FNA) de una masa mamaria. Describen las características de los núcleos celulares presentes en la imagen.

url: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.56.707&rep=rep1&type=pdf

Caso 2: Predicción del cáncer de mama

Información de atributo:

- radio (media de las distancias desde el centro a los puntos en el perímetro)
- textura (desviación estándar de los valores de escala de grises)
- perímetro
- área
- suavidad (variación local en las longitudes de los radios)
- compacidad (perímetro**2/área-1.0)
- concavidad (severidad de las porciones cóncavas del contorno)
- puntos cóncavos (número de porciones cóncavas del contorno)
- simetría
- dimensión fractal (“aproximación de la línea de costa” - 1)
- Diagnóstico (M= maligno, B= benigno)

Caso 2: Predicción del cáncer de mama

```
# Importar el conjunto de datos
cancer<-read.csv("https://raw.githubusercontent.com/VictorGuevaraP/Estadistica-R/master/cancer.csv",
                 sep = ";", stringsAsFactors = T)
```

```
#Mostrar los primeros registros
head(cancer)
```

```
##   radio textura perimetro   area Suavidad compacidad concavidad puntos_conc
## 1 17.99   10.38    122.80 1001.0  0.11840    0.27760     0.3001     0.14710
## 2 20.57   17.77    132.90 1326.0  0.08474    0.07864     0.0869     0.07017
## 3 19.69   21.25    130.00 1203.0  0.10960    0.15990     0.1974     0.12790
## 4 11.42   20.38     77.58  386.1  0.14250    0.28390     0.2414     0.10520
## 5 20.29   14.34    135.10 1297.0  0.10030    0.13280     0.1980     0.10430
## 6 12.45   15.70     82.57  477.1  0.12780    0.17000     0.1578     0.08089
##   simetria dim_fractal peor_radio peor_perimetro peor_area peor_concav
## 1   0.2419     0.07871      25.38          184.60    2019.0      0.7119
## 2   0.1812     0.05667      24.99          158.80    1956.0      0.2416
## 3   0.2069     0.05999      23.57          152.50    1709.0      0.4504
## 4   0.2597     0.09744      14.91           98.87     567.7      0.6869
## 5   0.1809     0.05883      22.54          152.20    1575.0      0.4000
## 6   0.2087     0.07613      15.47          103.40     741.6      0.5355
##   peor_pun_conc diagnostico
## 1        0.2654           M
## 2        0.1860           M
## 3        0.2430           M
```

Estadística descriptiva

Estadística descriptiva

- Las estadísticas descriptivas se utilizan para describir las características básicas de los datos en un estudio.
- Proporcionan resúmenes sencillos sobre la muestra y las medidas.
- Junto con el análisis gráfico simple, forman la base de prácticamente todos los análisis cuantitativos de datos.

Tablas de frecuencia

En estadística, se le llama distribución de frecuencias a la agrupación de datos en categorías mutuamente excluyentes que indican el número de observaciones en cada categoría.

Tabla N° 01
Consumo anual per cápita en Litros

	f_i	h_i	p_i
Agua de Mesa	6	0.15	15%
Gaseosa	28	0.7	70%
Néctar	4	0.1	10%
Refrescos Fluidos	2	0.05	5%
TOTAL	40	1	100%

Para variables cualitativas

Tabla N° ..
Distribución de vendedores según el número de ventas realizadas

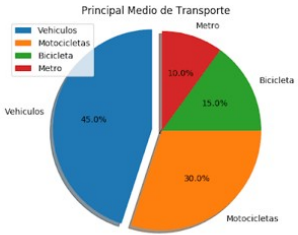
Número de Ventas	Frecuencia absoluta simple	Frecuencia absoluta acumulada	Frecuencia relativa simple	Frecuencia relativa acumulada	Frecuencia porcentual simple	Frecuencia porcentual acumulada
1	3	3	0.12	0.12	12%	12%
2	6	9	0.24	0.36	24%	36%
3	8	17	0.32	0.68	32%	68%
4	5	22	0.20	0.88	20%	88%
5	2	24	0.08	0.96	8%	96%
6	1	25	0.04	1.00	4%	100%
TOTAL	25		1.00		100%	

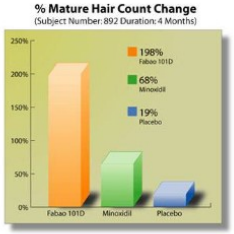
Para variables Cuantitativas

Presentación de datos (Gráficos estadísticos)

- Una representación visual de la información, que permite mostrarla a través de formas que corresponden a la magnitud de los datos medidos. Además de poder ver el comportamiento de la variable en análisis.

