Grado en Estadística

TRABAJO FIN DE GRADO

Análisis de datos obtenidos a través de cuestionarios con ayuda de R

Celia Romero Gustos

Sevilla, Junio de 2020

Índice general

–>

Resumen

El objetivo de este trabajo es el desarrollo de un procedimiento en R-Program para el análisis de datos obtenidos a través de cuestionarios que contenga la construcción de tablas, resúmenes numéricos básicos y visualización gráfica. Además, se realiza un análisis de las relaciones bivariadas y medidas de asociación para variables cuantitativas, ordinales y nominales.

Para ello, con el fin de ilustrar el análisis, se clasifican las variables según sus características y se aplica este procedimiento a distintos conjuntos de datos.

Por último, se crea un paquete en R donde se explican y describen todas las funciones necesarias para realizar un estudio completo, así como sus argumentos y usos.

Abstract

The aim of this work is to develop a procedure in R-Program for the analysis of questionnaire data. The study contains frequency tables, basic numerical summaries, and graphical visualization. In addition, an analysis of bivariate relationships and association measures for quantitative, ordinal, and nominal variables is performed.

For this, the variables are classified according to their characteristics. This procedure is applied to different data sets in order to illustrate the study.

Finally, a package in R with all the functions used for the different types of analysis and variables is created. In this package in R, necessary functions to carry out a complete study are explained and described, as well as their arguments and uses.

Capítulo 1

Introducción

La estadística se utiliza a diario en áreas muy diferentes, desde investigaciones con millones de datos hasta pequeñas muestras. La encuesta es uno de los procedimientos de investigación más usado en multitud de campos científicos. Se ha convertido en una actividad cotidiana en la que todos participamos en algún momento. La encuesta ha alcanzado una gran popularidad por los efectos positivos que puede llegar a tener, genera mucha información, que, usada de manera correcta, es de gran utilidad para conocer la opinión pública acerca de distintos ámbitos. Podemos ver la importancia de las encuestas en la sociedad actual observando la presencia de éstas en los medios de comunicación.

La encuesta se considera una técnica de recogida de datos a través de un interrogatorio a los sujetos cuya finalidad es la de obtener información acerca de un tema previamente planteado. El cuestionario, es el instrumento de recogida de esos datos, en él aparecen las preguntas de forma sistemática y ordenada. (López-Roldán & Fachelli 2015)

Las encuestas carecerán de tanta utilidad si no existe un análisis posterior que permita obtener conclusiones tanto analíticas como gráficas.

Este capítulo consta de tres secciones. En la sección 1.1 se encuentran las definiciones de los tipos de variables que se obtienen de un cuestionario y la manera en la que se han agrupado para realizar su análisis estadístico.
La sección 1.2 presenta la base de datos que se utilizará como ilustración a lo largo del trabajo.
Finalmente, la sección 1.3 muestra una pequeña reseña de cada una de las librerías usadas en los siguientes capítulos.

1.1. Tipos de variables.

Según los valores que tomen las variables aleatorias pueden clasificarse en cualitativas o cuantitativas.

Las **variables cualitativas** (también llamadas categóricas) son aquellas que no pueden asociarse de forma natural a un número. Aunque es frecuente asignarle un código numérico para volcarlo a una base de datos y facilitar su uso, este valor no es relevante. Estas variables a su vez se dividen en nominales u ordinales en función de la escala de medida.

- Las **variables nominales** permiten clasificar cada caso según su pertenencia a una u otra categoría establecida en la variable. Indica cualidad y no tienen un orden ni

relación entre las categorías prefijadas (ciudad de origen). Si las variables nominales sólo tiene dos categorías se llaman **variables dicotómicas** o binarias (fumador o no fumador).

- Las **variables ordinales** son aquellas que además de distinguir entre categorías, establecen un orden entre ellas. Estas variables son muy utilizadas en las encuestas. Es el caso de las valoraciones de aspectos o productos, de niveles de acuerdo frente a una afirmación o propuesta, de niveles de satisfacción ante un impulso, vivencias, acciones, etc. Son, por tanto, las cuestiones que llevan asociadas las opciones de respuesta del tipo "Muy bien, bien, regular,..."Muy deacuerdo, de acuerdo,...","Muy satisfecho, satisfecho,...", etc. Incluso se podrían incluir las cuestiones numéricas que se han agrupado en intervalos, como edad, niveles de renta, etc.

Las **variables cuantitativas** son aquellas que adoptan valores numéricos, representando cantidades, medidas (la estatura o el número de hijos). Pueden distinguirse dos subtipos: continuas y discretas.

- Las **variables continuas**, toman sus valores en un espacio continuo, generalmente dentro de un intervalo numérico, de forma que entre dos valores posibles, siempre existe un posible valor intermedio. Obviamente, esta propiedad queda restringida a la precisión de la medida que se esté realizando (peso de un bebé al nacer).
- Las **variables discretas** tienen sus valores "aislados", es decir, el número de valores posibles entre dos valores dados es finito. Habitualmente se representan a través de los números enteros, de forma que pueden enumerarse y existen valores "consecutivos" entre los que no puede haber otro (número de páginas de un libro).

Es importante esta clasificación porque dependiendo del tipo de variable se aplicará un análisis u otro para su mejor visualización.

1.2. Descripción de la base de datos

Para probar las funciones creadas, se ha usado un fichero *Excel* con datos aleatorios simulando un cuestionario. Cada variable ha sido generado con funciones de *Excel, excepto las variables altura y peso que han sido alteradas manualmente para generar un gráfico más intuitivo.

Los datos contienen la respuesta de 249 personas encuestadas.

A continuación, se describen las diferentes preguntas del cuestionario, así como sus posibles respuestas o explicación y el tipo de variable.

Cuadro 1.1: Descripción de los datos

Variables	**Tipo**	**Descripción**
V1 (Sexo)	Dicotómica	0-Hombre, 1-Mujer
V2 (Ingresos)	Continua	Ingresos brutos mensuales
V3 (Grupos de edad)	Ordinal	1- 18 a 29 , 2- 30 a 49, 3- 50 a 69, 4- 70 años y más
V4 (Ciudad de Origen)	Nominal	1-Sevilla, 2-Madrid, 3-Barcelona, 4-NSNC
V5 (Valore comida)	Ordinal	1-Muy bueno, 2-Bueno, 3-Regular, 4-Malo, 5-Muy malo
V6 (Valore limpieza)	Ordinal	1-Muy bueno, 2-Bueno, 3-Regular, 4-Malo, 5-Muy malo
V7 (Valore personal)	Ordinal	1-Muy bueno, 2-Bueno, 3-Regular, 4-Malo, 5-Muy malo
V8 (Altura)	Continua	Altura actual del individuo
V9 (Hijos)	Discreta	Número de hijos
V10 (Peso)	Continua	Peso actual del individuo

Todas las variables son de tipo numérico, o bien, codificadas con códigos numéricos para una mayor facilidad del encuestador, de esta manera se disminuye el número de valores perdidos por utilizar términos erróneos. Para algunas funciones se necesita que algunas variables sean de tipo factor, por tanto, se transformarán.

En la siguiente tabla se incluye una muestra de los primeros casos de la base de datos generada.

Cuadro 1.2: Previsualización datos

V1	V2	V3	V4	V5	V6	V7	V8	V9	V10
1	1778	3	1	4	1	1	1.75	4	75
0	1087	4	2	4	5	3	1.77	0	77
0	1214	4	2	4	2	3	1.62	1	60
1	1173	4	1	1	4	2	1.54	2	60
0	2276	1	2	4	1	2	1.74	2	55
1	1953	2	1	2	3	3	1.85	4	78

1.3. Librerías

Para este trabajo se han usado las siguientes librerías:

- **readxl**: Importa datos de excel(.xlsx y .xls) a R. Se ha usado la función *read_excel* (Wickham & Bryan 2019)
- **dplyr**: Una herramienta rápida y consistente para trabajar con marcos de datos como objetos. Se ha usado la función *mutate*. (Wickham *et al.* 2020)
- **ggplot2**: Permite crear gráficos de manera sencilla a partir de unos datos dados. Todos los gráficos de este paquete funcionan de manera similar, se empieza llamando a ggplot() que proporciona datos por defecto y asignaciones estéticas especificadas por aes (). Luego agrega capas, escalas, coordenadas y facetas antecedido por '+'. (Wickham 2016)
- **knitr**: Proporciona una herramienta de uso general para la generación dinámica de informes en R. La función más usada en este trabajo es *kable* que crea tablas en

latex, html y markdown. (Xie 2020)

- **sjPlot**: Colección de funciones análisis gráfico y salida de tablas para visualización de datos (Lüdecke 2020)
- **tables**: Calcula y muestra tablas complejas de estadísticas resumidas. La salida puede estar en LaTeX, HTML, texto sin formato o una R matriz para su posterior procesamiento. La función usada es *tabular* para crear tablas de resúmenes numéricos. (Murdoch 2020)
- **DescTools**: Una colección de diversas funciones estadísticas básicas y envoltorios convenientes para describir datos de manera eficiente. La función usada es *Lambda* para estudiar la asociación entre dos variables. (Signorell & mult.al. 2020)
- **Corrplot**: Crea una visualización gráfica de una matriz de correlación, intervalo de confianza o matriz general. *Corrplot.mixed* es la función de ese paquete utilizada para visualizar una matriz de correlación. (Wei & Simko 2017)

Capítulo 2

Análisis univariante

2.1. Introducción

En este capítulo se incluyen las funciones elaboradas para el análisis univariante de cada tipo de variable de un cuestionario. Se compone de dos secciones, la primera incluye las tablas de frecuencia relativa, absoluta y acumulada y tablas de resúmenes numéricos básicos; en la segunda sección de este capítulo, se muestran los gráficos para el análisis univariante de las distintas variables. Se han usado distintas librerías explicadas en el primer capítulo e ilustrado las funciones con la base de datos detallada en el capítulo anterior.

Para cada función, se explican los argumentos de entrada, el cuerpo de la función y seguidamente, se ilustra con los datos de ejemplo su utilidad.

2.2. Tablas

La primera acción, o al menos una de las primeras, de un análisis estadístico es ordenar, agrupar y resumir la información que proporciona el conjunto de datos. Una de las principales estrategias para ello es la construcción de tablas estadísticas, entendiendo por tal a la disposición de forma ordenada y agrupada de los valores y frecuencias de la distribución correspondiente a una variable asociada a una cuestión o medida incluida en el conjunto de datos.

2.2.1. Variables dicotómicas

La siguiente función devuelve la tabla de frecuencias relativas y absolutas para variables binarias o dicotómicas.

Uso:

freq.dic (datos, X, etiquetas, título)

Argumentos:

- Datos: data.frame con los datos a analizar.
- X: vector de respuestas de la variable dicotómica a analizar (tamaño n)
- Etiquetas: vector de cadena de caracteres con las posibles respuestas.

- Título: cadena de caracteres indicando el nombre que se le quiere dar a la tabla.

```
freq.dic= function(datos, X, etiquetas, título) {
   datos1 <- mutate(datos, v1 = factor(X, labels = etiquetas))
   Frec.abs=table(datos1$v1)
   Frec.rel=round(prop.table(Frec.abs),2)
   tabla=cbind(Frec.abs,Frec.rel )
   tabla=rbind(tabla, Total = colSums(tabla))
   kable( tabla, caption = título)
  }
```

Para ilustrar la función se considera la variable V1.

```
etiquetas=c("Hombre", "Mujer")
título= "Variable dicotómica"
freq.dic(datos, V1, etiquetas, título)
```

Cuadro 2.1: Variable dicotómica

	Frec.abs	Frec.rel
Hombre	108	0.43
Mujer	141	0.57
Total	249	1.00

2.2.2. Variables ordinales

La siguiente función genera una tabla con las frecuencias relativas, absolutas y acumuladas para variables ordinales.

