**Áreas que conforman la Estadística**

* Descriptiva → Encargada de la presentación de datos mediante tablas y gráficos y del cálculo de medidas. (Describir)
* Inferencial → Busca estimar características de una población mediante los datos extraídos de una muestra (Inferir)

**Conceptos de población y muestra**

* Población → Conjunto de todas las unidades de análisis con características similares
* Muestra → Subconjunto de la población

**Variable y Dato**

* Variable → Características de interés que se estudia de cada unidad de una población o muestra
* Datos → Valores obtenidos de la toma de variable de cada unidad de análisis

**Escalas de medición**

1. **Cualitativa**

* Nominal → No tienen significado numérico, y no importa el orden (soltero, casado, divorciado, viudo)
* Ordinal → No tienen valor numérico pero sí importa el orden (Oro, plata, bronce)

1. **Cuantitativa**

* Discreta → Números infinitos enteros(11,0,4,12)
* Continua → Números infinitos de no enteros (1.73,1.89,1.54)

***MEDIDAS DE TENDENCIA CENTRAL, POSICIÓN, DISPERSIÓN y FORMA***

**Media**

Promedio aritmético

Variables Cuantitativas

Desventaja: influenciado por valores extremos. Sirve, siempre y cuando no haya valores extremos.

nota <- c(14,18,12,16,14,15,16,18,10,12)  
  
 mean(nota)

## [1] 14.5

**Mediana**

Valor central

Variables Cuantitativas

No es afectada por valores extremos

nota <- c(14,18,12,16,14,15,16,18,10,12)  
  
 median(nota)

## [1] 14.5

**Moda**

Valor que más se repite

Puede no existir o haber más de una moda

library(modeest)

nota <- c(14,18,12,16,14,15,16,18,10,12)

mfv(nota)

## [1] 12 14 16 18

**Medidas de posición**

**Cuartil**

Dividen un grupo de datos ordenados en partes iguales (cuatro quantiles)

*# Primer cuartil*quantile(nota,0.25)

## 25%  
## 12.5

Interpretación: El 25% de las notas llegan hasta 12.5, mientras que el 75% (resto) es mayor a 12.5

**Percentil**

Divide un conjunto ordenado en 100 partes

*# Percentil 60*quantile(nota,0.60)

## 60%  
## 15.4

Interpretación: El 60% de las notas son como máximo 15.4 y el resto mayor a 15.4

#Percentil 80: quinto superior

#Percentil 90: décimo superior

**Medidas de dispersión**

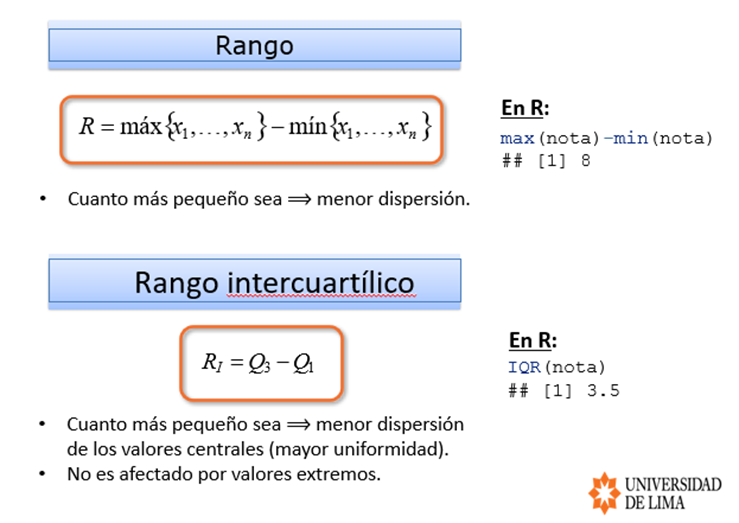
Las medidas de dispersión nos indican el nivel de separación/dispersión en un conjunto de datos.

