Análisis de asociación en R

Sesión 1

Natalie Julian - www.nataliejulian.com

Estadística UC y Data Scientist en Zippedi Inc.

Introducción

Muchas veces al tener una base de datos, no sólo es relevante analizar el *target* (variable de interés), también es necesario considerar la relación de dicha variable con las variables explicativas, y a la vez, la relación entre las variables explicativas.

Otra consideración importante es el tipo de variables a trabajar, no es lo mismo analizar el tipo de relación entre un par de variables de tipo numérica continua a un par de variables de tipo categórica.

En las siguientes láminas, veremos conceptos importantes para analizar la relación entre distinto tipo de variables.

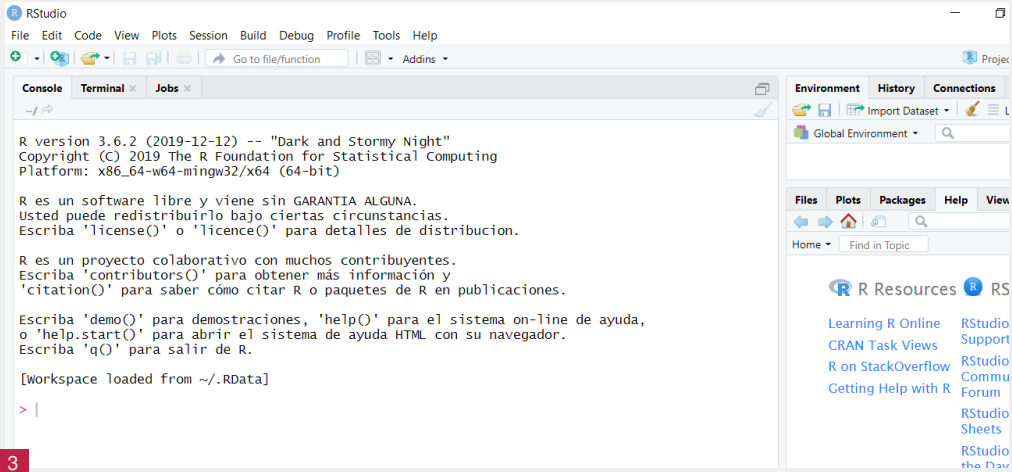
breast-cancer

La base de datos `breast-cancer.csv` contiene información de imágenes digitalizadas de tumores de mama. Interesa realizar análisis entre las variables de la data y obtener algunas inferencias iniciales de la data.

https://www.kaggle.com/uciml/breast-cancer-wisconsin-dataDescripción de las variables aquí

Vista de RStudio

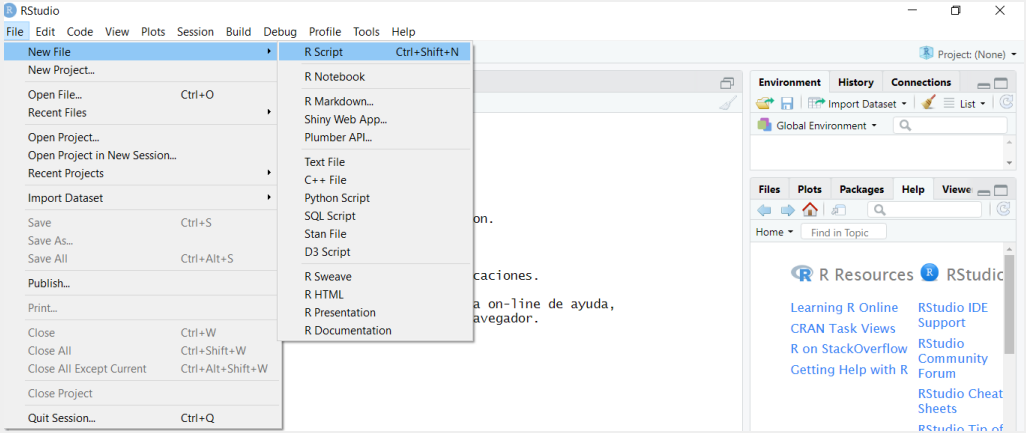
Utilizaremos RStudio, un software donde utilizaremos el lenguaje de programación R.