Uso:

freq.ord(datos, X, etiquetas, título)

Argumentos:

- Datos: data.frame con los datos a analizar.
- X: vector de respuestas de la variable ordinal a analizar (tamaño n).
- Etiquetas: vector de cadena de caracteres con las posibles respuestas.
- Título: cadena de caracteres indicando el nombre que se le quiere dar a la tabla.

```
freq.ord= function(datos, X, etiquetas, título) {
  datos1 <- mutate(datos, v1 = factor(X, labels = etiquetas))

   Frec.abs=table(datos1$v1)
   Frec.rel=round(prop.table(Frec.abs),2)
   Frec.abs.acum = cumsum(Frec.abs)
   Frec.rel.acum = cumsum(Frec.rel)
   tabla=cbind(Frec.abs,Frec.rel ,Frec.abs.acum, Frec.rel.acum )
   kable( tabla, caption = título)
  }
```

Se considera la variable V5 de tipo ordinal para ilustrar esta función.

```
etiquetas=c("Muy malo", "Malo", "Regular", "Bueno", "Muy bueno")
título= "Variable ordinal"
freq.ord(datos, V5, etiquetas, título)
```

Cuadro 2.2: Variable ordinal

	Frec.abs	Frec.rel	Frec.abs.acum	Frec.rel.acum
Muy malo	45	0.18	45	0.18
Malo	43	0.17	88	0.35
Regular	47	0.19	135	0.54
Bueno	61	0.24	196	0.78
Muy bueno	53	0.21	249	0.99

2.2.3. Variables nominales

La siguiente función genera la tabla de frecuencias relativas, absolutas y acumuladas para variables nominales.

Uso:

freq.nom(datos, X, etiquetas, título)

Argumentos:

- Datos: data.frame con los datos a analizar.
- X: vector de respuestas de la variable nominal a analizar (tamaño n).
- Etiquetas: vector de cadena de caracteres con las posibles respuestas.
- Nombre: cadena de caracteres indicando la pregunta realizada o nombre de X (No admite espacios).
- Título: cadena de caracteres indicando el nombre que se le quiere dar a la tabla.

```
freq.nom= function(datos, X, etiquetas, título) {

  datos1 <- mutate(datos, v1 = factor(X, labels = etiquetas))
  Frec.abs=table(datos1$v1)
  Frec.rel=round(prop.table(Frec.abs),2)
  Frec.abs.acum = cumsum(Frec.abs)
  Frec.rel.acum = cumsum(Frec.rel)
  tabla=cbind(Frec.abs,Frec.rel ,Frec.abs.acum, Frec.rel.acum )
  kable( tabla, caption = título )


  }
```

Considerando la variable nominal V4 de mis datos para ilustrar la función:

```
etiquetas=c("Sevilla", "Madrid", "Barcelona","NSNC")
título= " Variable Nominal "
freq.nom(datos, V4, etiquetas, título)
```

Cuadro 2.3: Variable Nominal

	Frec.abs	Frec.rel	Frec.abs.acum	Frec.rel.acum
Sevilla	101	0.41	101	0.41
Madrid	73	0.29	174	0.70
Barcelona	68	0.27	242	0.97
NSNC	7	0.03	249	1.00

2.2.4. Variables cuantitativas continuas

La siguiente función devuelve una tabla con las medidas más relevantes del análisis de variables numéricas, n (tamaño muestral), media, mediana, desv (desviación estándar), IC1, IC2, el intervalo de confianza con un nivel del 95 %.

Uso:

desc.cont(datos, x, nombre, título)

Argumentos:

- Datos: data.frame con los datos a analizar.
- X: vector de respuestas de la variable continua a analizar (tamaño n).
- Nombre: cadena de caracteres indicando la pregunta realizada. (No admite espacios).
- Título: cadena de caracteres indicando el nombre que se le quiere dar a la tabla.

```
desc.cont= function (datos, x, nombre, título) {
  IC1 <- function(x) {
    mean(x) - qt( 0.95, df = length(x) - 1) * sd(x) / sqrt(length(x))}
  IC2 <- function(x){
    mean(x) + qt( 0.95, df = length(x) - 1) * sd(x) / sqrt(length(x))}
  Media <- function(x) {mean(x)}
  Mediana<- function (x) {median(x)}
  Desv<- function (x) {sd(x)}

  tabla <- "~ (n=1)+(nombre =x)*(Media+Mediana+Desv+IC1+IC2)"
  tt <- tabular (tabla, data = datos)

  df <- data.frame(matrix(unlist(tt), nrow=1, byrow=F))
  colnames(df) = c("n", "Media", "Mediana", "Desv", "IC1", "IC2")
  rownames(df)= nombre

  kable(df, digits = 2, caption = título)

 }
```

```
título= "Variable Continua"
desc.cont(datos, V8, "Estatura", título)
```

Cuadro 2.4: Variable Continua

	n	Media	Mediana	Desv	IC1	IC2
Estatura	249	1.67	1.67	0.14	1.66	1.69

2.2.5. Variables cuantitativas discretas

La siguiente función devuelve una tabla con las medidas descriptivas más relevantes del análisis de variables numéricas discretas, n (tamaño muestral), media, mediana, desv (desviación estándar), IC1, IC2, el intervalo de confianza con un nivel del 95 %.

Uso:

desc.disc(datos, x, nombre, título)

Argumentos:

- Datos: data.frame con los datos a analizar.
- X: vector de respuestas de la variable discreta a analizar (tamaño n).
- Título: cadena de caracteres indicando el nombre que se le quiere dar a la tabla.

```
desc.disc= function (datos, x, título) {

  IC1 <- function(x) {
    mean(x) - qt( 0.95, df = length(x) - 1) * sd(x) / sqrt(length(x))}
  IC2 <- function(x){
    mean(x) + qt( 0.95, df = length(x) - 1) * sd(x) / sqrt(length(x))}
  Media <- function(x){mean(x)}
  Mediana<- function (x) {median(x)}
  Desv<- function (x) {sd(x)}

  tabla <- "~(n=1)+(nombre =x)*(Media+Mediana+Desv+IC1+IC2)"
  tt <- tabular (tabla, data = datos)

  df <- data.frame(matrix(unlist(tt), nrow=1, byrow=F))
  colnames(df) = c("n", "Media", "Mediana", "Desv", "IC1", "IC2")

  kable(df, digits = 2, caption = título)
   }
```

```
título="Tabla de frecuencias variable hijos"
desc.disc(datos, V9, título)
```

Cuadro 2.5: Tabla de frecuencias variable hijos

n	Media	Mediana	Desv	IC1	IC2
249	2.11	2	1.53	1.95	2.27

2.2.6. Otras funciones para variables cuantitativas

2.2.6.1. Dividida en dos partes

Esta función devuelve una tabla de resúmenes numéricos básicos, tamaño muestral, media, mediana, desviación estándar e intervalo de confianza al 95 % de una variable cuantitativa dada dividida en dos partes, la primera menor que un valor y la segunda mayor que el mismo valor dado.

Uso:

desc.cuan.subg(datos, x, nombre,valor, título)

Argumentos:

- Datos: data.frame con los datos a analizar.
- X: vector de respuestas de la variable cuantitativa a analizar (tamaño n).
- Nombre: cadena de caracteres indicando la pregunta realizada.
- Valor: de tipo numérico, valor dado para separar nuestra variable en dos partes.
- Título: cadena de caracteres indicando el nombre que se le quiere dar a la tabla.

```
desc.cuan.subg= function (datos, x, nombre, valor, título ) {
   IC1 <- function(x) {
     mean(x) - qt( 0.95, df = length(x) - 1) * sd(x) / sqrt(length(x))}
   IC2 <- function(x){
      mean(x) + qt( 0.95, df = length(x) - 1) * sd(x) / sqrt(length(x))}
  Media <- function(x){mean(x)}
  Mediana<- function (x) {median(x)}
  Desv<- function (x) {sd(x)}

  tabla <- "( Mayorquevalor = x > valor ) + ( Menorquevalor= x <= valor )~
  (n=1)+((nombre=x))* (Media+Mediana+Desv+IC1+IC2)"

  tt <- tabular (tabla, data = datos)

  df <- data.frame(matrix(unlist(tt), nrow=2, byrow=F))
  colnames(df) = c("n", "Media", "Mediana", "Desv", "IC1", "IC2")
  rownames(df) = c("mayor que" , "menor que")
  kable(df, digits = 2, caption = título)

   }
```

```
nombre="Peso"
título= "Variable cuantitativa dividida "
desc.cuan.subg(datos, V10, nombre , 68, título)
```

Cuadro 2.6: Variable cuantitativa dividida

	n	Media	Mediana	Desv	IC1	IC2
mayor que	117	80.28	81	7.55	79.13	81.44
menor que	132	55.65	56	7.74	54.54	56.77

2.2.6.2. Dos variables cuantitativas por columnas

Esta función calcula las medidas más relevantes del análisis descriptivo de dos variables cuantitativas, ya sean continuas, tamaño muestral, media, mediana, desviación estándar e intervalo de confianza por columnas.

Uso:

desc.2cuan.colum(datos, x1, x2, nombrex1, nombrex2, título)

Argumentos:

- Datos: data.frame con los datos a analizar.
- X1: vector de respuestas de la variable cuantitativa 1 a analizar (tamaño n).
- X2: vector de respuestas de la variable cuantitativa 2 a analizar (tamaño n).
- Título: cadena de caracteres indicando el nombre que se le quiere dar a la tabla.

```
desc.2cuan.colum= function (datos,x1,x2, título) {
  IC1 <- function(x) {
    mean(x) - qt( 0.95, df = length(x) - 1) * sd(x) / sqrt(length(x))}
  IC2 <- function(x){
    mean(x) + qt( 0.95, df = length(x) - 1) * sd(x) / sqrt(length(x))}
  Media <- function(x){mean(x)}
  Mediana<- function (x) {median(x)}
  Desv<- function (x) {sd(x)}

  tabla <- "~(n=1)+Format(digits=2)*((nombrex1=x1)+(nombrex2=x2))*
            (Media+Mediana+Desv+IC1+IC2)"

  tt <- tabular (tabla, data = datos)

  df <- data.frame(matrix(unlist(tt), nrow=1, byrow=F))
  colnames(df) = c("n", "Media", "Mediana", "Desv", "IC1", "IC2",
                        "Media", "Mediana", "Desv", "IC1", "IC2")

  kable(df, digits = 2, caption = título) }
```

```
título= " Dos variables cuantitativas de Altura y Peso"
desc.2cuan.colum(datos,V8, V10, título )
```

Cuadro 2.7: Dos variables cuantitativas de Altura y Peso

n	Media	Mediana	Desv	IC1	IC2	Media	Mediana	Desv	IC1	IC2
249	1.67	1.67	0.14	1.66	1.69	67.22	67	14.49	65.71	68.74

2.2.6.3. Dos variables cuantitativas por filas

Esta función calcula las medidas más relevantes del análisis descriptivo de dos variables cuantitativas, nos muestra el tamaño muestral, media, mediana, desviación estándar e intervalo de confianza por filas.

Uso:

desc.2cuan.filas (x1, x2, nombrex1, nombrex2, título)

Argumentos:

- Datos: data.frame con los datos a analizar.
- X1: vector de respuestas de la variable cuantitativa 1 a analizar (tamaño n).
- X2: vector de respuestas de la variable cuantitativa 2 a analizar (tamaño n).
- Nombrex1: cadena de caracteres indicando la pregunta realizada para la variable x1 (No admite espacios).
- Nombrex2: cadena de caracteres indicando la pregunta realizada para la variable x2 (No admite espacios).
- Título: cadena de caracteres indicando el nombre que se le quiere dar a la tabla.

```
desc.2cuan.filas= function (x1,x2, nombrex1, nombrex2, título) {
  IC1 <- function(x) {
    mean(x) - qt( 0.95, df = length(x) - 1) * sd(x) / sqrt(length(x))}
  IC2 <- function(x){
    mean(x) + qt( 0.95, df = length(x) - 1) * sd(x) / sqrt(length(x))}
  Media <- function(x){mean(x)}
  Mediana<- function (x) {median(x)}
  Desv<- function (x) {sd(x)}

  tabla <- "((nombrex1=x1)+(nombrex2=x2))~(n=1)+
            (Media+Mediana+Desv+IC1+IC2)"

  tt <- tabular (tabla, data = datos)
  df <- data.frame(matrix(unlist(tt), nrow=2, byrow=F))
  colnames(df) = c("n", "Media", "Mediana", "Desv", "IC1", "IC2")
  rownames(df)= c(nombrex1, nombrex2)

  kable(df, digits = 2, caption = título)
   }
```

```
desc.2cuan.filas(V8, V10, "Altura","Peso", "Dos variables cuantitativas")
```

Cuadro 2.8: Dos variables cuantitativas

	n	Media	Mediana	Desv	IC1	IC2
Altura	249	1.67	1.67	0.14	1.66	1.69
Peso	249	67.22	67.00	14.49	65.71	68.74

2.3. Visualización gráfica.

2.3.1. Variables cualitativas

Para las variables cualitativas, ya sean nominales u ordinales, el análisis gráfico univariante se realiza con la función graf.cual, que genera un diagrama de barras y diagrama de sectores.