**Absolutas**

* Rango
* Rango intercuartílico
* Varianza
* Desviación estándar

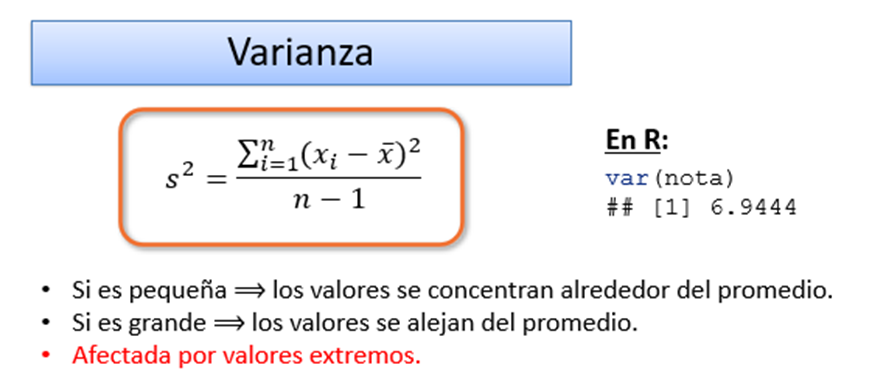
**Relativas**

* Coeficiente de variación

****

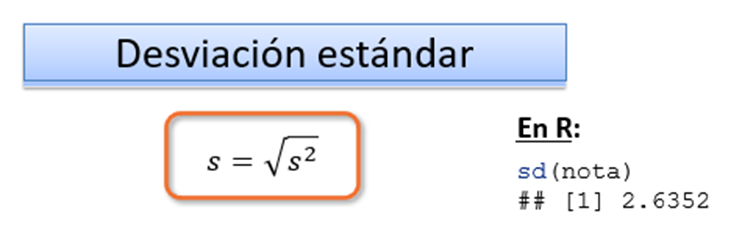
Se trabaja con el “Rango intercuartílico” porque no es afectado pro datos out layer, si no que trabaja con el 50% central.

Rango intercuartílico = Percentil 75 – Percentil 25

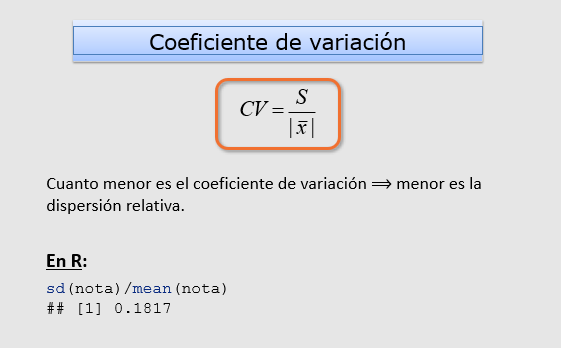


Si es grande, es porque hay mucha dispersión en los datos.

Te da la unidad al cuadrado (ejem. Soles al cuadrado, puntos al cuadrado, etc.)



Te va a dar la unidad de medida original. La varianza te da la unidad al cuadrado.



CV = Desviación Estándar / Valor Absoluto de la Media

**Medidas de forma**

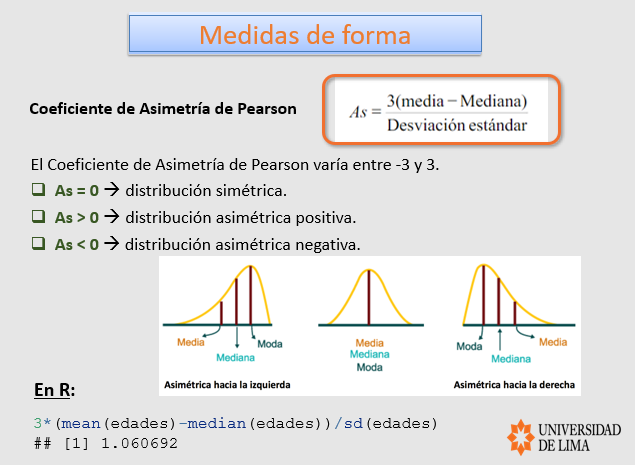
Las medidas de forma proporcionan información sobre el aspecto de la distribución (sin representarla gráficamente)

**Asimetría**

* Coeficiente de asimetría de Pearson
* Coeficiente de asimetría de Fisher

**Curtosis**

* Coeficiente de curtosis
* **Coeficiente de Asimetría de Pearson**

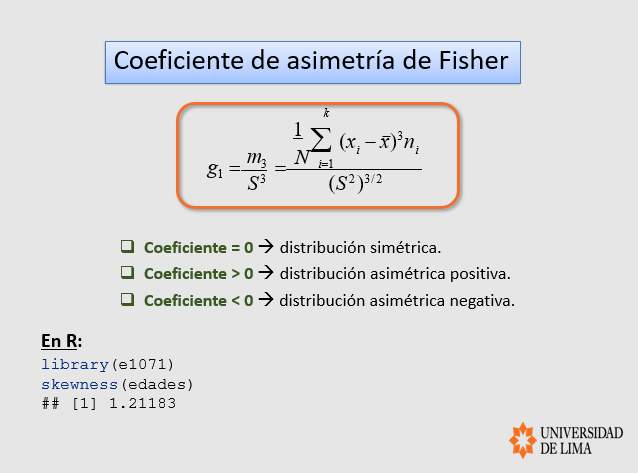
****

Distribución simétrica: 0 o muy cerca de 0. (As = 0)