Datos Cualitativos: Gráficos





Presentación de datos (Gráficos estadísticos)

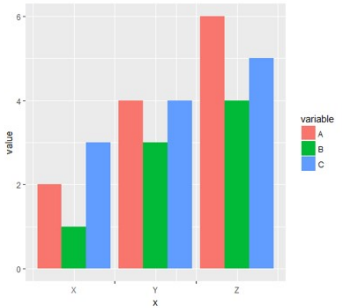


Gráfico de Barras Comparativas

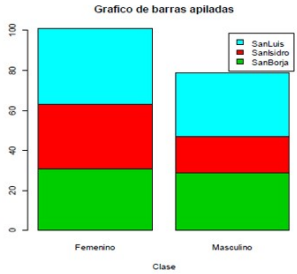
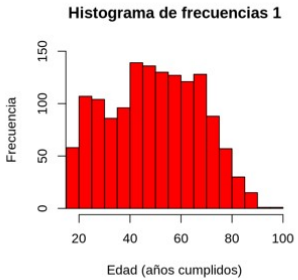
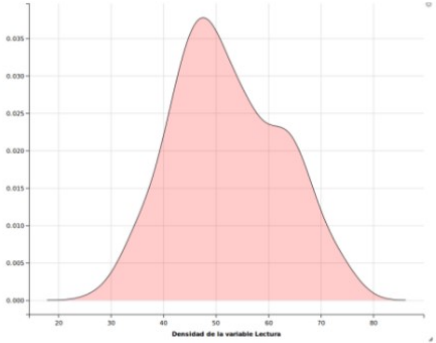


Gráfico de Barras Apiladas

Presentación de datos (Gráficos estadísticos)

- **Gráficos para variables cuantitativos**





Medidas de tendencia central

Media o Promedio

Se calcula tomando la suma agregada de los valores dividida por un conteo de esos valores. La media es sensible a los valores atípicos en los datos.

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Valor atípico

Un valor atípico es el valor de un conjunto o columna que es altamente desviado de los muchos otros valores en los mismos datos; Suele tener valores muy altos o bajos.

Mediana

Dado que las observaciones en una muestra son x1, x2, . . . , xn, acomodadas en orden de magnitud creciente, la mediana es.

$$m_e = \begin{cases} x_{(n+1)/2}, & \text{si } n \text{ es impar} \\ \frac{1}{2}\left(x_{n/2} + x_{n/2+1}\right), & \text{si } n \text{ es par} \end{cases}$$

Moda

La moda es el valor que ocurre con mayor frecuencia en los datos

Medidas de dispersión

Varianza

Es la media de las desviaciones cuadradas de la media. La dimensión de la varianza es el cuadrado de los valores reales.

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

Desviación estandar

Esta es la raíz cuadrada de la varianza. Al aplicar la raíz cuadrada en la varianza, medimos la dispersión con respecto a la variable original.

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

Medidas de dispersión

El coeficiente de variación

El C.V. es la relación entre la desviación típica de una muestra y su media.

$$CV = \frac{S_X}{|\bar{x}|}$$

Rango

El Rango o recorrido es el intervalo entre el valor máximo y el valor mínimo.

$$R = X_{\max} - X_{\min}$$

Medidas de posición

Cuantiles

Estos son simplemente fragmentos idénticos de los datos. Los cuantiles cubren percentiles, deciles, cuartiles, etc. Estas medidas se calculan después de organizar los datos en orden ascendente.

Percentiles:

Son los valores que dividen a la distribución en 100 partes iguales, cada una de las cuales engloba el 1% de las observaciones.

Déciles:

Son los valores de la variable que dividen a la distribución en las partes iguales, cada una de las cuales engloba el 10% de los datos.

Rango intercuartil

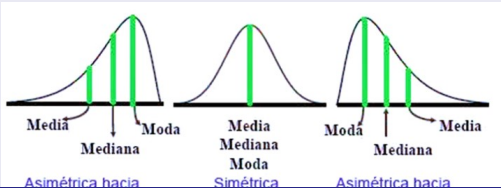
Rango intercuartil Esta es la diferencia entre el tercer cuartil y el primer cuartil. Es eficaz para identificar valores atípicos en los datos. El rango intercuartil describe el 50 por ciento medio de los puntos de datos.

Asimetría y Curtosis

Las medidas de distribución nos permiten identificar la forma en que se separan o aglomeran los valores de acuerdo a su representación gráfica. Estas medidas describen la manera como los datos tienden a reunirse de acuerdo con la frecuencia con que se hallen dentro de la información.

Asimetría

El coeficiente de asimetría mide la asimetría de una distribución. Se basa en la noción del momento de la distribución. Este coeficiente es una de las medidas de asimetría.



Coeficiente de asimetría de Pearson:

El coeficiente de asimetría de Pearson (Basado en al relación existente entre media, mediana y moda) de un conjunto de observaciones se define como:

$$As = \frac{3(\bar{x} - me)}{S} \qquad \text{As } = \frac{(\bar{x} - mo)}{S}$$

Interpretación

- Si as=0, entonces la distribución es simétrica.
- Si as<0 entonces la distribución es asimétrica negativa.
- Si as>0 entonces la distribución es asimétrica positiva.

Coeficiente de asimetría de Fisher:

El coeficiente de asimetría de Fisher evalúa la proximidad de los datos a su media.