Uso:

graf.cual(datos, x, etiquetas, nombre, título)

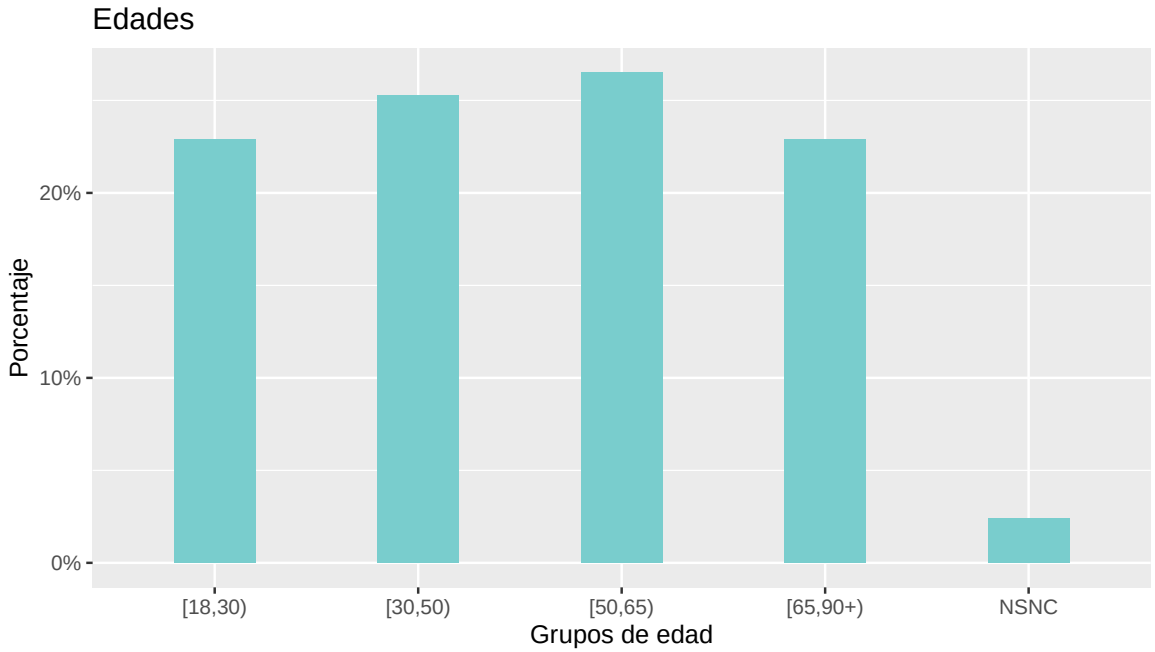
Argumentos:

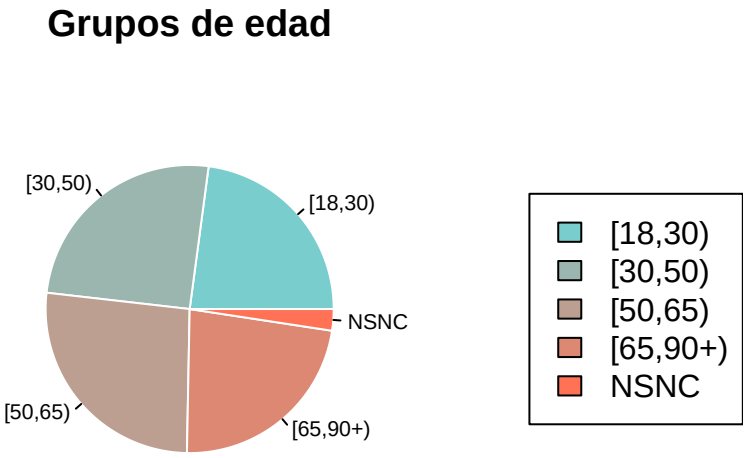
- Datos: data.frame donde se encuentra X.
- X: vector de respuestas de la variable cualitativas a analizar (tamaño n).
- Etiquetas: vector de cadena de caracteres con las posibles respuestas no numéricas
- Nombre: cadena de caracteres indicando la pregunta realizada o nombre de X.
- Título: cadena de caracteres indicando el nombre que se le quiere dar a los gráficos.

```
graf.cual= function(datos, x , etiquetas, nombre, título) {

  #Crea una nueva columna como factores de la variable X
  datos <- mutate(datos, factor = factor(x, labels = etiquetas))

  #Se calcula la longitud del vector etiquetas
  netiq= length(etiquetas)

  #Paleta de color.
  mis.colores <- colorRampPalette(c( "darkslategray3","coral1" ))

  #Diagrama de barras

  print(ggplot(datos, aes(x = factor)) +
       geom_bar(width = 0.4,
                aes(y = (..count..)/sum(..count..)),
                fill = mis.colores(1)) +
       scale_x_discrete(nombre) +
       scale_y_continuous("Porcentaje", labels=scales::percent) +
       labs(title = título))

  #Diagrama de sectores
```

```
  pie(table(datos$factor), main=nombre, border="white",
      radius=0.75,
      cex=0.7, col = mis.colores(netiq))
  legend("right", etiquetas, fill=mis.colores(netiq), cex=1)

}
```

```
etiquetas=c("[18,30)","[30,50)","[50,65)","[65,90+)","NSNC")
nombre= "Grupos de edad"
graf.cual(datos, datos$V3, etiquetas, nombre, "Edades")
```





2.3.2. Visualización gráfica de Escalas Likert.

Para las variables ordinales de valoración existe una herramienta de medición llamada 'Escala de likert', que nos permite medir aptitudes y saber el grado de satisfacción del encuestado acerca de una afirmación dada. (Devlin 2016)

2.3.2.1. Para una variable

La siguiente función genera un gráfico de barras para una variable ordinal escala Likert.

Uso:

graf.val1(datos, x, etiquetas, nombre)

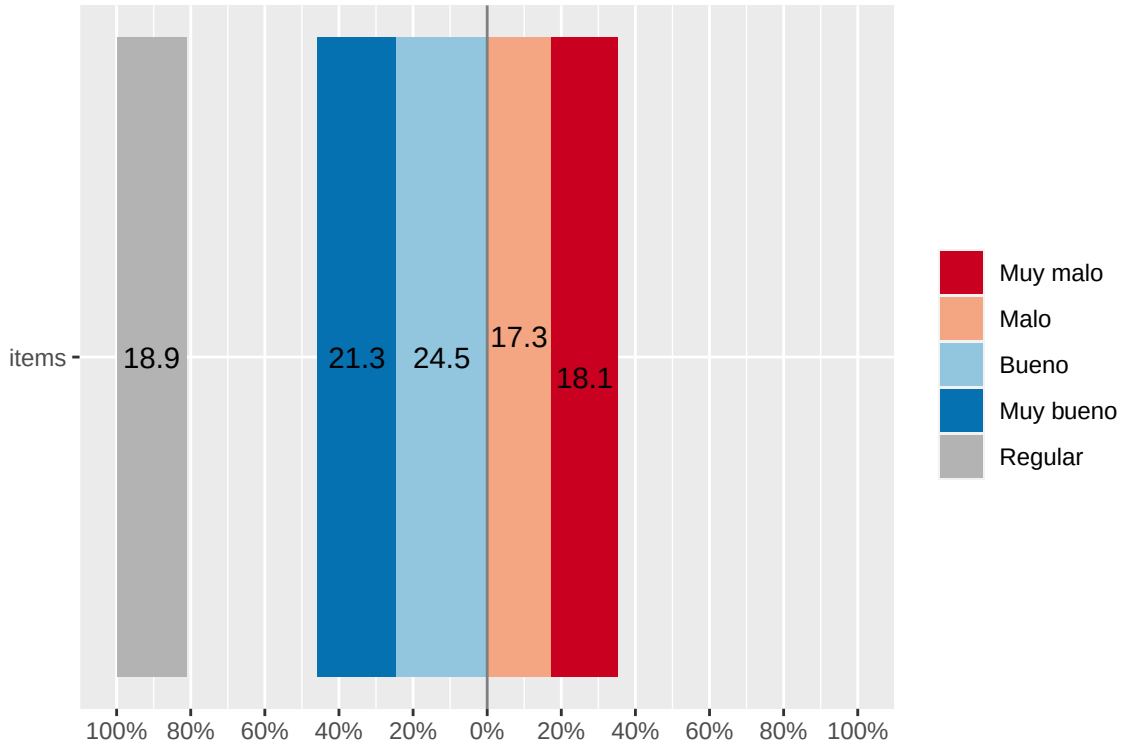
Argumentos:

- Datos: data.frame donde se encuentra X.
- X: vector de respuestas de la variable cualitativas a analizar (tamaño n).
- Etiquetas: vector de cadena de caracteres con las posibles respuestas no numéricas.
- Nombre: cadena de caracteres indicando la pregunta realizada o nombre de X.

```
graf.val1=function(datos, x, etiquetas, nombre) {

        datos <- mutate(datos, factor = factor(x, labels = etiquetas))
        data=cbind.data.frame(datos$factor)
        names(data)= nombre


         #Gráfico de barras
         plot_likert(data, show.n=FALSE,
                     geom.colors = "RdBu",
                     show.prc.sign = FALSE,
                     cat.neutral = 3)



}
```

Por ejemplo:

```
etiquetas=c("Muy malo", "Malo", "Regular", "Bueno", "Muy bueno")
nombre="Calidad"
graf.val1(datos, V5, etiquetas, nombre)
```



2.3.2.2. Para bloques de tres variables ordinales con las mismas etiquetas

Con esta función se obtienen diagramas de barra bloques de tres variables ordinales de valoración con las mismas etiquetas, es decir, mismas posibles respuestas, para así obtener de manera más visual la comparación de éstas. El gráfico de densidad además de mostrar la distribución de los valores, aparece una línea vertical que representa la media.