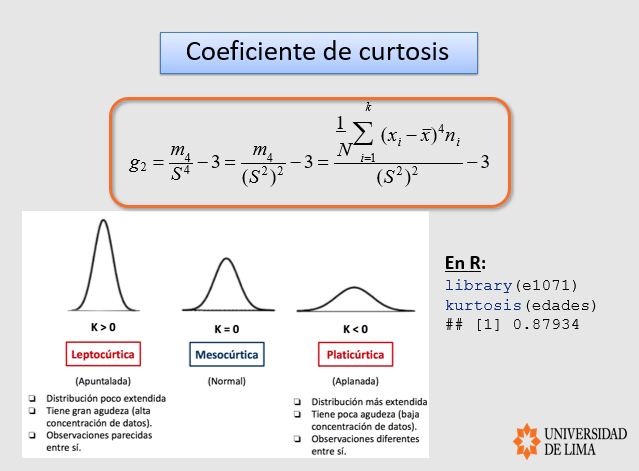
Distribución asimétrica positiva: mayor a 0. (As > 0)

Distribución asimétrica negativa: menor a 0. (As < 0)

* **Coeficiente de asímetria Fisher**

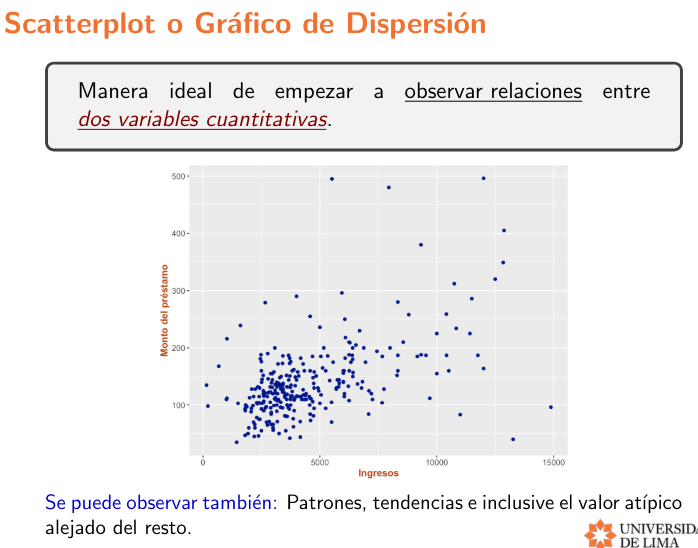
****

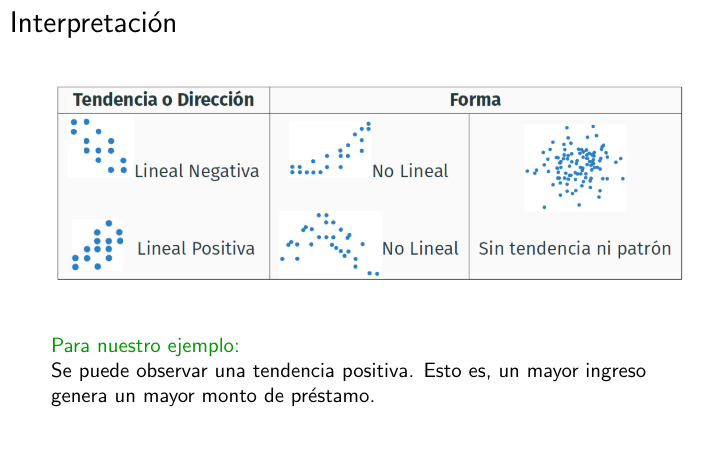
* **Coeficiente Curtosis**

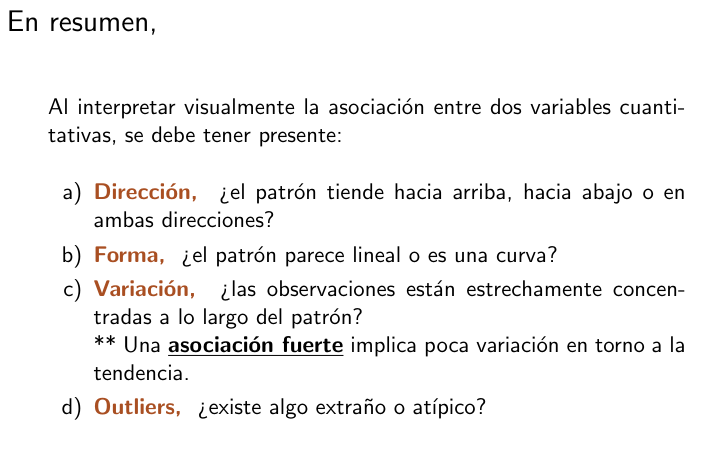
****

***ASOCIACIÓN ENTRE VARIABLES***

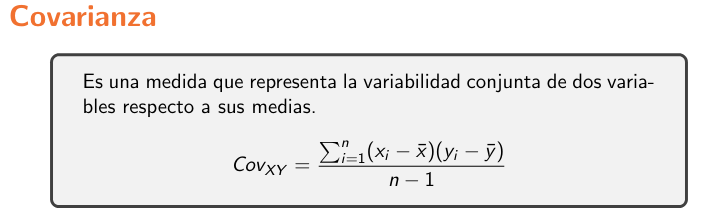
**Asociación entres Variables Cuantitativas**