Permite interpretar la forma de la distribución, respecto a ser o no simétrica.

$$As = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{ns^3}$$

Hallamos los coeficientes utilizando las librerias

```
library(e1071)
#Ésta es la definición típica que se utiliza en muchos libros de texto ant
skewness(cancer$radio, type=1)
```

```
## [1] 0.9398934
```

```
#Utilizado en SAS y SPSS.
skewness(cancer$radio, type=2)
```

```
## [1] 0.9423796
```

```
#Utilizado en MINITAB y BMDP.
skewness(cancer$radio, type=3)
```

Coeficiente de asimetría de Pearson:

```
coef_asimetria <- function(x){
  resultado <- 3*(mean(x)-median(x))/sd(x)
  return(resultado)
}
coef_asimetria(cancer$radio)
```

```
## [1] 0.6446776
```

Coeficiente de asimetría de Pearson:

```
library(modeest)
coef_asimetria2 <- function(x){
  resultado <- (mean(x)-mfv(x))/sd(x)
  return(resultado)
}
coef_asimetria2(cancer$radio)
```

```
## [1] 0.50717
```

Coeficiente de asimetría de Fisher:

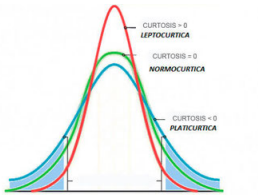
```
asimetria_fisher <- function(x){
  resultado <- sum((x-mean(x))^3)/(length(x)*sd(x)^3)
  return(resultado)
}
asimetria_fisher(cancer$radio)
```

```
## [1] 0.9374168
```

Coeficiente de Curtosis

La curtosis es el grado de concentración de un conjunto de datos (concentración central), con relación a la media aritmética, midiendo el grado de aplastamiento o apuntamiento de la gráfica de la distribución de la variable estadística:

$$K = \frac{\frac{\sum (x_i - \bar{x})^4}{n}}{s^4} - 3$$



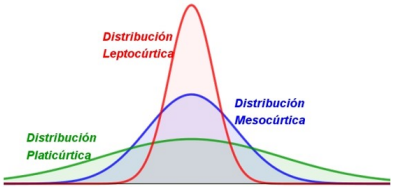
Coeficiente de curtosis de Kelley

$$k = \frac{\frac{1}{2}(q_3 - q_1)}{(d_9 - d_1)}$$

Distribución Platicúrtica
$K < 0.25$

Distribución Mesocúrtica
$K \approx 0.25$

Distribución Leptocúrtica
$K > 0.25$



Hallamos los coeficientes utilizando las librerias

```
library(e1071)
#Ésta es la definición típica que se utiliza en muchos libros de texto ant
kurtosis(cancer$radio, type=1)
```

```
## [1] 0.8275837
```

```
#Utilizado en SAS y SPSS.
kurtosis(cancer$radio, type=2)
```

```
## [1] 0.8455216
```

```
#Utilizado en MINITAB y BMDP.
kurtosis(cancer$radio, type=3)
```

```
coef_curtosis <- function(x){
  resultado <- (sum((x-mean(x))^4)/(length(x)*sd(x)^4))-3
  return(resultado)
}
coef_curtosis(cancer$radio)
```

```
## [1] 0.8141418
```

```
coef_curtosis2 <- function(x){
  resultado <- ((1/2)*(quantile(x,0.75)-quantile(x,0.25)))/
    (quantile(x,0.90)-quantile(x,0.10))
  return(resultado)
}

print(paste("Curtosis:",coef_curtosis2(cancer$radio)))
```

```
## [1] "Curtosis: 0.220064724919094"
```

Análisis Bi-variado y multivariado

Correlación

la correlación indica la fuerza y la dirección de una relación lineal y proporcionalidad entre dos variables estadísticas.

Diagrama de dispersión

Una gráfica de dispersión puede ser usada para datos en la forma de parejas ordenadas de números. El resultado será un montón de puntos dispersos alrededor del plano.

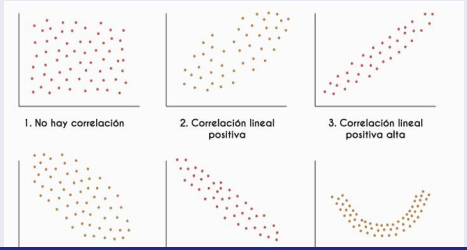
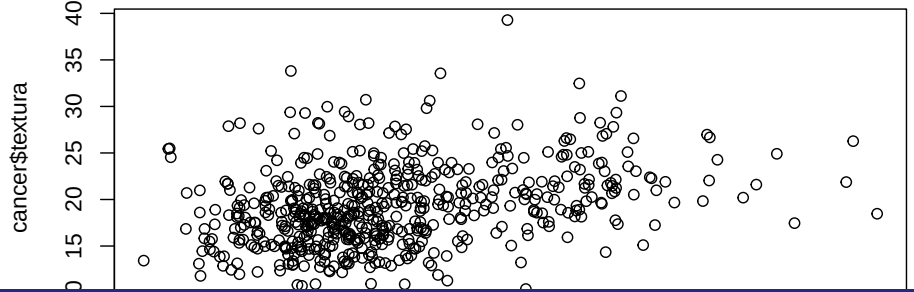
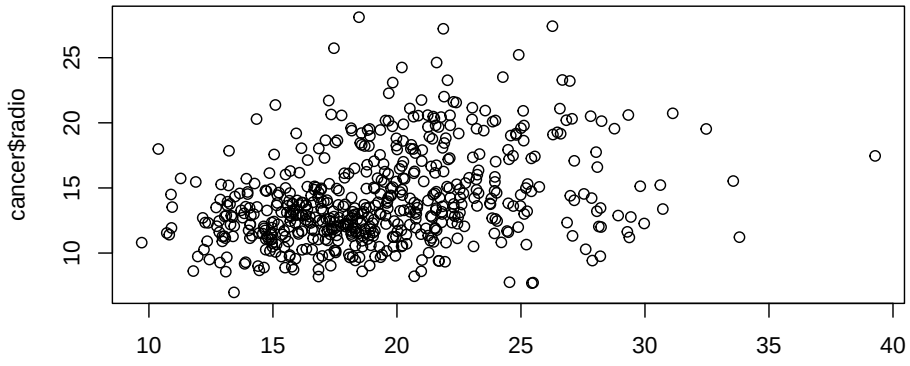