Uso:

graf.val3(datos, x1, x2, x3, etiquetas, preguntas)

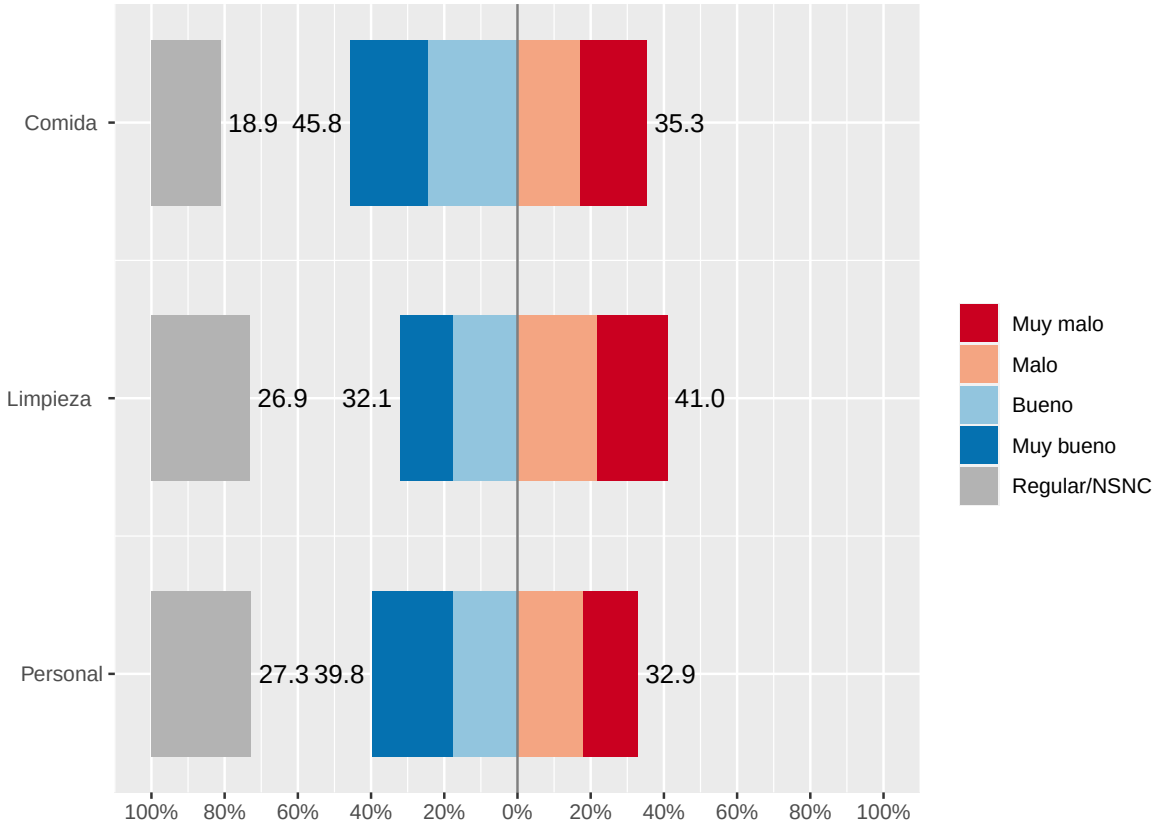
Argumentos:

- Datos: data.frame donde se encuentra X.
- X1: vector de respuestas de la variable ordinal 1 a analizar (tamaño n).
- X2: vector de respuestas de la variable ordinal 2 a analizar (tamaño n).
- X3: vector de respuestas de la variable ordinal 3 a analizar (tamaño n).
- Etiquetas: vector de cadena de caracteres con las posibles respuestas no numéricas de las 3 variables dadas.
- Preguntas: vector de cadena de caracteres con las preguntas o nombres, en orden, de x1, x2 y x3.

```
graf.val3=function(datos,x1,x2,x3, etiquetas, preguntas) {
        datos <- mutate(datos, factor1 = factor(x1, labels = etiquetas))
        datos <- mutate(datos, factor2 = factor(x2, labels = etiquetas))
        datos <- mutate(datos, factor3 = factor(x3, labels = etiquetas))

        data=cbind.data.frame(datos$factor1,datos$factor2, datos$factor3)
        names(data)=preguntas
```

```
        #Gráfico de barras

        plot_likert(data, show.n=FALSE,
                    geom.colors = "RdBu",
                    show.prc.sign = FALSE,
                    value = "sum.outside",
                    cat.neutral = 3)



}
```

```
datos <- read_excel("datos1.xlsx")
etiquetas=c("Muy malo", "Malo", "Regular/NSNC", "Bueno", "Muy bueno")
preguntas=c("Comida ","Limpieza  ","Personal")
graf.val3(datos, V5, V6, V7, etiquetas, preguntas)
```



2.3.3. Variables cuantitativas continuas

La siguiente función genera diagrama de cajas, gráfico de densidad, histograma con media, histograma con densidad de una variable continua.

Uso:

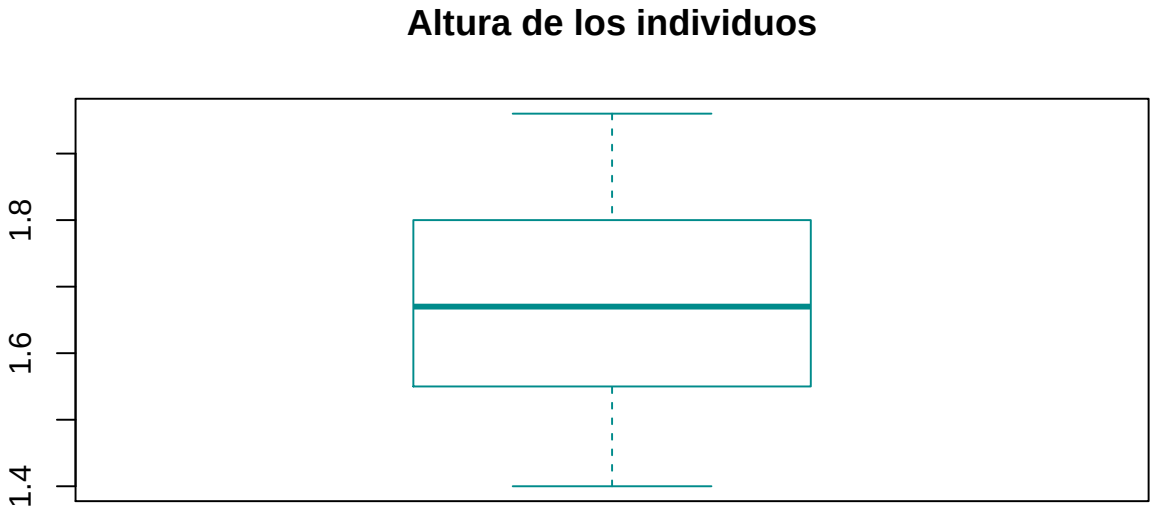
graf.cont(datos, X, nombre, título)

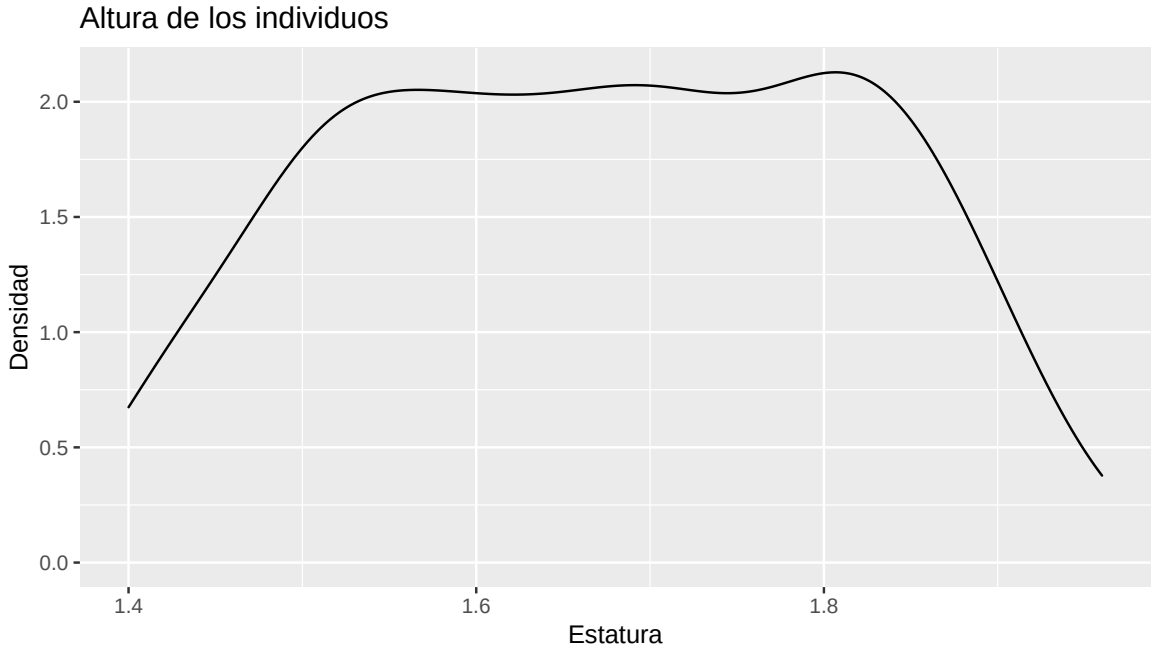
Argumentos:

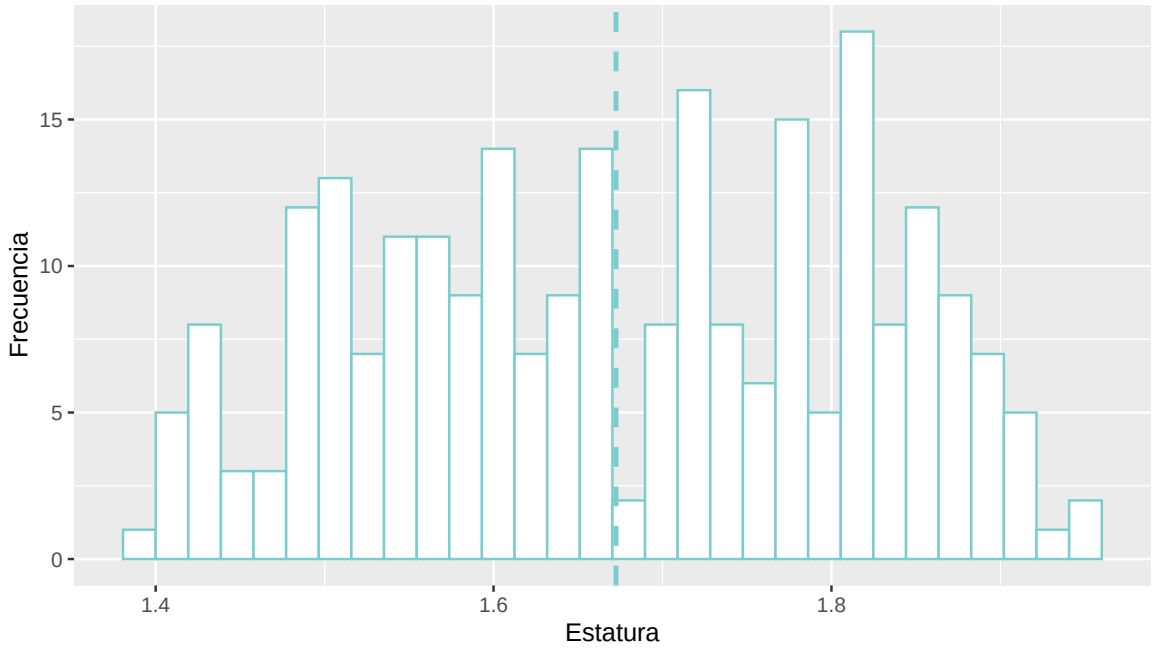
- Datos: data.frame con los datos a analizar.
- X: vector de respuestas de la variable cuantitativa a analizar (tamaño n).
- Nombre: cadena de caracteres indicando el nombre que se le da a X.
- Título: cadena de caracteres indicando el nombre que se le quiere dar a los gráficos.