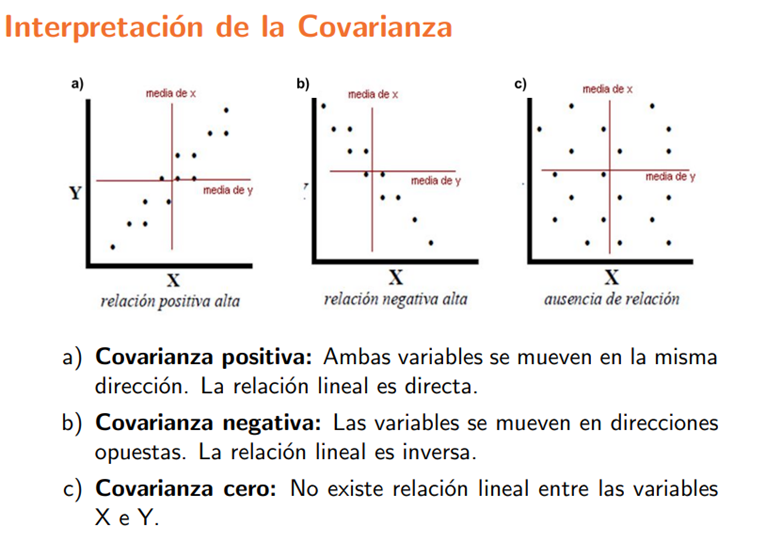




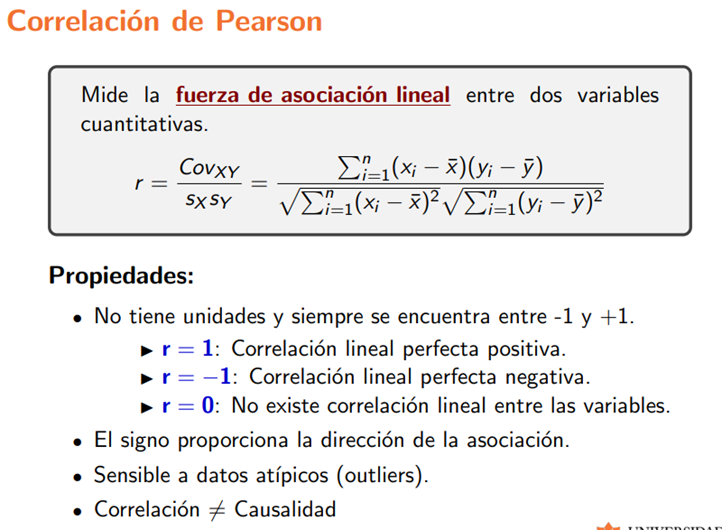


* ***Covarianza:***

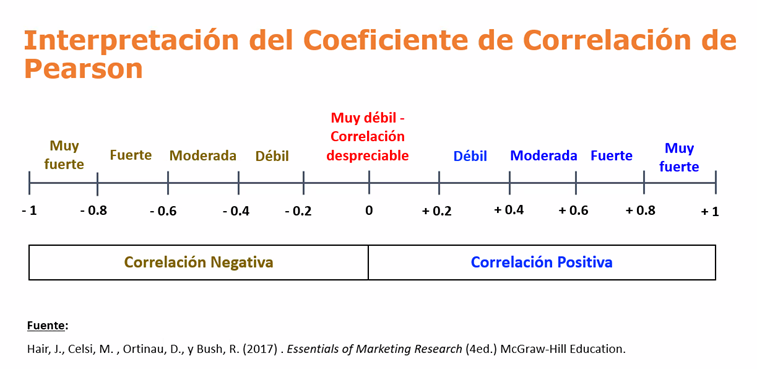


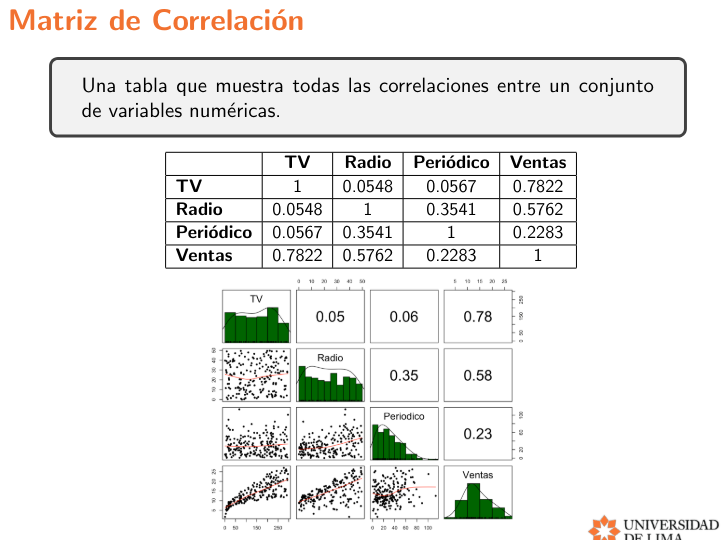


* ***Correlación de Pearson:***

****

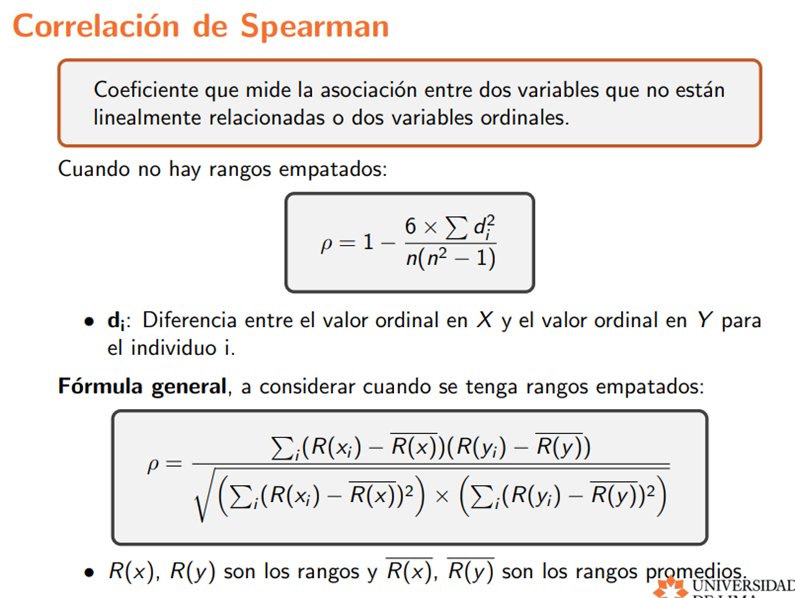
Solo se usa cuando las variables son cuantitativas continuas y hay el supuesto de normalidad. Cuando la relación es lineal.

****

****

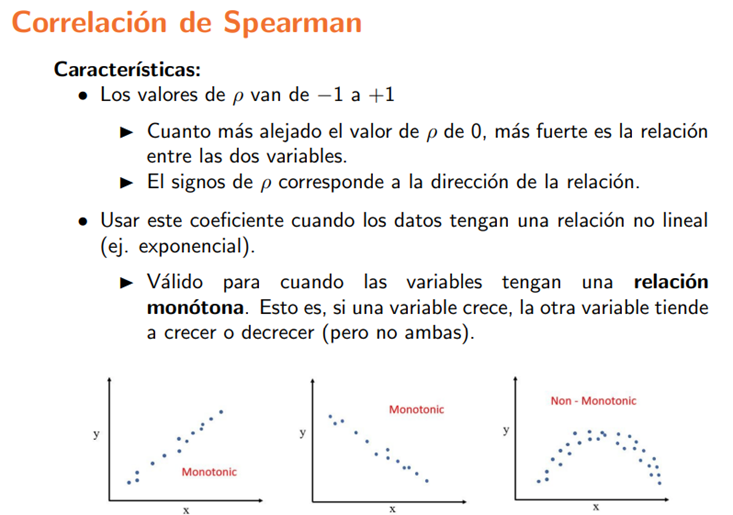
Solo se puede con variables numéricas.

* ***Correlación de Spearman:***

****

Sirve para cualquier tipo de asociación. Pero Pearson es más poderoso. (más te dice si hay asociación o no)

Variables ordinales.



**RStudio:**

1. **Pasar los datos a numéricos**

datos$Edad <- **factor(datos$Edad,levels = c("Joven","Media","Adulta")**,**ordered=TRUE)**

datos$Genero <- **as.factor(datos$Genero)**

datos$Vivienda <- as.factor(datos$Vivienda)

datos$Ecivil <- as.factor(datos$Ecivil)

datos$Ubicacion <- as.factor(datos$Ubicacion)

datos$Historial <- factor(datos$Historial,levels = c("Bajo","Medio","Alto"),ordered=TRUE)

**¿Cómo se usa?**

* **factor(“datos”, levels = c(“ “), ordered = TRUE) →** es para transformar ordinales a números
* **as.factor(“datos”) →** es para transformar nominales a números

1. **Gráfico de dispersión**

plot(datos$Salario, datos$Monto)

install.packages("ggplot2")

library(ggplot2)

ggplot(datos, aes(x = Salario, y = Monto)) +

geom\_point(size = 1.5, colour = "darkblue") +

xlab("Salario") + ylab("Monto") +

theme(axis.title = element\_text(face = "bold", colour = "#b83f04", size = 12),

axis.text = element\_text(size = 10))