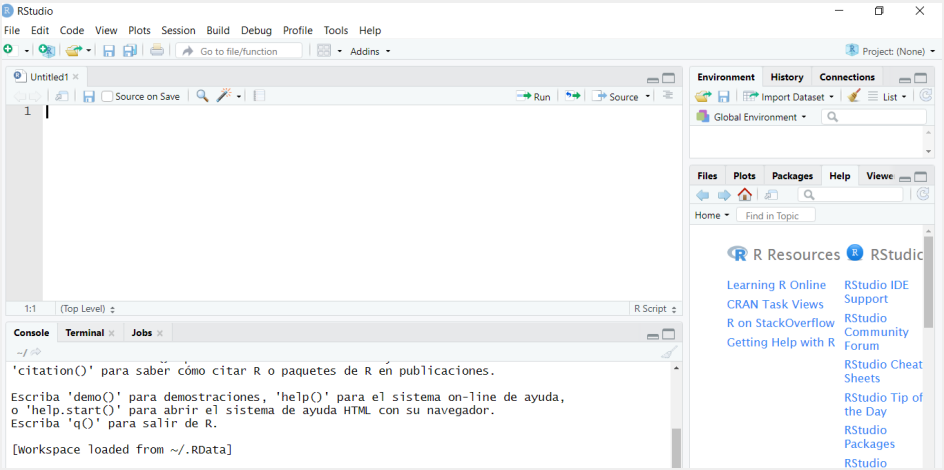
Un script

Al realizar análisis en cualquier software de programación, necesitamos guardar un *script* o *código* con las funciones utilizadas, de manera tal de poder compartir nuestro trabajo o replicarlo posteriormente.

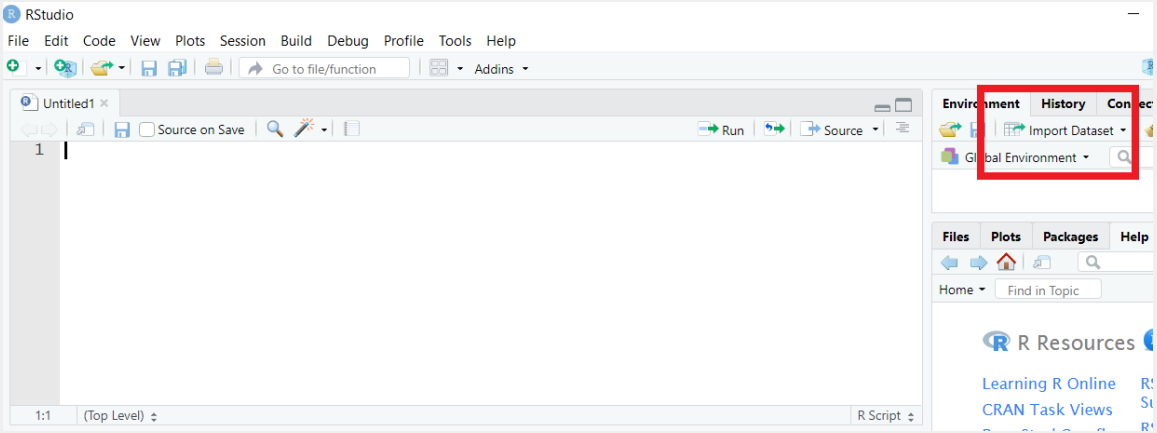
Creando un script en R



Vista del script



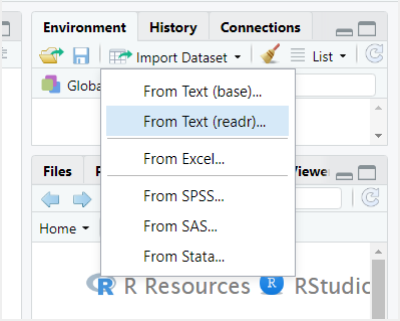
Cargando una data en RStudio



Cargando la data breast-cancer.csv

Existen distinto tipo de bases de datos, algunas en formato *.txt*, *.xlsx*, *.dta*, *.csv*, entre otros.

Recordemos que la data `breast-cancer` se encuentra en formato *csv*:



Una vez cargada la data, podemos empezar a trabajar con ella en R.