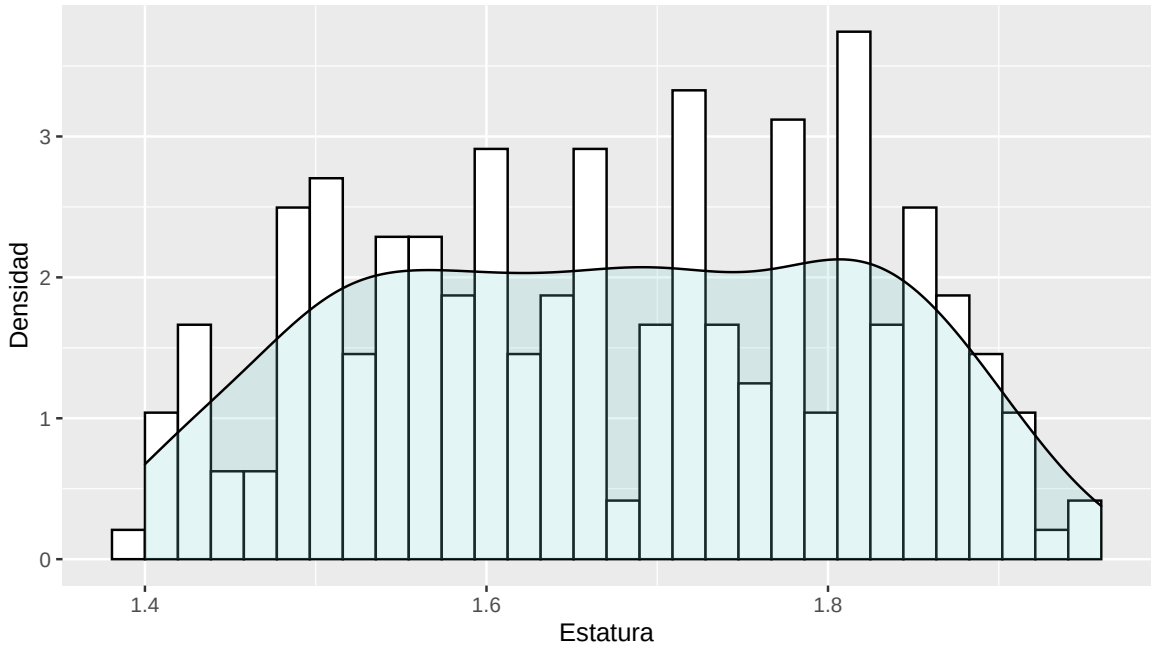
```
graf.cont=function(datos, X, nombre, título){
  mis.colores <- colorRampPalette(c("darkslategray3", "coral1"))

  #Diagrama de caja
  boxplot(X, xlab=nombre, main = título, border= "cyan4")

  #Densidad
   print(ggplot(datos, aes(x = X)) +
    geom_density() +
    scale_x_continuous(nombre)+
    scale_y_continuous("Densidad")+
    labs(title = título))

  #Histograma con media
  print(ggplot(datos, aes(x=X)) +
    geom_histogram(bins=30, color=mis.colores(1), fill="white") +
    scale_x_continuous(nombre) +
    scale_y_continuous("Frecuencia")+
    geom_vline(aes(xintercept=mean(X)),
                  color=mis.colores(1), linetype="dashed", size=1))

  #Histograma + Densidad

  print(ggplot(datos, aes(x=X)) +
    geom_histogram(bins=30,aes(y=..density..),
                    colour="black", fill="white")+
    scale_x_continuous(nombre) +
    scale_y_continuous("Densidad")+
    geom_density(alpha=.2, fill=mis.colores(1)) )

}
```

```
nombre= "Estatura"
graf.cont(datos, V8, nombre, "Altura de los individuos")
```


Altura de los individuos
1.4
1.6
1.8
Estatura


Altura de los individuos
Densidad
0.0
0.5
1.0
1.5
2.0
1.4
1.6
1.8
Estatura


Frecuencia
15
10
5
0
1.4
1.6
1.8
Estatura


Densidad
3
2
1
0
1.4
1.6
1.8
Estatura

2.3.4. Variables cuantitativas discretas

Esta función es similar a la de variables continuas, recogida en la subsección anterior, pero sin incorporar histogramas.

Uso:

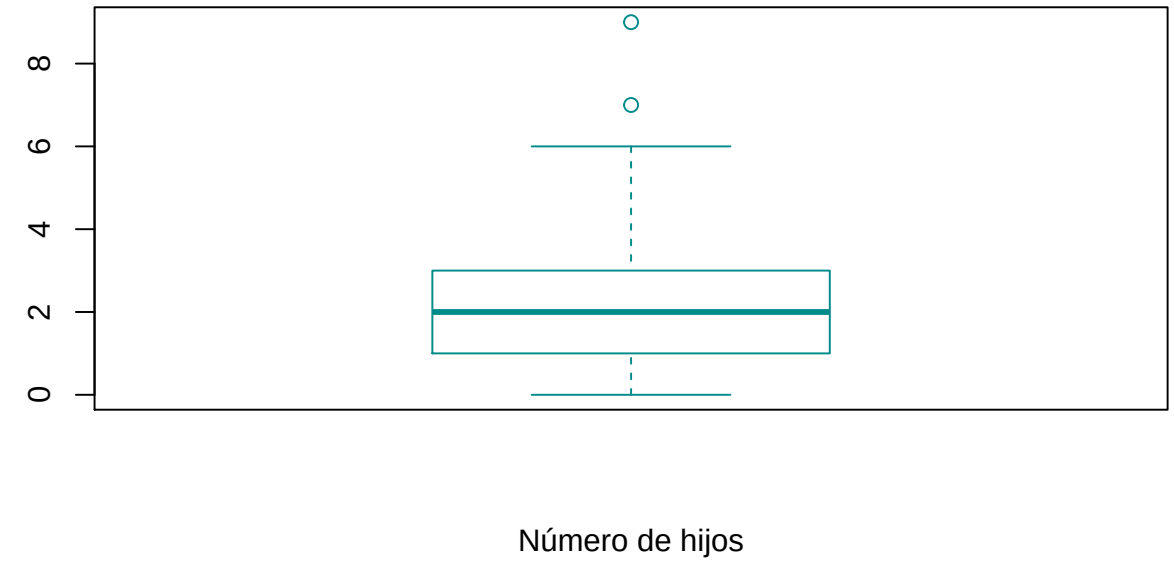
graf_disc(datos, X, nombre, título)

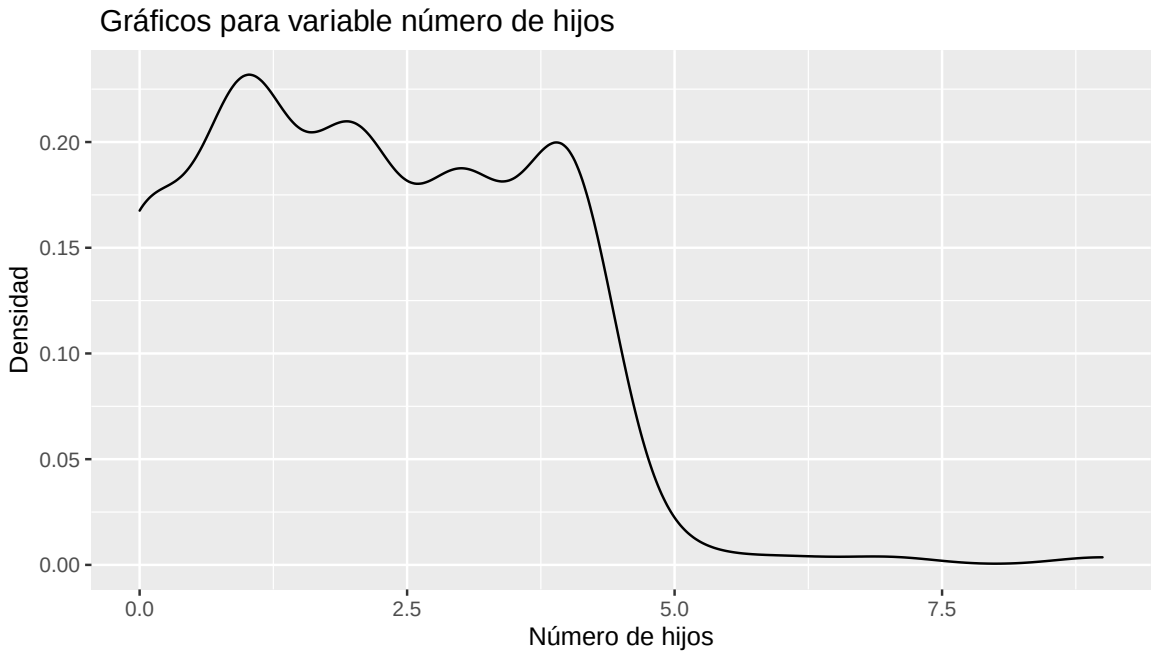
Argumentos:

- Datos: data.frame con los datos a analizar.
- X: vector de respuestas de la variable cuantitativa a analizar (tamaño n).
- Nombre: cadena de caracteres indicando la pregunta realizada.
- Título: cadena de caracteres indicando el nombre que se le quiere dar a los gráficos.

```
graf_disc=function(datos, X, nombre, título){

  #Diagrama de caja
  boxplot(X, xlab= nombre, main = título,
          border= "cyan4")

  #Densidad
  print(ggplot(datos, aes(x = X)) +
          geom_density() +
          scale_x_continuous(nombre) +
          scale_y_continuous("Densidad") +
          labs(title = título))

}
```

```
nombre= "Número de hijos"
título=" Gráficos para variable número de hijos"
graf_disc(datos, V9, nombre, título)
```




Gráficos para variable número de hijos
Densidad
0.20
0.15
0.10
0.05
0.00
0.0
2.5
5.0
7.5
Número de hijos

Capítulo 3

Análisis conjunto de variables

3.1. Introducción.

Este capítulo se centra en el análisis estadístico para dos o más variables, especialmente en la relación que existe entre ellas.
En la primera parte, se representan las tablas de contingencia y de resúmenes numéricos para las variables y, en la segunda, se encuentra la visualización gráfica de la relación entre los distintos tipos de variable.

Se usa el mismo esquema que en el capítulo anterior, primero se explican los argumentos de entrada de cada función, el cuerpo, es decir, las funciones que componen cada función y, finalmente, la salida con los datos.

3.2. Tablas.

3.2.1. Tablas para dos variables

Se incluyen las tablas de frecuencia y de resúmenes numéricos de dos variables cruzando las diferentes modalidades.

3.2.1.1. Variable cualitativa con respecto a otra cualitativa

Esta función devuelve la tabla de frecuencias absolutas y relativas del cruce de dos variables cualitativas.

Uso:

freq.2cual(datos, X1,X2, etiquetasx1, etiquetasx2, nombrex1, nombrex2, título1, título2)

Argumentos:

- Datos: data.frame con los datos a analizar.
- X1: vector de respuestas de la variable cualitativa 1 (filas) (tamaño n).
- X2: vector de respuestas de la variable cualitativa 2 (columnas) (tamaño n).
- EtiquetasX1: vector de cadena de caracteres con las posibles respuestas no numéricas de la variable 1.
- EtiquetasX2: vector de cadena de caracteres con las posibles respuestas no numéricas de la variable 2.

- NombreX1: cadena de caracteres indicando la pregunta realizada para la variable x1.
- NombreX2: cadena de caracteres indicando la pregunta realizada para la variable x2.
- Título1: cadena de caracteres indicando el nombre que se le quiere dar a la tabla de frecuencias absolutas.
- Título2: cadena de caracteres indicando el nombre que se le quiere dar a la tabla de frecuencias relativas.

Es importante el orden en el que se introducen los argumentos.

```
freq.2cual= function(datos, X1, X2, etiquetasx1, etiquetasx2,
                     nombrex1, nombrex2, título1, título2 ) {
  datos1 <- mutate(datos, v1 = factor(X1, labels = etiquetasx1))
  datos2 <- mutate(datos, v2 = factor(X2, labels = etiquetasx2))
  tabla <- "(nombre1 = datos1$v1) + ( Total = 1 ) ~
            (nombre2 = datos2$v2) + ( Total = 1 )"
  tabla <- gsub("nombre1", nombrex1, tabla)
  tabla <- gsub("nombre2", nombrex2, tabla)
  tt <- tabular (tabla, data = datos1)


   Frec.abs=table(datos1$v1, datos2$v2)
   Frec.rel=round(prop.table(Frec.abs),2)
   tabla=cbind(Frec.rel )
   tabla=cbind(tabla, Total = rowSums(tabla))
   tabla=rbind(tabla, Total = colSums(tabla))
   kable(tabla, caption = título2) }
```

```
etiquetas1=c("Sevilla", "Madrid", "Barcelona","NSNC")
etiquetas2=c("Muy malo", "Malo", "Regular", "Bueno", "Muy bueno")
nombre1 = "Origen"
nombre2 = "Valoración"
título1 = "Frecuencia  absoluta"
título2 = "Frecuencia relativa"

freq.2cual(datos, V4, V5, etiquetas1,etiquetas2, nombre1, nombre2,
            título1, título2)
```

Cuadro 3.1: Frecuencia relativa

	Muy malo	Malo	Regular	Bueno	Muy bueno	Total
Sevilla	0.09	0.06	0.08	0.10	0.07	0.40
Madrid	0.03	0.06	0.05	0.08	0.07	0.29
Barcelona	0.06	0.04	0.06	0.05	0.07	0.28
NSNC	0.00	0.01	0.00	0.01	0.00	0.02
Total	0.18	0.17	0.19	0.24	0.21	0.99

3.2.1.2. Variable cuantitativa según variable cualitativa

Esta función devuelve la tabla de resúmenes numéricos básicos de la variable cuantitativa respecto de la variable cualitativa, incluyendo el tamaño muestral, media, mediana, desviación estándar e intervalo de confianza al 95 %.