Diagrama de dispersión en R

```
plot(cancer$radio,cancer$textura)
```

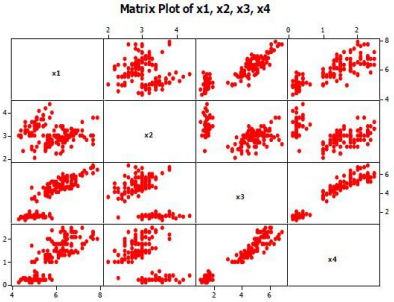


```
plot(cancer$radio ~ cancer$textura)
```



Matriz de dispersión

Es un gráfico que presenta el diagrama de dispersión de varias variables por pares de variables. La matriz gráfica es simétrica es decir la parte superior a la diagonal de la matriz es similar a la parte inferior de la diagonal de la matriz.



Matriz de dispersión en R

```
plot(cancer[,1:6])
```

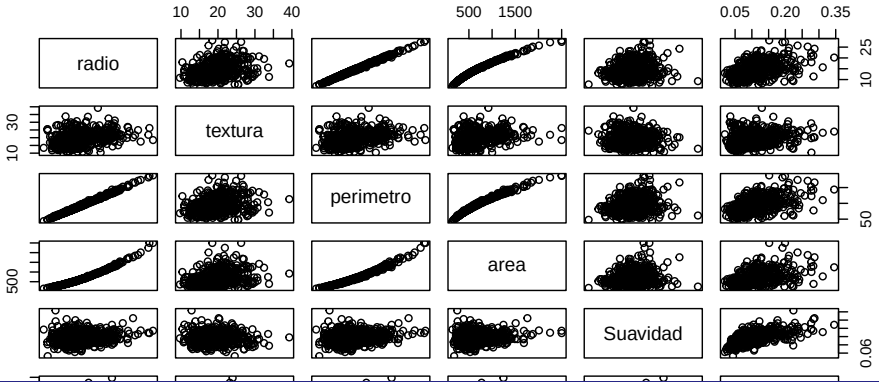
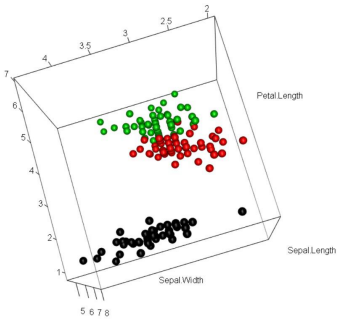
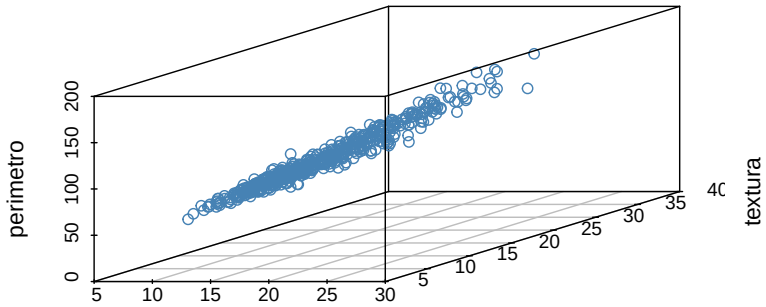


Diagrama de dispersión 3D

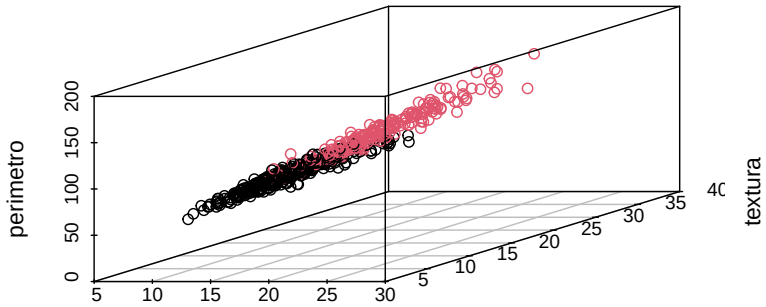
Es un gráfico que permite analizar la relación existente entre tres variables de tipo cuantitativa. En cada uno de los ejes X, Y, Z se ubican los valores de cada una de las variables.



```
library("scatterplot3d")
scatterplot3d(cancer[,1:3], angle = 25, color = "steelblue")
```



```
library("scatterplot3d")
scatterplot3d(cancer[,1:3], angle = 25, color = as.numeric(cancer$diagnosti
```



Medidas bivariadas

Covarianza

La covarianza entre dos variables, s_{xy} , nos indica si la posible relación entre dos variables es directa o inversa.