Funciones aplicables a datas en R

Función	Descripción
`View()`	Muestra en una ventana aparte la data
`head()`	Muestra los primeros registros de la data
`tail()`	Muestra los últimos registros de la data
`dim()`	Dimensión de una data
`nrow()`	Número de filas (registros)
`ncol()`	Número de columnas (variables)
`names()`	Nombre de las columnas (variables)
`str()`	Muestra el tipo de variables que contiene la data
`summary()`	Resumen estadístico de las variables

Matriz de variables numéricas

```
#Conclusiones inciales de la data:

#Variable id corresponde a un identificador

table(breast_cancer$id)   #Cada id aparece una y solo una vez, no precisa análisis

#Variable diagnosis indica el diagnostico del tumor, es la unica variable cualitativa
#de la base de datos

table(breast_cancer$diagnosis)

#Crearemos una matriz de las variables cuantitativas:

numvar<-breast_cancer[,-c(1,2)]   #Quitamos las dos primeras variables
```

Covarianza

La covarianza es un valor que indica el grado de variación conjunta entre dos variables aleatorias. Cuando **X** aumenta, ¿cuánto y cómo cambia **Y**?

$$\sigma_{XY} = E[(X - \mu_x)(Y - \mu_y)] = \frac{\sum_{i=1}^{n} (x_i - \mu_x)(y_i - \mu_y)}{n}$$

Interpretación:

- Si una de las variables aumenta y la otra disminuye, existe una relación negativa ($\sigma_{XY} < 0$).
- Si una de las variables aumenta y la otra aumenta, existe una relación positiva ($\sigma_{XY} > 0$).
- No se puede establecer ninguno de los dos patrones anteriores ($\sigma_{XY} = 0$).

Matriz de varianzas-covarianzas

```
cov(numvar)    #Matriz de varianzas-covarianzas

head(cov(numvar))
                  radius_mean  texture_mean perimeter_mean    area_mean smoothness_mean
radius_mean      1.241892e+01   4.907581564   8.544714e+01 1.224483e+03    0.0084544598
texture_mean     4.907582e+00  18.498908679   3.443976e+01 4.859938e+02   -0.0014147788
perimeter_mean   8.544714e+01  34.439759167   5.904405e+02 8.435772e+03    0.0708360652
area_mean        1.224483e+03 485.993786656   8.435772e+03 1.238436e+05    0.8761781263
smoothness_mean  8.454460e-03  -0.001414779   7.083607e-02 8.761781e-01    0.0001977997
compactness_mean 9.419706e-02   0.053766806   7.147141e-01 9.264931e+00    0.0004895739

                 compactness_mean concavity_mean concave points_mean symmetry_mean
radius_mean          0.0941970568   1.901276e-01        1.124751e-01  0.0142731729
texture_mean         0.0537668058   1.036923e-01        4.897693e-02  0.0084188757
perimeter_mean       0.7147141251   1.387234e+00        8.023604e-01  0.1219215828
area_mean            9.2649307889   1.924492e+01        1.124196e+01  1.4595958865
smoothness_mean      0.0004895739   5.852428e-04        3.021671e-04  0.0002150545
compactness_mean     0.0027891874   3.718135e-03        1.703233e-03  0.0008725181

                 fractal_dimension_mean    radius_se    texture_se perimeter_se
radius_mean                -7.753706e-03  0.663650325 -1.891886e-01 4.803550e+00
texture_mean               -2.321158e-03  0.329037393  9.166951e-01 2.449449e+00
perimeter_mean             -4.485888e-02  4.661401017 -1.162988e+00 3.405303e+01
area_mean                  -7.034264e-01 71.490944748 -1.286717e+01 5.170100e+02
smoothness_mean             5.806859e-05  0.001175770  5.307283e-04 8.419558e-03
compactness_mean            2.108131e-04  0.007285822  1.346135e-03 5.861195e-02
```

Problemas de la Covarianza

- La covarianza depende de las escalas de las variables, por lo que, es difícil establecer un grado de asociación lineal entre las variables sólo observando la covarianza.
- ¿Opciones?: El Coeficiente de Correlación de Pearson

Correlación de Pearson

Una medida para analizar la relación entre variables cuantitativas continuas es *el Coeficiente de correlación de Pearson*. Se dice que es la versión estandarizada de la Covarianza.