Uso:

desc.cuan.cual(X1, X2, etiquetasX2, nombreX1, nombreX2, título)

Argumentos:

- Datos: data.frame con los datos a analizar.
- X1: vector de respuestas de la variable cuantitativa (tamaño n).
- X2: vector de respuestas de la variable cualitativa (tamaño n).
- EtiquetasX2: vector de cadena de caracteres con las posibles respuestas no numéricas de la variable cualitativa.
- Título: cadena de caracteres indicando el nombre que se le quiere dar a la tabla.

```
desc.cuan.cual= function (datos,X1,X2, etiquetasX2, título) {
  datos1 <- mutate(datos, v1 = factor(X2, labels = etiquetas2))

  IC1 <- function(X1){
    mean(X1) - qt(0.95, df = length(X1)- 1) * sd(X1) / sqrt(length(X1))}
  IC2 <- function(X1){
    mean(X1) + qt(0.95, df = length(X1)- 1) * sd(X1) / sqrt(length(X1))}
  Media <- function(x){mean(x)}
  Mediana<- function (x) {median(x)}
  Desv<- function (x) {sd(x)}

  tabla <- " ( nombre2 =datos1$v1 ) + ( Total = 1 ) ~
            (n=1)+ ( nombre1 = X1 ) * (Media + Mediana + Desv + IC1+ IC2)"

  tt <- tabular (tabla, data = datos1)

  df <- data.frame(matrix(unlist(tt), nrow=length(etiquetasX2)+1, byrow=F))
  colnames(df) = c("n", "Media", "Mediana", "Desv", "IC1", "IC2")
  rownames(df) = c(etiquetasX2,"Total")

  kable(df, digits = 2, caption = título)


}
```

```
etiquetas2=c("Hombre", "Mujer")
título= "Resumen medidas de variable cuantitativa
          respecto cualitativa: Estatura y Sexo"
desc.cuan.cual(datos, V8, V1, etiquetas2, título)
```

Cuadro 3.2: Resumen medidas de variable cuantitativa respecto cualitativa: Estatura y Sexo

	n	Media	Mediana	Desv	IC1	IC2
Hombre	108	1.72	1.77	0.15	1.70	1.75
Mujer	141	1.63	1.64	0.13	1.62	1.65
Total	249	1.67	1.67	0.14	1.66	1.69

3.2.2. Tablas para tres o más variables.

3.2.2.1. Variable cualitativa con respecto a otras dos variables cualitativas.

La tabla de frecuencias teniendo en cuenta las tres variables se genera mediante la función freq.cual.2cual.

Uso:

freq.cual.2cual(datos, X1, X2, X3, etiquetasx1, etiquetasx2, etiquetasx3, nombre1, nombre2, nombre3, título)

Argumentos:

- Datos: data.frame donde se encuentra X1, X2, y X3.
- X1: vector de respuestas de la variable cualitativa 1 (tamaño n).
- X2: vector de respuestas de la variable cualitativa 2 (tamaño n).
- X3: vector de respuestas de la variable cualitativa 3 (tamaño n).
- EtiquetasX1: cadena de caracteres con las posibles respuestas no numéricas de la variable 1.
- EtiquetasX2: cadena de caracteres con las posibles respuestas no numéricas de la variable 2.
- EtiquetasX3: cadena de caracteres con las posibles respuestas no numéricas de la variable 3.
- Nombre1: cadena de caracteres con el nombre de la variable 1.
- Nombre2: cadena de caracteres con el nombre de la variable 2.
- Nombre3: cadena de caracteres con el nombre de la variable 3.
- Título: cadena de caracteres indicando el nombre que se le quiere dar a la tabla.

```
freq.cual.2cual = function(datos,X1,X2, X3, etiquetasx1, etiquetasx2,
                           etiquetasx3, nombre1, nombre2,nombre3,título) {
  datos1 <- mutate(datos, v1 = factor(X1, labels = etiquetasx1))
  datos2 <- mutate(datos, v2 = factor(X2, labels = etiquetasx2))
  datos3 <- mutate(datos, v3 = factor(X3, labels = etiquetasx3))

  tabla <- "(nombre1 = datos1$v1 ) * ( nombre2 = datos2$v2 ) +
  ( Total = 1 ) ~  (  nombre3 = datos3$v3 )+ ( Total = 1 )"
  tabla <- gsub("nombre1", nombre1, tabla)
  tabla <- gsub("nombre2", nombre2, tabla)
```

```
  tabla <- gsub("nombre3", nombre3, tabla)
  tt <- tabular (tabla, data = datos1)

  tt

  }
```

```
etiquetas1=c("Hombre", "Mujer")
etiquetas2=c("Sevilla", "Madrid", "Barcelona", "NSNC")
etiquetas3=c(" Muy malo", "Malo", "Regular", "Bueno", "Muy bueno")
nombre1="Sexo"
nombre2= "Origen"
nombre3= "valoración"
título= "Frecuencias abs. de tres var.cualitativas"
freq.cual.2cual (datos, datos$V1, datos$V4, datos$V5, etiquetas1,
       etiquetas2, etiquetas3, nombre1, nombre2, nombre3, título)
```

				valoración			
Sexo	Origen	Muy malo	Malo	Regular	Bueno	Muy bueno	Total
Hombre	Sevilla	4	6	6	11	8	35
	Madrid	5	9	5	9	13	41
	Barcelona	7	4	6	7	7	31
	NSNC	0	1	0	0	0	1
Mujer	Sevilla	18	10	13	15	10	66
	Madrid	3	7	7	11	4	32
	Barcelona	7	5	9	6	10	37
	NSNC	1	1	1	2	1	6
	Total	45	43	47	61	53	249

3.2.2.2. Variable cuantitativa según dos variables cualitativas

Esta función devuelve la tabla de resúmenes numéricos básicos de la variable cuantitativa respecto de dos variables cualitativas, incluyendo el tamaño muestral, media, mediana, desviación estándar e intervalo de confianza al 95 %.

Uso

desc.cuan.2cual(datos, X1, X2, X3, etiquetas2, etiquetas3, nombre1, nombre2, nombre3, título)

Argumentos:

- Datos: data.frame donde se encuentra X1, X2, y X3.
- X1: vector de respuestas de la variable cuantitativa 1 (tamaño n).
- X2: vector de respuestas de la variable cualitativa 2 (tamaño n).
- X3: vector de respuestas de la variable cualitativa 3 (tamaño n).
- Etiquetas2: cadena de caracteres con las posibles respuestas no numéricas de la variable 2.
- Etiquetas3: cadena de caracteres con las posibles respuestas no numéricas de la variable 3.

- Nombre1: cadena de caracteres con el nombre de la variable 1 (No admite espacios).
- Nombre2: cadena de caracteres con el nombre de la variable 2 (No admite espacios).
- Nombre3: cadena de caracteres con el nombre de la variable 3 (No admite espacios).
- Título: cadena de caracteres indicando el nombre que se le quiere dar a la tabla.

```
desc.cuan.2cual= function (datos, X1, X2, X3, etiquetas2, etiquetas3,
                           nombre1, nombre2, nombre3, título) {
  datos1 <- mutate(datos, v1 = factor(X2, labels = etiquetas2))
  datos2 <- mutate(datos, v2 = factor(X3, labels = etiquetas3))

  IC1 <- function(X1){
    mean(X1) - qt(0.95, df = length(X1)- 1) * sd(X1) / sqrt(length(X1))}
  IC2 <- function(X1){
    mean(X1) + qt(0.95, df = length(X1)- 1) * sd(X1) / sqrt(length(X1))}
  Media <- function(x){mean(x)}
  Mediana<- function (x) {median(x)}
  Desv<- function (x) {sd(x)}

  tabla <- "(nombre2 =datos1$v1 )+( nombre3 =datos2$v2 )+( Total = 1 ) ~
            (n=1)+ ( nombre1 = X1 )*(Media + Mediana + Desv + IC1+ IC2)"
  tabla <- gsub("nombre1", nombre1, tabla)
  tabla <- gsub("nombre2", nombre2, tabla)
  tabla <- gsub("nombre3", nombre3, tabla)
  tt <- tabular (tabla, data = datos1)

  tt
   }
```

```
etiquetas2=c("Hombre", "Mujer")
etiquetas3=c("Sevilla", "Madrid", "Barcelona","NSNC")
nombre1="Estatura"
nombre2="Sexo"
nombre3= "Origen"
título= "Resumen medidas de variable
             cuantitativa respecto dos cualitativas"
desc.cuan.2cual(datos,datos$V8, datos$V1, datos$V4 , etiquetas2,
               etiquetas3 ,nombre1, nombre2, nombre3, título)
```

		Estatura					
		n	Media	Mediana	Desv	IC1	IC2
Sexo	Hombre	108	1.724	1.77	0.1494	1.700	1.748
	Mujer	141	1.633	1.64	0.1271	1.615	1.651
Origen	Sevilla	101	1.653	1.66	0.1456	1.629	1.678
	Madrid	73	1.696	1.72	0.1469	1.667	1.725
	Barcelona	68	1.669	1.66	0.1390	1.641	1.697
	NSNC	7	1.734	1.71	0.1133	1.651	1.817
	Total	249	1.672	1.67	0.1441	1.657	1.688

3.3. Visualización gráfica

3.3.1. Gráficos para dos variables.

3.3.1.1. Variable cualitativa en función de otra cualitativa

Esta función genera un gráfico de barras para ver la relación existente entre dos variables cualitativas.