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

$$Directa : S_{xy} > 0$$

$$Inversa : S_{xy} < 0$$

$$Incorreladas : S_{xy} = 0$$

```
cov(cancer$radio, cancer$textura)
```

```
## [1] 4.907582
```

```
mi_covarianza <- function(x,y){
  resultado <- sum((x-mean(x))*(y-mean(y)))/(length(x)-1)
  return(resultado)
}
mi_covarianza(cancer$radio, cancer$textura)
```

```
## [1] 4.907582
```

Matriz de Varianzas y covarianzas

Es una representación ordenada de las varianzas y las covarianzas entre las variables.

$$[S_{X,X_i}] = \begin{bmatrix} S^2_{X_1} & S_{X_1X_2} \\ S_{X_2X_1} & S^2_{X_2} \end{bmatrix}$$

```
cov(cancer[,1:5])
```

```
##                  radio       textura    perimetro          area       Suav
## radio      1.241892e+01   4.907581564 8.544714e+01 1.224483e+03  0.008454
## textura    4.907582e+00  18.498908679 3.443976e+01 4.859938e+02 -0.001414
## perimetro  8.544714e+01  34.439759167 5.904405e+02 8.435772e+03  0.070830
## area       1.224483e+03 485.993786656 8.435772e+03 1.238436e+05  0.876178
## Suavidad   8.454460e-03  -0.001414779 7.083607e-02 8.761781e-01  0.000197
```

Coeficiente de correlación de Pearson (r)

El Coeficiente de Correlación de Pearson es una medida de la correspondencia o relación lineal entre dos variables cuantitativas aleatorias. Se puede definir tambien como un índice utilizado para medir el grado de relación lineal que tienen dos variables, ambas cuantitativas.

$$r = \frac{S_{xy}}{S_x S_y}$$

- Si r está cercano a 1; entonces X y Y tienen correlación lineal positiva fuerte
- Si r está cercano a -1; entonces X y Y tienen correlación lineal negativa fuerte
- Si r está cercano a 0; entonces X y Y no están correlacionadas linealmente, o es muy débil

```
cor(cancer$radio, cancer$textura)
```

```
## [1] 0.3237819
```

Mi propia función

```
mi_correlacion <- function(x,y){
  covarianza <- sum((x-mean(x))*(y-mean(y)))/(length(x)-1)
  resultado <- covarianza/(sd(x)*sd(y))
  return(resultado)
}
mi_correlacion(cancer$radio, cancer$textura)
```

```
## [1] 0.3237819
```

Matriz de correlaciones

Es una matriz simétrica. Los valores en la diagonal principal son iguales a 1

$$[\mathrm{r_{ij}}] = \begin{bmatrix} \mathrm{r}_{1,1} & \mathrm{r}_{1,2} \\ \mathrm{r}_{2,1} & \mathrm{r}_{2,2} \end{bmatrix}$$

Matriz en r

```
cor(cancer[,1:5])
```

```
##                radio     textura perimetro      area    Suavidad
## radio      1.0000000  0.32378189 0.9978553 0.9873572  0.17058119
## textura    0.3237819  1.00000000 0.3295331 0.3210857 -0.02338852
## perimetro  0.9978553  0.32953306 1.0000000 0.9865068  0.20727816
## area       0.9873572  0.32108570 0.9865068 1.0000000  0.17702838
## Suavidad   0.1705812 -0.02338852 0.2072782 0.1770284  1.00000000
```

coef. de correlación de Spearman, ρ (rho)

Se basa en los rangos de los datos en lugar de hacerlo en los valores reales. Resulta apropiada para datos ordinales, o los de intervalo que no satisfagan el supuesto de normalidad.

$$\rho = 1 - \frac{6 \sum D^2}{N\left(N^2 - 1\right)}$$

```r
cor(cancer[,1:5], method = "spearman")
```

```
##                radio    textura perimetro      area   Suavidad
## radio      1.0000000 0.34095627 0.9978017 0.9996020 0.14850987
## textura    0.3409563 1.00000000 0.3481419 0.3441451 0.02464927
## perimetro  0.9978017 0.34814189 1.0000000 0.9970683 0.18292299
## area       0.9996020 0.34414508 0.9970683 1.0000000 0.13805304
## Suavidad   0.1485099 0.02464927 0.1829230 0.1380530 1.00000000
```

coeficiente de correlación por rangos de Kendall

Es una medida no paramétrica de asociación para variables ordinales o de rangos que tiene en consideración los empates

$$\hat{\tau} = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$

```r
cor(cancer[,1:5], method = "kendall")
```

```
##                radio    textura perimetro       area   Suavidad
## radio      1.00000000 0.22915938 0.9633203 0.98556490 0.09954914
## textura    0.22915938 1.00000000 0.2343527 0.23082876 0.01713488
## perimetro 0.96332028 0.23435270 1.0000000 0.95696526 0.12243357
## area       0.98556490 0.23082876 0.9569653 1.00000000 0.09254106
## Suavidad   0.09954914 0.01713488 0.1224336 0.09254106 1.00000000
```