Sean dos variables cuantitativas continuas **X** y **Y**, el coeficiente de correlación de Pearson $\rho_{X,Y}$ consiste en:

$$\rho_{X,Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

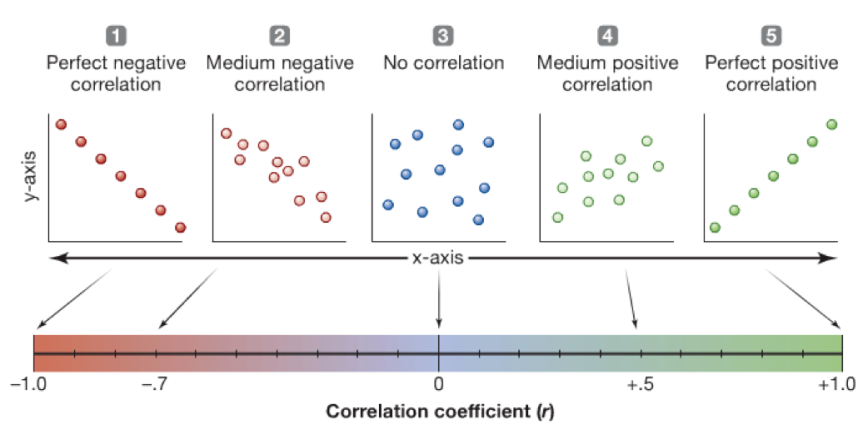
Donde,

σ_{XY} es la covarianza entre **X** e **Y**
σ_X es la desviación estándar de **X**
σ_Y es la desviación estándar de **Y**

Interpretación $\rho_{X,Y}$

- Si $\rho_{X,Y} = 1$ indica una relación lineal positiva perfecta entre las variables, es decir, al graficar ambas variables, puede dibujarse una recta con pendiente positiva y pasar por todos los puntos.
- Si $0 < \rho_{X,Y} < 1$ existe una asociación lineal positiva, el grado de dicha asociación aumenta a medida que se acerca al valor 1.
- Si $\rho_{X,Y} = 0$ no se puede establecer una asociación lineal de ningún tipo, ni positiva ni negativa, lo que no quiere decir que las variables no se encuentren relacionadas de otra forma.
- Si $\rho_{X,Y} = -1$ indica una relación lineal negativa perfecta entre las variables, es decir, al graficar ambas variables, puede dibujarse una recta con pendiente negativa y pasar por todos los puntos.
- Si $-1 < \rho_{X,Y} < 0$ existe una asociación lineal negativa, el grado de dicha asociación aumenta a medida que se acerca al valor -1.

Relaciones en Pearson



Matriz de correlación de Pearson

```
cor(numvar)    #Matriz de correlación de Pearson

head(cor(numvar))
                radius_mean texture_mean perimeter_mean area_mean smoothness_mean
radius_mean       1.0000000   0.32378189      0.9978553 0.9873572      0.17058119
texture_mean      0.3237819   1.00000000      0.3295331 0.3210857     -0.02338852
perimeter_mean    0.9978553   0.32953306      1.0000000 0.9865068      0.20727816
area_mean         0.9873572   0.32108570      0.9865068 1.0000000      0.17702838
smoothness_mean   0.1705812  -0.02338852      0.2072782 0.1770284      1.00000000
compactness_mean  0.5061236   0.23670222      0.5569362 0.4985017      0.65912322

                compactness_mean concavity_mean concave points_mean symmetry_mean
radius_mean            0.5061236      0.6767636           0.8225285    0.14774124
texture_mean           0.2367022      0.3024178           0.2934641    0.07140098
perimeter_mean         0.5569362      0.7161357           0.8509770    0.18302721
area_mean              0.4985017      0.6859828           0.8232689    0.15129308
smoothness_mean        0.6591232      0.5219838           0.5536952    0.55777479
compactness_mean       1.0000000      0.8831207           0.8311350    0.60264105

                fractal_dimension_mean radius_se  texture_se perimeter_se   area_se
radius_mean                -0.31163083 0.6790904 -0.09731744    0.6741716 0.7358637
texture_mean               -0.07643718 0.2758687  0.38635762    0.2816731 0.2598450
perimeter_mean             -0.26147691 0.6917650 -0.08676108    0.6931349 0.7449827
area_mean                  -0.28310981 0.7325622 -0.06628021    0.7266283 0.8000859
smoothness_mean             0.58479200 0.3014671  0.06840645    0.2960919 0.2465524
compactness_mean            0.56536866 0.4974734  0.04620483    0.5489053 0.4556529
```

Relación entre un par de variables

Supongamos que nos interesa estudiar la relación entre las variables `radius_mean` y `perimeter_mean`. Recordemos lo que miden:

- `radius_mean`: mean of distances from center to points on the perimeter
- `perimeter_mean`: mean size of the core tumor

Intuitivamente tiene sentido que estas variables esten asociadas, a mayor tamaño del tumor central, naturalmente exisitiría mayor distancia entre el centro y los puntos del perímetro. Veamos qué obtenemos en R:

Cálculos en R

```
cov(numvar$radius_mean,numvar$perimeter_mean)
[1] 85.44714

#La covarianza entre las variables es positiva, por lo cual, existe
# una asociacion  positiva, si una aumenta, la otra tambien

#No se puede comentar sobre el grado de asociacion con la covarianza,
# pero si con la correlacion de Pearson:

cor(numvar$radius_mean,numvar$perimeter_mean)
[1] 0.9978553

#La correlacion es practicamente 1, por lo cual, se tendria una
# relacion lineal practicamente perfecta entre ambas variables
# y de tipo positiva
```

Gráfico de dispersión

```
#Grafico de dispersion basico entre radius_mean y perimeter_mean

plot(numvar$radius_mean,numvar$perimeter_mean,
main="Relacion entre radius mean y perimeter mean",xlab="Radius mean",
ylab="Perimeter mean",las=1)
```

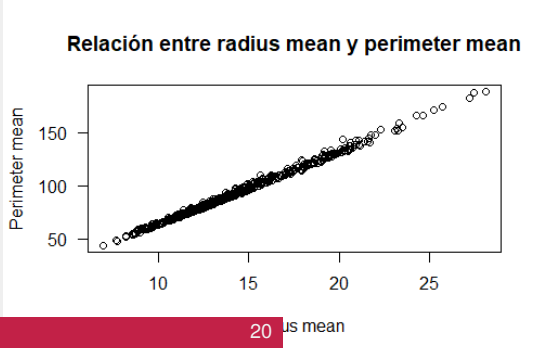
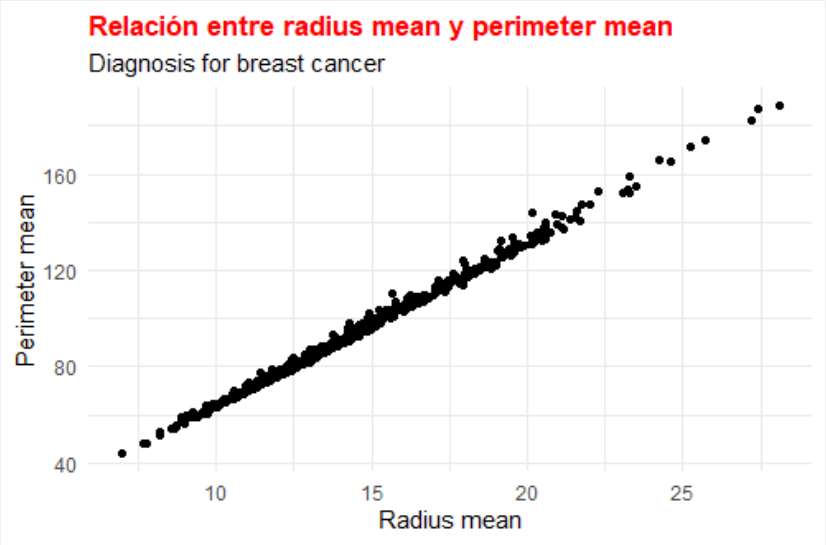


Gráfico de dispersión

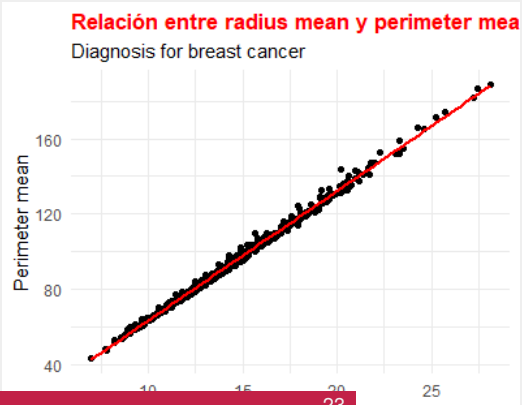
```
#Grafico de dispersion mejorado entre radius_mean y perimeter_mean

#Necesitaremos instalar el paquete ggplot2
install.packages("ggplot2")

#Carga el paquete
library(ggplot2)

graph<-ggplot(numvar, aes(x = radius_mean, y = perimeter_mean))+
geom_point()+ggtitle("Relacion entre radius mean y perimeter mean")+
xlab("Radius mean")+ ylab("Perimeter mean")+
labs(subtitle="Diagnosis for breast cancer")+theme_minimal()+theme(
plot.title = element_text(color = "red", size = 13, face = "bold"))

graph   #Muestra el grafico
```