1. **Gráfico de dispersión + Histograma**

# Guardar el gráfico ggplot (de dispersión) como objeto de nombre p

p <- ggplot(datos, aes(x = Salario, y = Monto)) +

geom\_point(size = 1.5, colour = "darkblue") +

xlab("Salario") + ylab("Monto") +

theme(axis.title = element\_text(face = "bold", colour = "#b83f04", size = 12),

axis.text = element\_text(size = 10))

# Gráfico de dispersión con los histogramas marginales.

install.packages("ggExtra")

library(ggExtra)

ggExtra::ggMarginal(p, type = "histogram")

1. **Covarianza**

cov(datos$Salario, datos$Monto)

1. **Correlación de Pearson**

cor(datos$Salario, datos$Monto)

1. **Matriz de correlación**

cor(datos[c("Salario", "Monto")])

¿Cómo se usa? Escribes de esa manera las variables que quieres relacionar.

1. **Gráficos adicionales**

# Gráfico de dispersión 2 a 2

pairs(datos[,c("Salario", "Monto")])

pairs(datos[,c("Salario", "Monto")], pch = 19)

# Histograma y correlaciones 2 a 2

install.packages("psych")

library(psych)

pairs.panels(datos[,c("Salario", "Monto")],

method = "pearson", # método de correlación

hist.col = "darkgreen",

density = TRUE, # mostrar gráficos de densidad

ellipses = FALSE # mostrar elipses de correlación

)

# Visualización de una matriz de correlación

install.packages("corrplot")

library(corrplot)

corrplot(cor(datos[,c("Salario", "Monto")]), method="circle")

corrplot(cor(datos[,c("Salario", "Monto")]),

method = "color",

type = "upper",

addCoef.col = "black", # Añadir coeficiente de correlación

tl.col = "black", tl.srt = 45, # Color de la etiqueta y rotación

)

# Gráficos interactivos

#https://plotly.com/r/

cor(datos$Salario, datos$Monto, method="spearman")

cor(datos$Salario, datos$Monto, method="pearson")

cor(as.numeric(datos$Historial), datos$Monto, method="spearman")

cor(as.numeric(datos$Edad), as.numeric(datos$Historial), method="spearman")

1. **Correlación de Pearson o Spearman**

cor(datos$Salario, datos$Monto, method="spearman")

cor(datos$Salario, datos$Monto, method="pearson")

**Solo debes escribir al final “method = “**

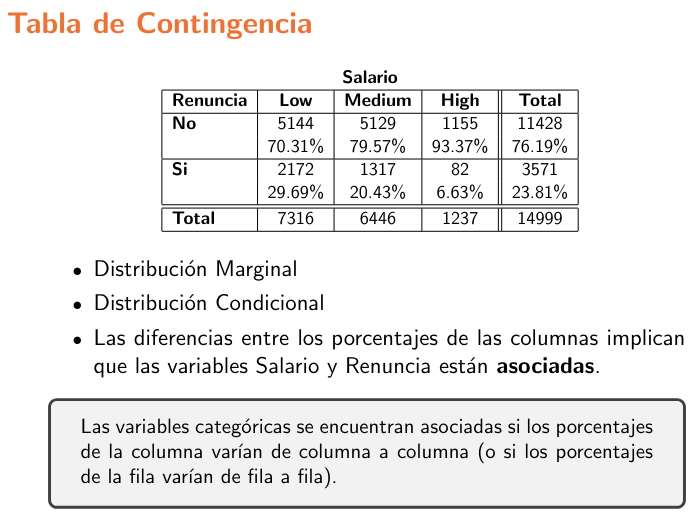
1. **Correlación de variables cuantitativas**

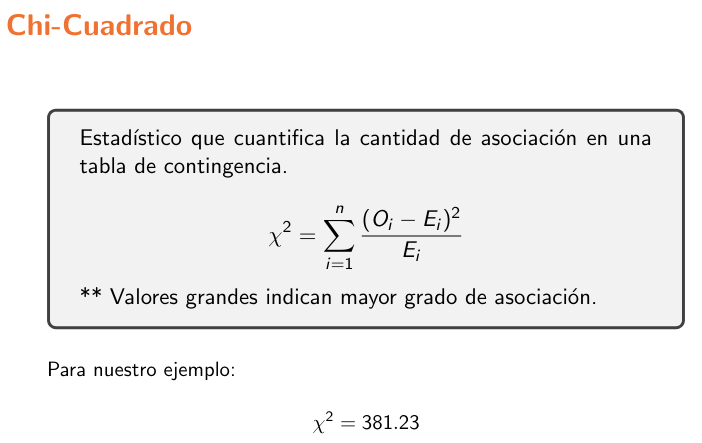
cor(as.numeric(datos$Historial), datos$Monto, method="spearman")

cor(as.numeric(datos$Edad), as.numeric(datos$Historial), method="spearman")