Uso:

graf.cual.cual(datos, x1, x2, etiquetasx1, etiquetasx2, nombrex1, nombrex2)

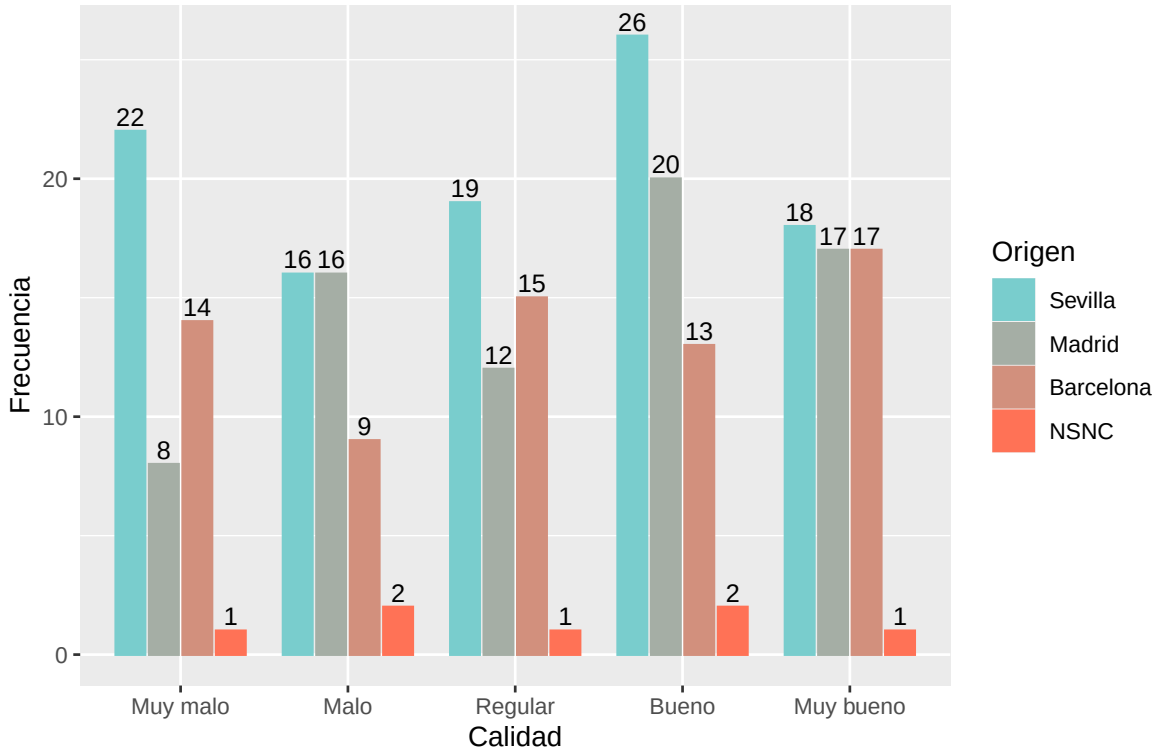
Argumentos:

- Datos: data.frame con los datos a analizar.
- X1: vector de respuestas de la variable cualitativa (tamaño n).
- X2: vector de respuestas de la variable cualitativa (tamaño n).
- EtiquetasX1: vector de cadena de caracteres con las posibles respuestas no numéricas de la variable 1.
- EtiquetasX2: vector de cadena de caracteres con las posibles respuestas no numéricas de la variable 2.
- NombreX1: cadena de caracteres indicando la pregunta realizada para la variable x1.
- NombreX2: cadena de caracteres indicando la pregunta realizada para la variable x2.

```
graf.cual.cual=function(datos,x1,x2, etiquetasx1,
                        etiquetasx2, nombrex1, nombrex2) {

   #Transforma como factores las variables originales x1 y x2
    datos <- mutate(datos, factor1 = factor(x1,labels = etiquetasx1))
    datos <- mutate(datos, factor2 = factor(x2,labels = etiquetasx2))

    mis.colores <- colorRampPalette(c("darkslategray3", "coral1"))
    n2=length(etiquetasx2)
    n1=length(etiquetasx1)

    datos3 <- datos %>%
    filter(factor2 %in% etiquetasx2 )%>%
    group_by(factor1, factor2) %>%
    summarise(Frecuencia = n())

 #Gráfico de barras
   print(ggplot(datos3, aes(x = factor1, y= Frecuencia )) +
         geom_bar(aes(color = factor2, fill = factor2),
                  stat = "identity", position = position_dodge(0.8),
                  width = 0.7) +
        labs(title = "", x = nombrex1)+
        scale_color_manual(name= nombrex2,values = mis.colores(n2))+
        scale_fill_manual(name= nombrex2,values = mis.colores(n2))+
```

```
        geom_text( aes(label = Frecuencia, group = factor2),
                   position = position_dodge(0.8),
                   vjust = -0.3, size = 3.5))


}
```

```
etiquetasx2=c("Sevilla", "Madrid", "Barcelona","NSNC")
etiquetasx1=c("Muy malo", "Malo", "Regular", "Bueno", "Muy bueno")
nombrex1="Calidad"
nombrex2="Origen"
graf.cual.cual(datos, datos$V5,datos$V4,
               etiquetasx1, etiquetasx2 , nombrex1, nombrex2)
```



3.3.1.2. Variable cuantitativa en función de una cualitativa.

Esta función genera el histograma, gráfico de densidad, caja y bigote y tipo jitter de la variable cuantitativa en función de la cualitativa.

Uso: graf.cont.cual (X1, X2, etiquetasx1, nombrex1,nombrex2, título)

Argumentos:

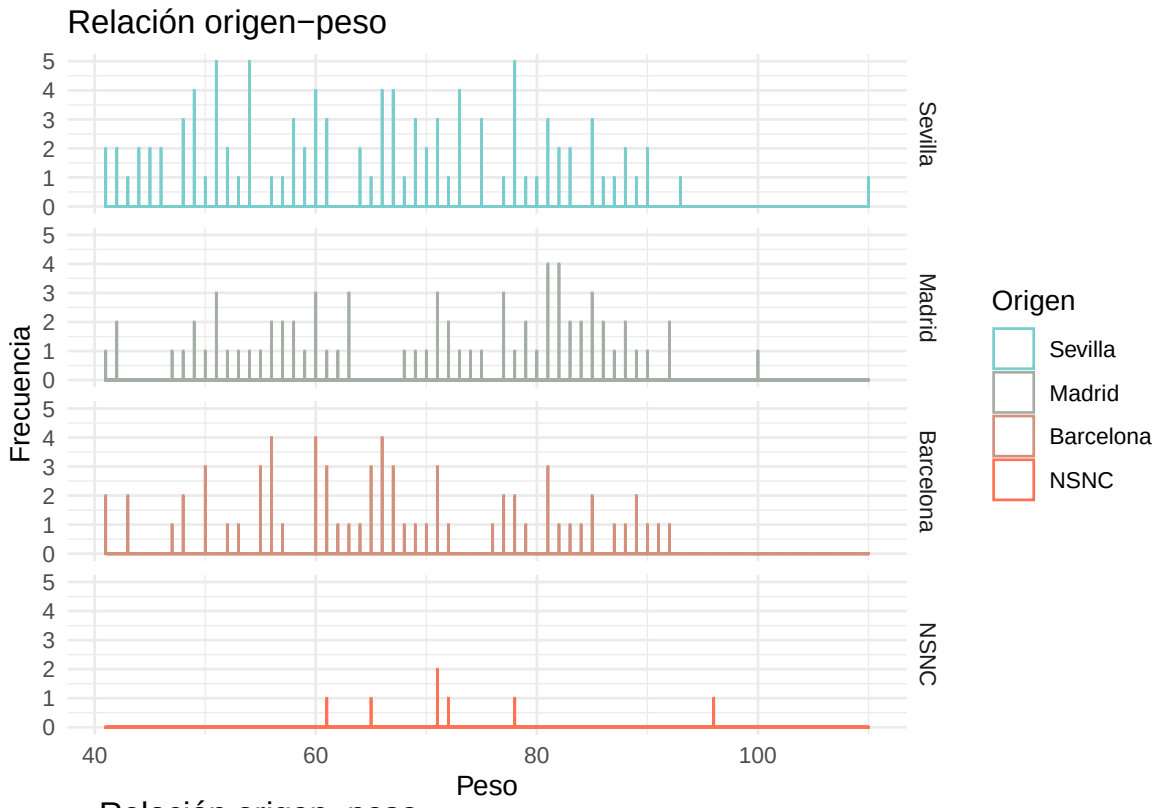
- Datos: data.frame con los datos a analizar.
- X1: vector de respuestas de la variable cualitativa (tamaño n).
- X2: vector de respuestas de la variable continua (tamaño n).
- EtiquetasX1: vector de cadenas de caracteres con las posibles respuestas no numéricas de la variable 1.
- NombreX1: cadena de caracteres indicando la pregunta realizada para la variable x1.

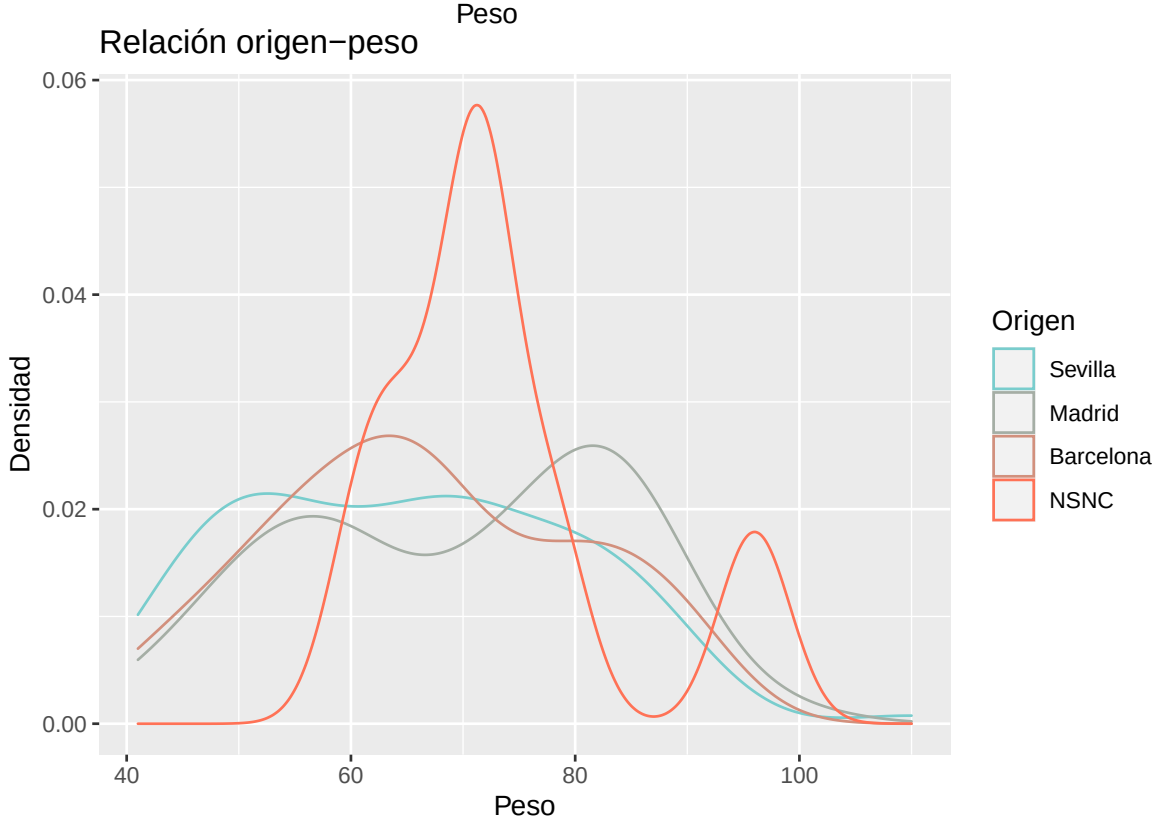
- NombreX2: cadena de caracteres indicando la pregunta realizada para la variable x2.
- Título: cadena de caracteres indicando el nombre que se le quiere dar a los gráficos.