Resultado


Relación entre radius mean y perimeter mean
Diagnosis for breast cancer
Perimeter mean
160
120
80
40
10
15
20
25
Radius mean

Evidenciar una relación lineal

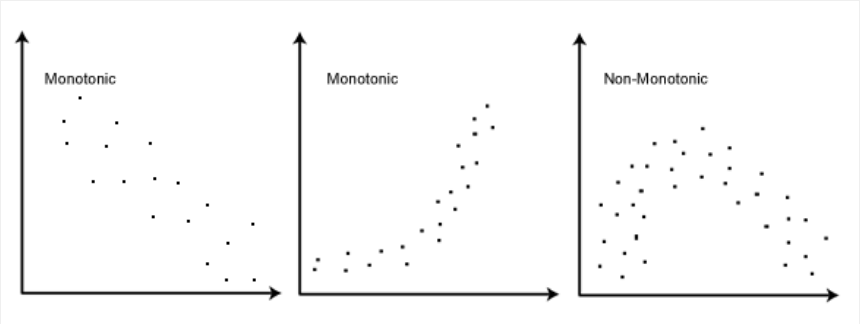
```
#Le agregamos una recta que represente la relacion lineal existente

graph+geom_smooth(method='lm', formula= y~x,col="red")
```



Correlación de Spearman

La correlación de Spearman utilizada para variables cuantitativas continuas u ordinales. ¿Cuándo usarla? Cuando nos interesa estudiar otro tipo de relaciones entre variables, pues, puede existir una correspondencia o relación monotónica entre ambas variables. Si una variable aumenta, la otra puede hacerlo pero no necesariamente que éste aumento sea a un ritmo lineal.

Relaciones en Spearman



Correlación de Spearman en R

En R, se obtiene de la siguiente forma:

```
cor(numvar$radius_mean,numvar$perimeter_mean)
[1] 0.9978553

cor(numvar$radius_mean,numvar$perimeter_mean,method="spearman")
[1] 0.9978553

#Cuando existe una fuerte asociacion lineal,
#Pearson y Spearman entregan resultados coincidentes.
#Pues una relacion lineal es monotonica.
#Pero no toda relacion monotonica es lineal.
```

Concordancia

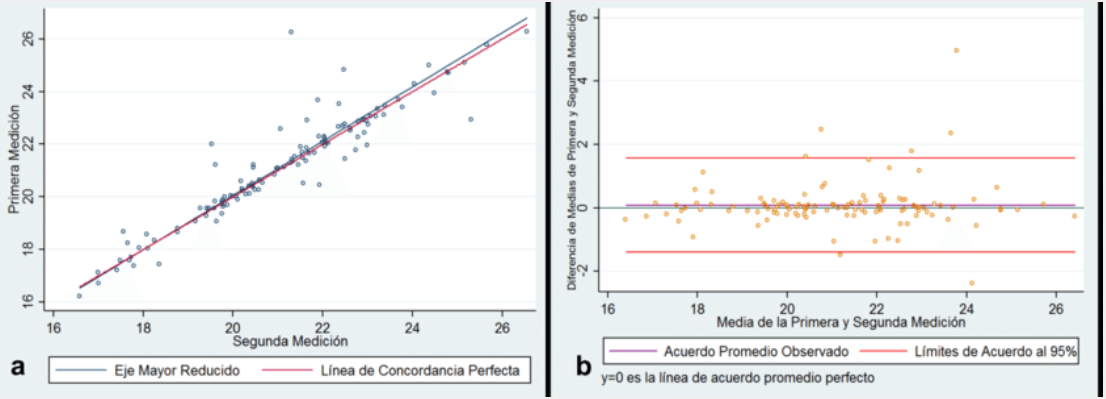
Cuando se desea medir la misma variable, en las mismas muestras o pacientes, con dos métodos, equipos o personas diferentes, para determinar si ambos métodos, equipos o personas producen resultados equivalentes, se realiza análisis de concordancia.

Note que la concordancia no es lo mismo que la correlación, si una de las mediciones tiene un error sistemático, por ejemplo si una de las mediciones tiene sistemáticamente cinco unidades menos que la otra medición, el coeficiente de correlación puede ser muy elevado aunque las diferencias en las mediciones sean importantes, es decir, las mediciones pueden no ser concordantes.