**Debes cambiar los datos a numéricos.**

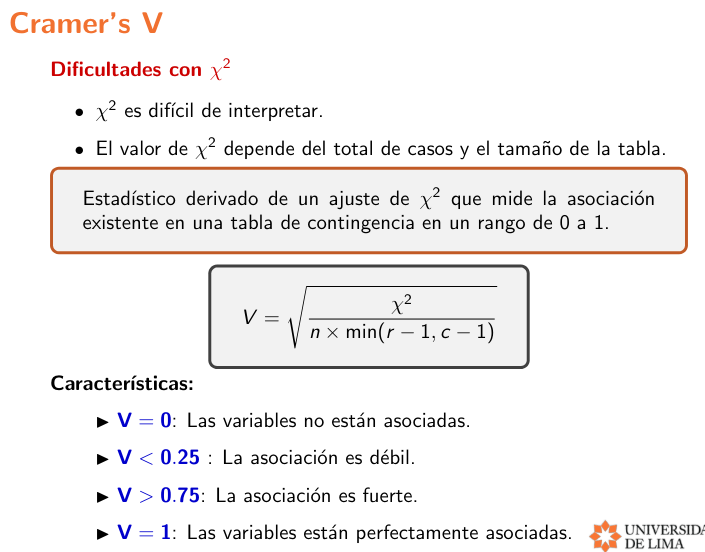
**Asociación entre Variables Cualitativas**



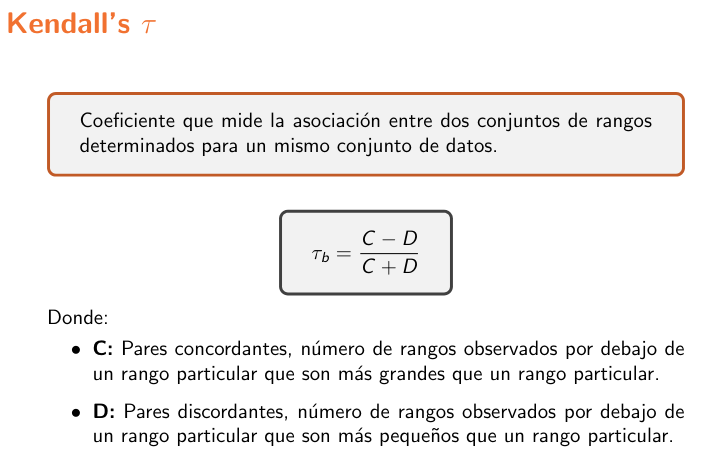


Solo te dice si hay o no hay asociación.

* ***Cuando las dos son cualitativas netas (nominales)***

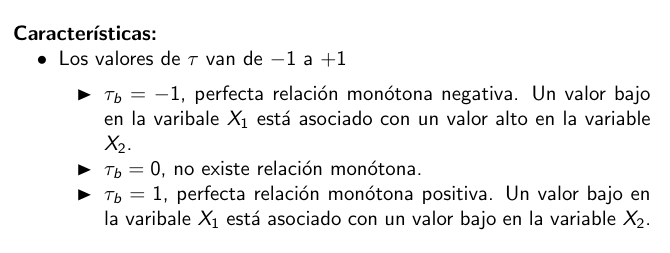


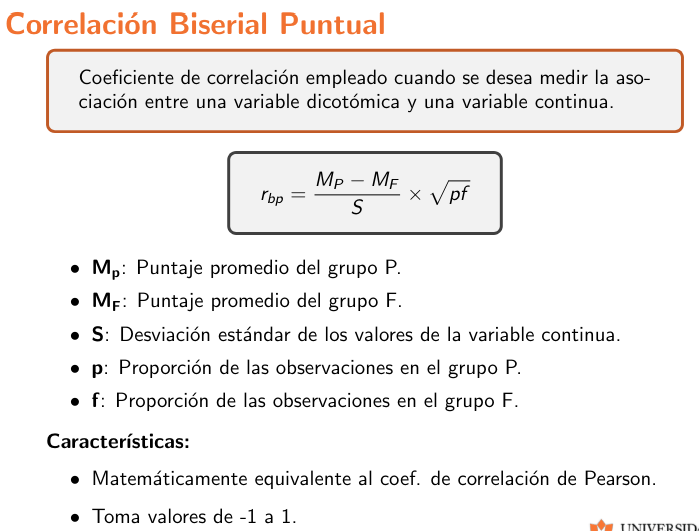
Este estadístico si te dice qué tan fuerte es la asociación.