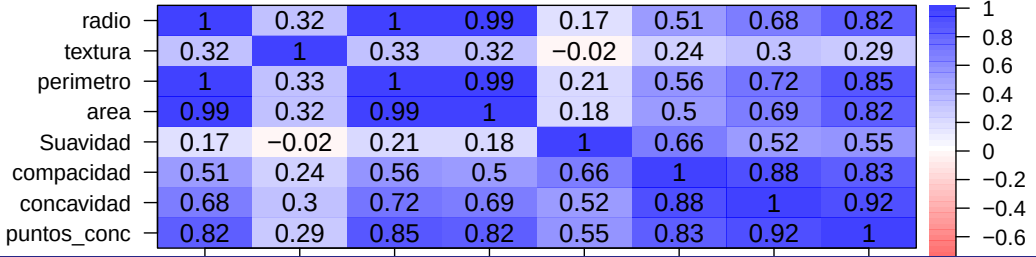
Análisis exploratorio de datos

```
library(psych)
cor(cancer[,1:5])
```

```
##                radio     textura perimetro      area    Suavidad
## radio      1.0000000  0.32378189 0.9978553 0.9873572  0.17058119
## textura    0.3237819  1.00000000 0.3295331 0.3210857 -0.02338852
## perimetro  0.9978553  0.32953306 1.0000000 0.9865068  0.20727816
## area       0.9873572  0.32108570 0.9865068 1.0000000  0.17702838
## Suavidad   0.1705812 -0.02338852 0.2072782 0.1770284  1.00000000
```

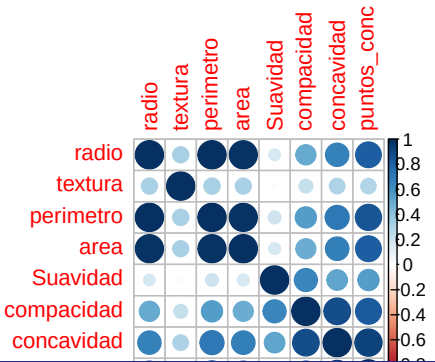
Análisis con la librería psych

```
library(psych)
cor.plot(cor(cancer[,1:8]))
```



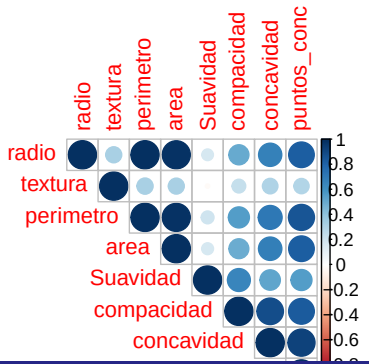
Análisis con la librería corrplot

```
library(corrplot)
corrplot(cor(cancer[,1:8]))
```



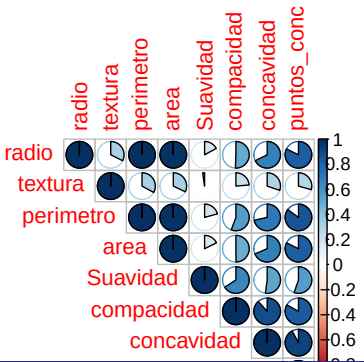
Análisis con la librería corrplot

```
library(corrplot)
corrplot(cor(cancer[,1:8]), type="upper")
```



Análisis con la librería corrplot

```
library(corrplot)
corrplot(cor(cancer[,1:8]), type="upper", method = "pie")
```



Análisis con la librería PerformanceAnalytics

```
library(PerformanceAnalytics)
chart.Correlation(cancer[,1:5])
```

```
## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter

## Warning in par(usr): argument 1 does not name a graphical parameter
```

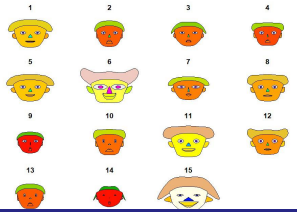
Análisis descriptivo detallado

```
library(Hmisc)
describe(cancer)
```

```
## cancer
##
##  16  Variables      569  Observations
## ---------------------------------------------------------------------------
## radio
##        n  missing distinct     Info     Mean      Gmd      .05      .10
##      569        0      456        1    14.13    3.848    9.529   10.260
##      .25      .50      .75      .90      .95
##   11.700   13.370   15.780   19.530   20.576
##
```

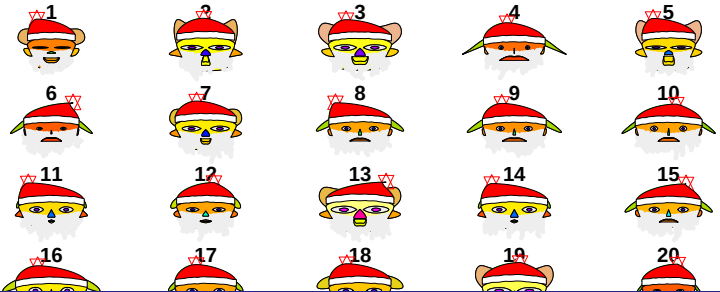
Las caras de Chernoff

En un análisis multivariado se quiere visualizar los datos en una dimensión baja. La presentación numérica de la estructura de datos usando coordenadas por lo tanto puede ser de a lo más en tres dimensiones. El tamaño de los elementos como las pupilas, ojos, cabellos, etc, son asignados a ciertas variables. La idea de usar caras proviene de Chernoff (1973) y ha sido desarrollado por Bernhard Flury.
https://www.rdocumentation.org/packages/aplpack/versions/1.3.3/topics/faces