Gráfico de Bland y Altman

El gráfico de Bland y Altman es un gráfico que sirve para medir las diferencias producidas entre dos métodos y evaluar y analizar la concordancia de éstas.

Gráfico


Primera Medición
16
18
20
22
24
26
Segunda Medición
a
Eje Mayor Reducido
Línea de Concordancia Perfecta
Diferencia de Medias de Primera y Segunda Medición
-2
0
2
4
6
Media de la Primera y Segunda Medición
Acuerdo Promedio Observado
Límites de Acuerdo al 95%
b
y=0 es la línea de acuerdo promedio perfecto

Explicación del gráfico

En variables concordantes, se puede observar en el gráfico de dispersión que es posible trazar una recta identidad y pasar por la mayoría de los puntos. Si los puntos se encuentran en su mayoría dentro de las bandas azules, esto se interpreta como que las diferencias entre una medición y la otra son bastante pequeñas. Además, si los puntos se dispersan en torno al eje rojo, esto da un gran indicio de que los métodos son concordantes.

Ejemplo

Suponga que para medir la deformación del núcleo celular promedio se utiliza `compactness_mean`. Suponga que un doctor le dice a usted que la variable `concavity_mean` también pudiera ser un buen método para discernir qué tan deformados están los núcleos celulares. Realice análisis de concordancia para aprobar o rechazar lo que el médico le indica.

Análisis previo

```
metodo_A<-breast_cancer$compactness_mean
metodo_B<-breast_cancer$concavity_mean

summary(metodo_A)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.01938 0.06492 0.09263 0.10434 0.13040 0.34540

summary(metodo_B)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.02956 0.06154 0.08880 0.13070 0.42680

cov(metodo_A,metodo_B)
[1] 0.003718135

cor(metodo_A,metodo_B)
[1] 0.8831207

cor(metodo_A,metodo_B,method="spearman")
[1] 0.8965184
```

Gráfico de dispersión en R

```
ggplot(data = datos, mapping = aes(x = metodo_A, y = metodo_B)) +
  geom_point(color = "black", size = 1) +
  labs(title = "Diagrama de dispersion", x = "metodo A", y = "metodo B") +
  geom_smooth(method = "lm", se = TRUE, color = "blue", lwd = 0.5) +
  geom_abline(intercept = 0, slope = 1, lwd = 0.7, col = "red") +
  theme(axis.line = element_line(colour = "black"),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        panel.background = element_blank()) +
  theme_bw() + theme(plot.title = element_text(hjust = 0.5))
```

Gráfico de dispersión

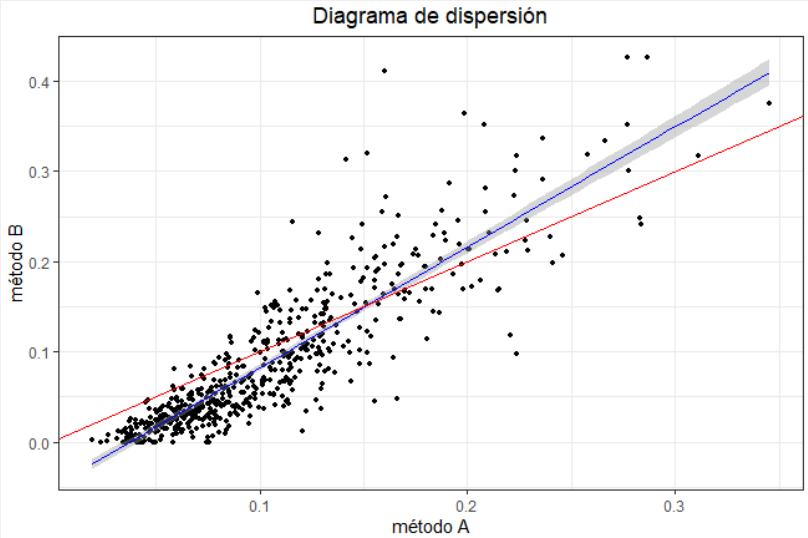
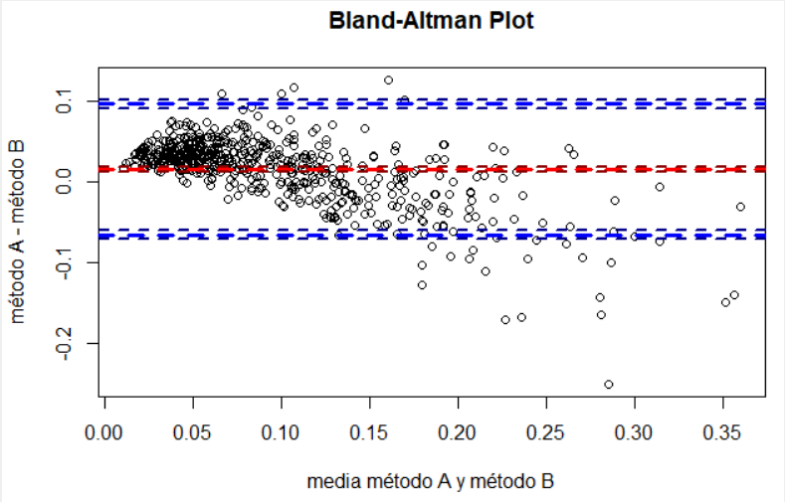