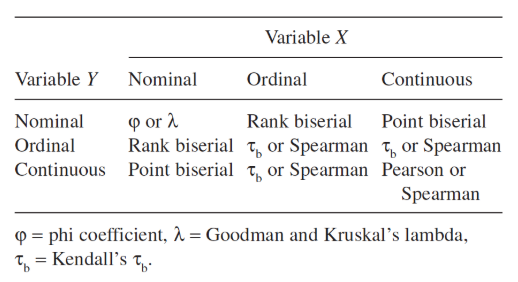
Sirve para ver rangos.

Sirve para datos ordinales.



****

RESUMEN:

****

**RStudio**

**Tabla de Contingencia**

table(datos$Vivienda, datos$Ecivil)

addmargins(table(datos$Vivienda, datos$Ecivil))

# margin = 2 o 1, las proporciones son calculadas por columna

prop.table(table(datos$Vivienda, datos$Ecivil), margin = 2)

**Gráfico de Barras Apilado**

library(ggplot2)

ggplot(datos) +

geom\_bar(aes(x = Ecivil, fill = Vivienda), position = "fill", colour = "white") +

theme\_classic() + xlab("Estado civil") + ylab(" ") +

labs(fill = "Vivienda")

**Chi-Cuadrado**

chisq.test(datos$Vivienda, datos$Ecivil, correct = FALSE)

**Cramer’s V**

install.packages("vcd")

library(vcd)

assocstats(table(datos$Vivienda, datos$Ecivil))

**Kendall’s τ**

# Las variables cualitativas deben estar codificadas de manera numérica

cor(as.numeric(datos$Edad), as.numeric(datos$Historial), method = "kendall")

**Correlación de Spearman**

# Las variables cualitativas deben estar codificadas de manera numérica

cor(as.numeric(datos$Edad), as.numeric(datos$Historial), method = "spearman")

**Correlación Biserial Puntual**

# Numéricamente igual que Pearson

cor(as.numeric(datos$Salario), as.numeric(datos$Vivienda))

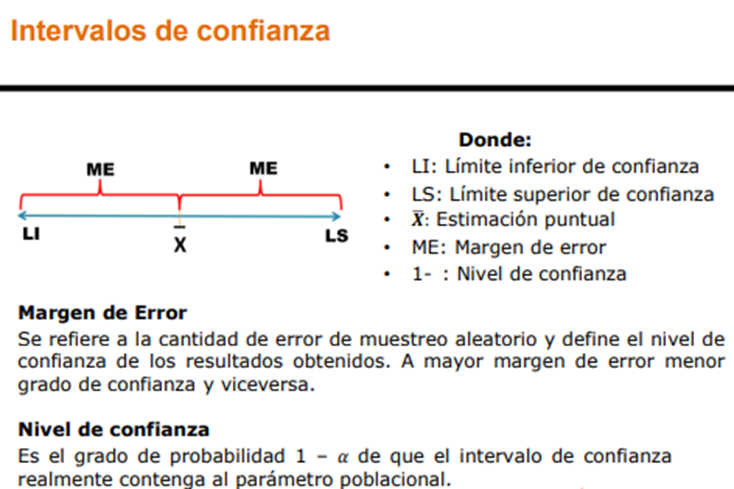
**EJERCICIOS:**

**Crear Tabla**

tabla <- matrix(c(245, 228, 177, 219, 31, 27, 13, 10), nrow=2, ncol = 4, byrow = TRUE, dimnames = list(c("Nutrición buena","Nutrición pobre"),c("< 80", "80 - 90", "90 - 99", "≥100")))

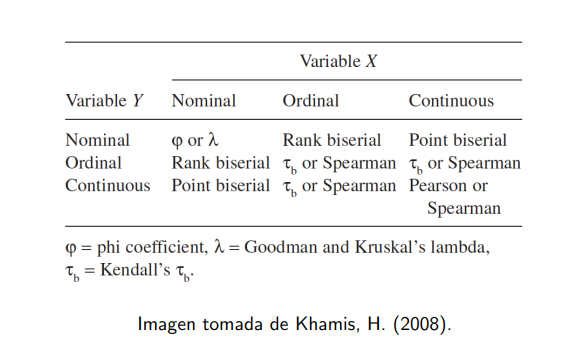
tabla

**Teoría Intervalos**

****

**as.factor**

**factor(datos$Edad, levels = c(“Joven”,”Adulto”))**

****

**NOTA: Biserial puntual → solo se usa con variables dicotómicas(es decir solo 2 variables, ejm: Masculino y Femenino, Propia y Alquilada, Dia y Noche)**

**→ Discreta / Discreta → Crammers**

**→ Nominal / Nominal → Crammers o chisq**

**Como hacer tabla en Rstudio:**

table <- matrix(c(1,2,3,4), nrow = 2, ncol = 2, byrow=TRUE, dimnames = list(c(“Nutrición buena”,”Nutrición pobre” ), c(“C.1”, “C.2”)))