Caras de Chernoff

```
library(aplpack)
faces(cancer[1:20,1:5],face.type=2)
```



Gráfica de estrellas

```
library(aplpack)
stars(cancer[1:20,1:5])
```

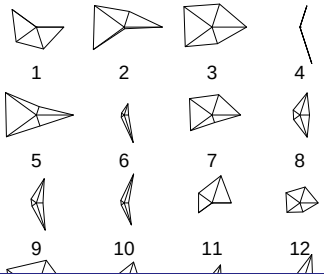
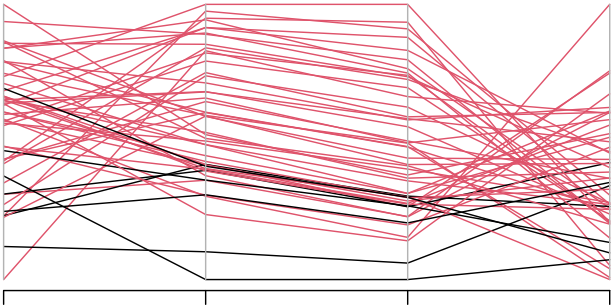


Gráfico de coordenadas paralelas

Las coordenadas paralelas son una forma común de visualizar y analizar conjuntos de datos de alta dimensión .

Para mostrar un conjunto de puntos en un espacio n-dimensional , se dibuja un telón de fondo que consta de n líneas paralelas , típicamente verticales e igualmente espaciadas. Un punto en el espacio n-dimensional se representa como una polilínea con vértices en los ejes paralelos; la posición del vértice en el i-ésimo eje corresponde a la i-ésima coordenada del punto.

```
library(MASS)
parcoord(cancer[1:50,2:5], col=as.numeric(cancer$diagnostico))
```



Valores Outliers (Valores atípicos)

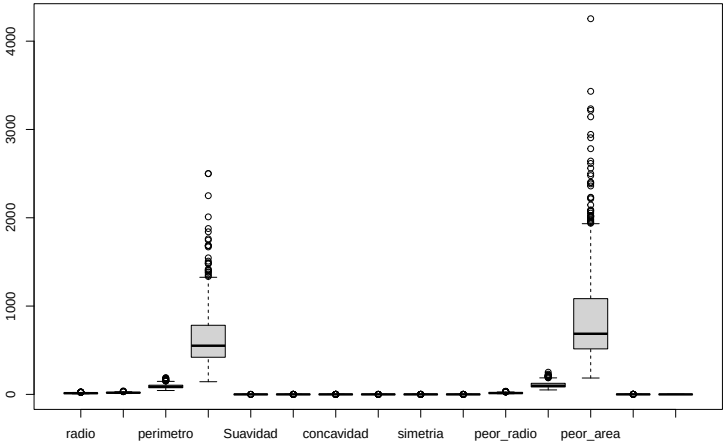
valores Outliers

- Los valores atípicos se definen como valores de una variable que están a una "distancia anormal" de otros valores en un conjunto de datos: extremadamente altos o extremadamente bajos.
- No siempre es fácil decir si un valor es un valor atípico, ya que puede depender del conocimiento del dominio.
- Los valores atípicos extremos pueden causar problemas y pueden distorsionar los modelos planteados
- Hay muchas técnicas gráficas y estadísticas que se pueden utilizar para detectar valores atípicos.

valores Outliers

- Los valores atípicos a veces se deben a errores de entrada de datos, por ejemplo, Cuando se agrega un cero o dos extra al valor por error.
- En otros casos, se puede identificar un patrón de valores atípicos que puede sugerir cómo manejarlos.
- En general, no es una buena idea simplemente eliminar los valores atípicos sin más investigación.

Un *outlier* es una observación que se desvía tanto de las otras observaciones como para crear la sospecha de que fue generado por un mecanismo diferente.


4000
3000
2000
1000
0
radio
perimetro
Suavidad
concavidad
simetria
peor_radio
peor_area

Estandarización de variables

Estandarización de variables

La estandarización es el proceso de poner diferentes variables en la misma escala. Este proceso le permite comparar puntuaciones entre diferentes tipos de variables. Normalmente, para estandarizar las variables, se calcula la media y la desviación estándar de una variable. Luego, para cada valor observado de la variable, resta la media y divide por la desviación estándar.