Gráfico de Bland y Altman en R

```
diferencia <- metodo_A - metodo_B
media <- (metodo_A + metodo_B) / 2
porcentaje <- ((diferencia / media) * 100)
datos <- data.frame(metodo_A, metodo_B, diferencia, media, porcentaje)


install.packages("BlandAltmanLeh")   #instala paquete

library(BlandAltmanLeh)
bland.altman.plot(metodo_A, metodo_B, main = "Bland-Altman Plot",
                  xlab = "media metodo A y metodo B",
                  ylab = "metodo A - metodo B", conf.int = .95)
```

Gráfico de Bland y Altman



Verdadero o Falso

1. Todas las variables numéricas son variables cuantitativas.
2. A todo par de variables cuantitativas tiene sentido aplicarles la correlación de Pearson.
3. La covarianza tiene un gran pero: No indica grado o magnitud de la asociación lineal a estudiar.
4. Toda relación lineal es monotónica, pero no toda relación monotónica es lineal.
5. Si la correlación de Pearson entre un par de variables resulta cero, significa que las variables no están asociadas.
6. Los resultados al analizar correlación, covarianza y gráfico de dispersión, son concordantes.
7. Al analizar concordancia de variables no es necesario realizar un análisis previo de cada una de las variables.
8. Suponga que una nutricionista tiene dos métodos para medir la altura de una persona, un método tiene un desfase de 20 centímetros respecto al otro método. Los métodos no son concordantes.
9. El gráfico de Bland y Altman puede utilizarse para analizar cualquier par de variables cuya concordancia sea posible.

Respuestas

1. Todas las variables numéricas son variables cuantitativas. *Falso, hay variables categóricas etiquetadas numéricamente.*
2. A todo par de variables cuantitativas tiene sentido aplicarles la correlación de Pearson. *Falso, la correlación de Pearson adquiere sentido cuando se le aplica a un par de variables cuantitativas continuas*.
3. La covarianza tiene un gran pero: No indica grado o magnitud de la asociación lineal a estudiar. *Verdadero.*
4. Toda relación lineal es monotónica, pero no toda relación monotónica es lineal. *Verdadero.*
5. Si la correlación de Pearson entre un par de variables resulta cero, significa que las variables no están relacionadas. *Falso. Existen distintos tipos de relaciones más allá de la lineal: cuadrática, exponencial, autoregresiva, etcétera.*

Respuestas

6. Los resultados al analizar correlación, covarianza y gráfico de dispersión, son concordantes. *Verdadero, correlación, covarianza y el gráfico de dispersión se utilizan para complementarse, nunca para oponerse.*
7. Al analizar concordancia de variables no es necesario realizar un análisis previo de cada una de las variables. *Falso, siempre es útil realizar análisis previo de las variables, estadísticas de éstas y ver cuánto difieren entre sí.*
8. Suponga que una nutricionista tiene dos métodos para medir la altura de una persona, al realizar un gráfico entre ambas mediciones la pendiente es 1 y el intercepto es 20. Los métodos no son concordantes. *Falso, se requiere más información, más que analizar los valores puntuales del intercepto y la pendiente, es necesario estudiar los intervalos de confianza de éstos.*
9. El gráfico de Bland y Altman puede utilizarse para analizar cualquier par de variables cuya concordancia sea posible. *Falso, en realidad, el gráfico de Bland y Altman tiene un supuesto detrás: La diferencia entre los valores de ambas variables debe cumplir el supuesto de normalidad. Además, este gráfico requiere que sean variables cuantitativas*.