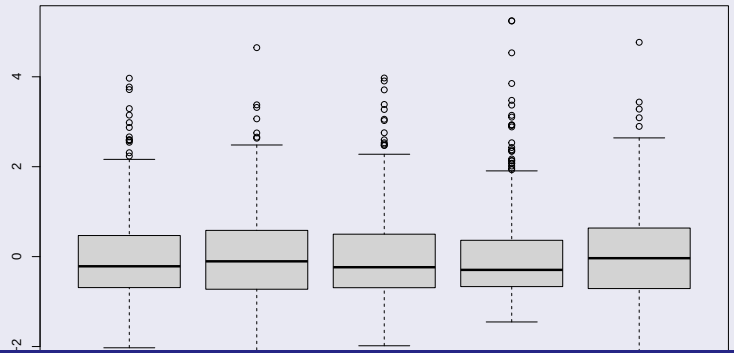
$$z = \frac{x_i - \mu}{\sigma}$$

```
#Usando la función scale
cancer_e1<-scale(cancer[,1:5])
head(cancer_e1)
```

```
##            radio    textura  perimetro       area   Suavidad
## [1,]  1.0960995 -2.0715123  1.2688173  0.9835095  1.5670875
## [2,]  1.8282120 -0.3533215  1.6844726  1.9070303 -0.8262354
## [3,]  1.5784992  0.4557859  1.5651260  1.5575132  0.9413821
## [4,] -0.7682333  0.2535091 -0.5921661 -0.7637917  3.2806668
## [5,]  1.7487579 -1.1508038  1.7750113  1.8246238  0.2801253
## [6,] -0.4759559 -0.8346009 -0.3868077 -0.5052059  2.2354545
```

Graficamos nuevamente

```
boxplot(cancer_e1[,1:5])
```



Manejo de valores outtliers

outtliers univariados:

- Se puede Considerar outliers valores que al aplicar $\frac{|x-\bar{x}|}{s} > k$
- Donde k es 2 ó 3 si consideramos normalidad.

Tambien:

- Considerando el Boxplot(Tukey, 1977), se considera outlier a los valores que caen fuera de este intervalo. $(Q_1 - 3 \times IQR, Q_3 + 3 \times IQR)$
  http://www.physics.csbsju.edu/stats/box2.html

```
#Observamos las filas donde hay valores outliers
rownames(cancer[abs(cancer_e1[,1])>2,])
```

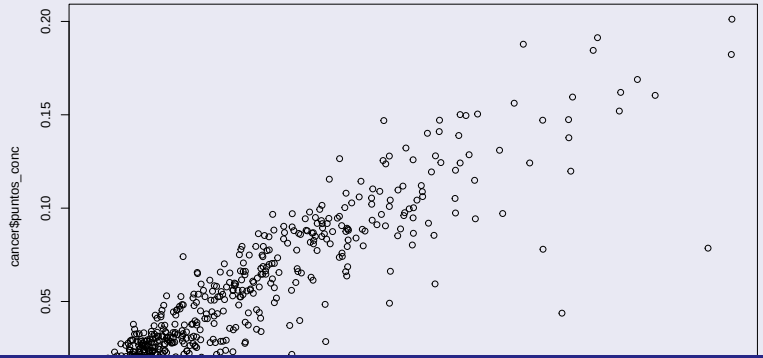
```
##  [1] "83"  "102" "109" "123" "165" "181" "203" "213" "237" "273" "340"
## [13] "369" "370" "373" "394" "462" "504" "522" "565"
```

```
rownames(cancer[abs(cancer_e1[,1])>3,])
```

```
## [1] "83"  "181" "213" "353" "462"
```

Outliers bivariado

```
plot(cancer$concavidad, cancer$puntos_conc)
```



Outliers Multivariados

Consideremos un conjunto de datos C con p variables y n instancias.

- Supongamos que también conocemos las clases a las cuales pertenecen cada una de las instancias.
- El objetivo es detectar todas las instancias que parecen ser no usuales, estas serán los outliers multivariados.
- Uno podría pensar que los outliers multivariados pueden ser detectados basados en los outliers univariados en cada una de las variables, pero no es cierto.
- Una instancia puede tener valores que son outliers en varias variables, pero la instancia como todo podrá no ser un outlier multivariado.

Outliers Multivariados

Métodos para detectar Outliers Multivariados * Métodos basados en estadística robusta * Métodos basados en clustering * Métodos basados en distancia, y * Métodos basados en densidad local.

Distancia de Mahalanobis

- Sea x una observación de un conjunto de datos multivariado consistente de n observaciones y p variables.
- Sea $\bar{x}$ el centroide del conjunto de datos, el cual es un vector p dimensional que tiene como componentes la media de cada variable.
- Sea $\tilde{x}$ la matriz del conjunto de datos original con columnas centradas por sus medias.

-Entonces la matriz $S = \frac{1}{n-1}\tilde{x}'\tilde{x}$ de orden $p \times p$ representa la matriz de covarianza de p variables. - La versión multivariada de la ecuación anterior es

$$D^2(x, \bar{x}) = (x - \bar{x})' S^{-1} (x - \bar{x}) > k$$

donde D^2 es llamada la distancia de Mahalanobis cuadrada estimada desde x al centroide del conjunto de datos. - Una observación con una distancia de Mahalanobis grande puede ser considerada como un outlier.

Outlier basado en densidad local

En este tipo de outliers la densidad de los vecinos de una distancia juega un crucial rol Además, una instancia no es explícitamente clasificada como outlier ó no-outlier; lo que se hace es calcular para cada instancia un factor de outlier local (LOF) y esta medida da una idea de que tan fuerte una instancia puede ser un outlier.
Practicamos...