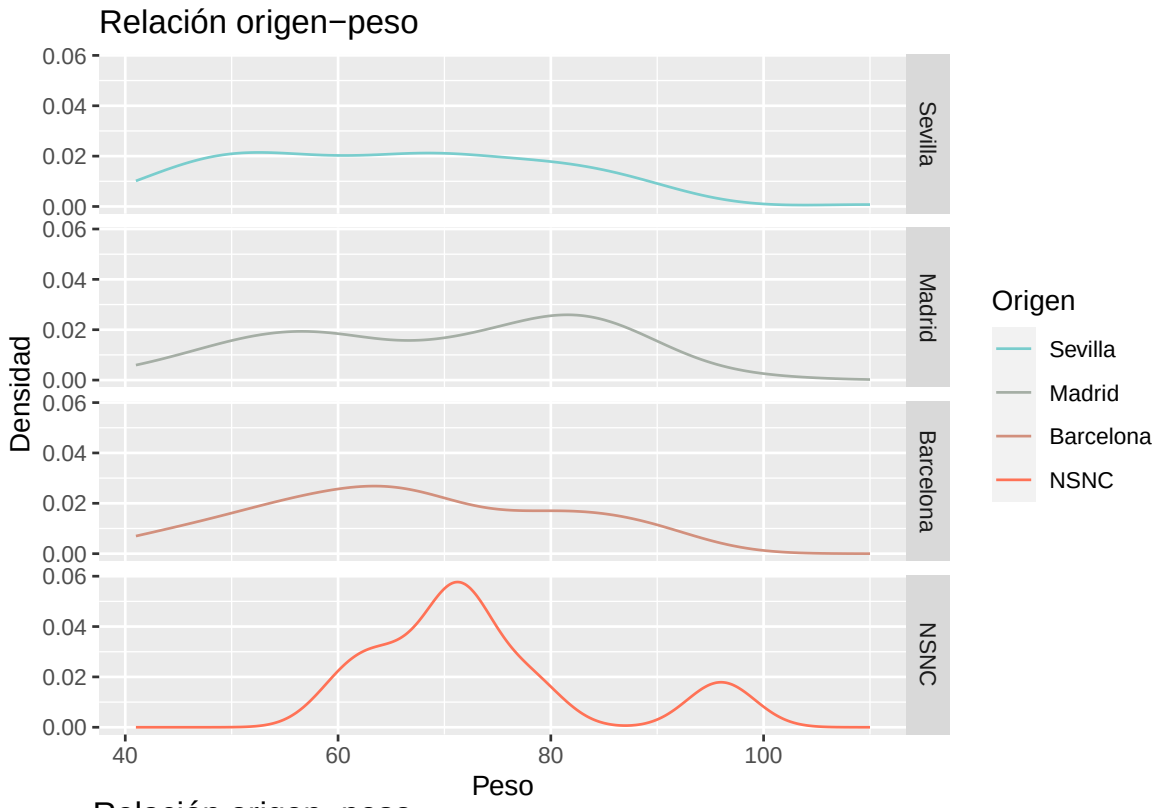
```
graf.cont.cual=function(X1, X2, etiquetasx1, nombrex1,nombrex2, título) {

  datos <- mutate(datos, factor1= factor(X1,labels = etiquetasx1))

  mis.colores <- colorRampPalette(c("darkslategray3", "coral1"))
  netiq=length(etiquetas)


  #Histograma
  print(ggplot(datos, aes(x=X2,color=factor1, fill=factor1)) +
    geom_histogram(binwidth = 0.01, fill="white") +
    labs(x = nombre2) +
    theme_minimal() +
    facet_grid(factor1 ~.)+
    scale_color_manual(name=nombre1,
                       values=mis.colores(netiq),labels=etiquetasx1)+
    scale_fill_manual(name=nombrex1,
                       values=mis.colores(netiq),labels=etiquetasx1)+
    labs(title = título, x = nombrex2, y = "Frecuencia")+
    theme_minimal())


   #Densidades por categorías en un mismo gráfico
  print(ggplot(datos, aes(x=X2, color=factor1)) +
    geom_density()+
    scale_color_manual(name=nombrex1,
                       values=mis.colores(netiq),labels=etiquetasx1)+
      labs(title = título , x = nombrex2, y = "Densidad"))

   # Densidades de la variable continua para cada
      # categoría en gráficos separados
  print(ggplot(datos, aes(x=X2, color=factor1)) +
          geom_line(stat="Density") +
          facet_grid(factor1 ~.)+
          scale_color_manual(name=nombrex1,
                       values=mis.colores(netiq),labels=etiquetasx1)+
          labs(title = título , x = nombrex2, y = "Densidad"))

   # Gráfico de caja y bigote por categorías
  print(ggplot(datos, aes(factor1,X2)) +
        geom_boxplot(aes(color = factor1 ))+
        scale_color_manual(name=nombrex1, values = mis.colores(netiq))+
        labs(title = título,
          x = nombrex1, y = nombrex2)+
```

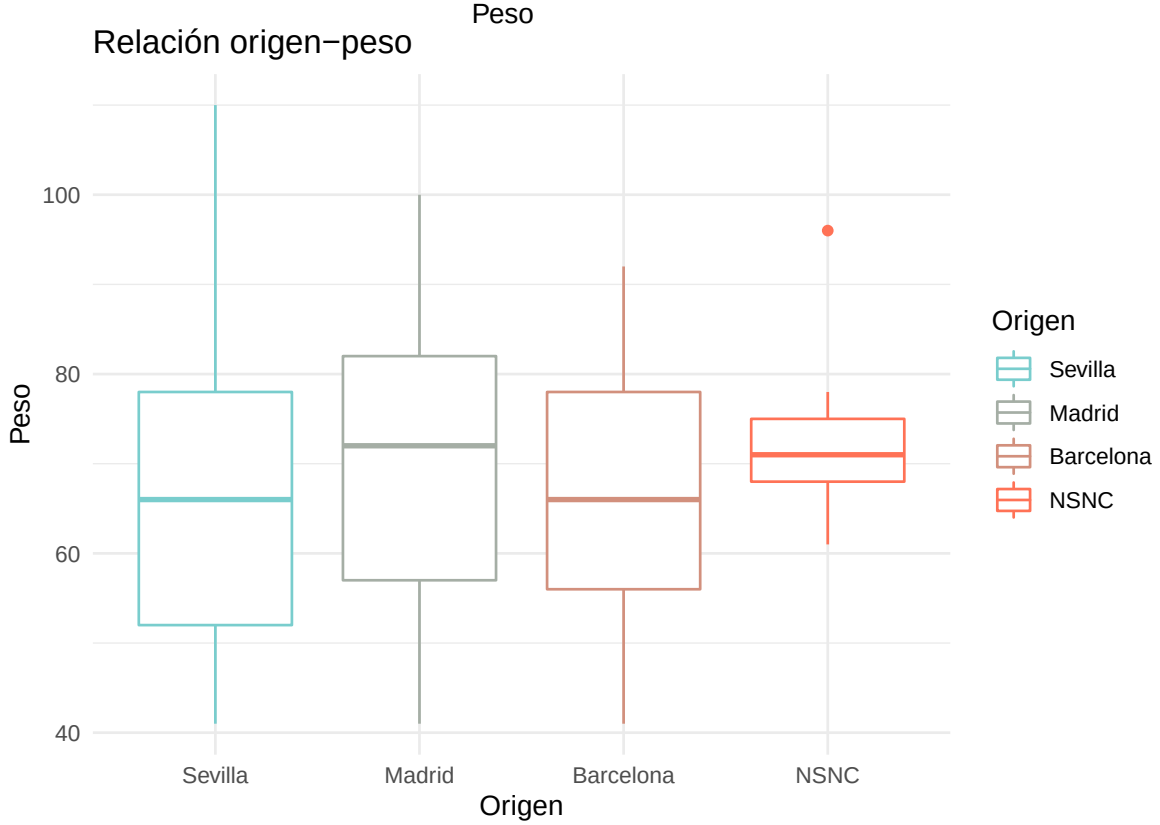
```
        theme_minimal())

   #Gráfico tipo jitter
  print(ggplot(datos, aes(x=factor1, y= X2)) +
        geom_jitter(aes(colour = factor1))+
        scale_color_manual(name=nombrex1,
                           values=mis.colores(netiq),labels=etiquetasx1)+
        labs(title = título , x = nombrex1, y = nombrex2))

    #Mezcla gráfico caja y bigote y tipo jitter.
  print(ggplot(datos, aes(x = factor1, y = X2)) +
        geom_jitter(aes(color = factor1), size = 1, alpha = 0.7) +
        geom_boxplot(aes(color = factor1), alpha = 0.7) +
        scale_color_manual(name=nombrex1,values=mis.colores(netiq),
        labels=etiquetasx1)+
        xlab(nombrex1) +
        ylab(nombrex2) +
        ggtitle(título) +
        theme_minimal())


 }
```

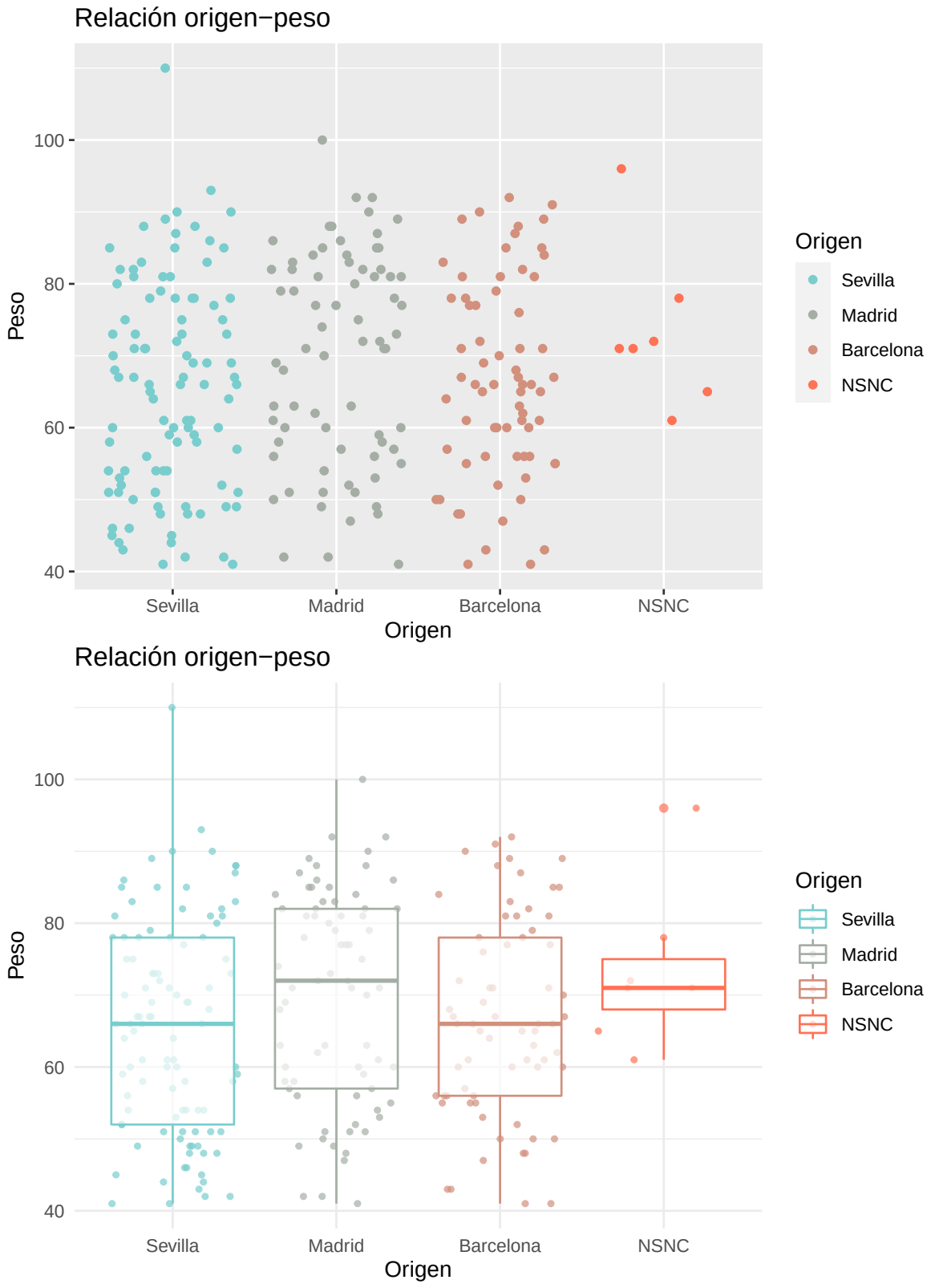
```
etiquetas=c("Sevilla", "Madrid", "Barcelona","NSNC")
nombre1="Origen"
nombre2="Peso"
título= "Relación origen-peso"
graf.cont.cual(datos$V4, datos$V10 ,etiquetas, nombre1, nombre2, título)
```


Relación origen−peso
Frecuencia
Peso
Sevilla
Madrid
Barcelona
NSNC
Origen
Sevilla
Madrid
Barcelona
NSNC


Relación origen−peso
Densidad
Peso
Origen
Sevilla
Madrid
Barcelona
NSNC


Relación origen−peso
Densidad
Peso
Sevilla
Madrid
Barcelona
NSNC
Origen
0.06
0.04
0.02
0.00
40
60
80
100


Relación origen−peso
Peso
Origen
Sevilla
Madrid
Barcelona
NSNC
100
80
60
40



3.3.1.3. Variable continua respecto de otra variable continua

La siguiente función muestra el gráfico de dispersión entre dos variables continuas dadas.

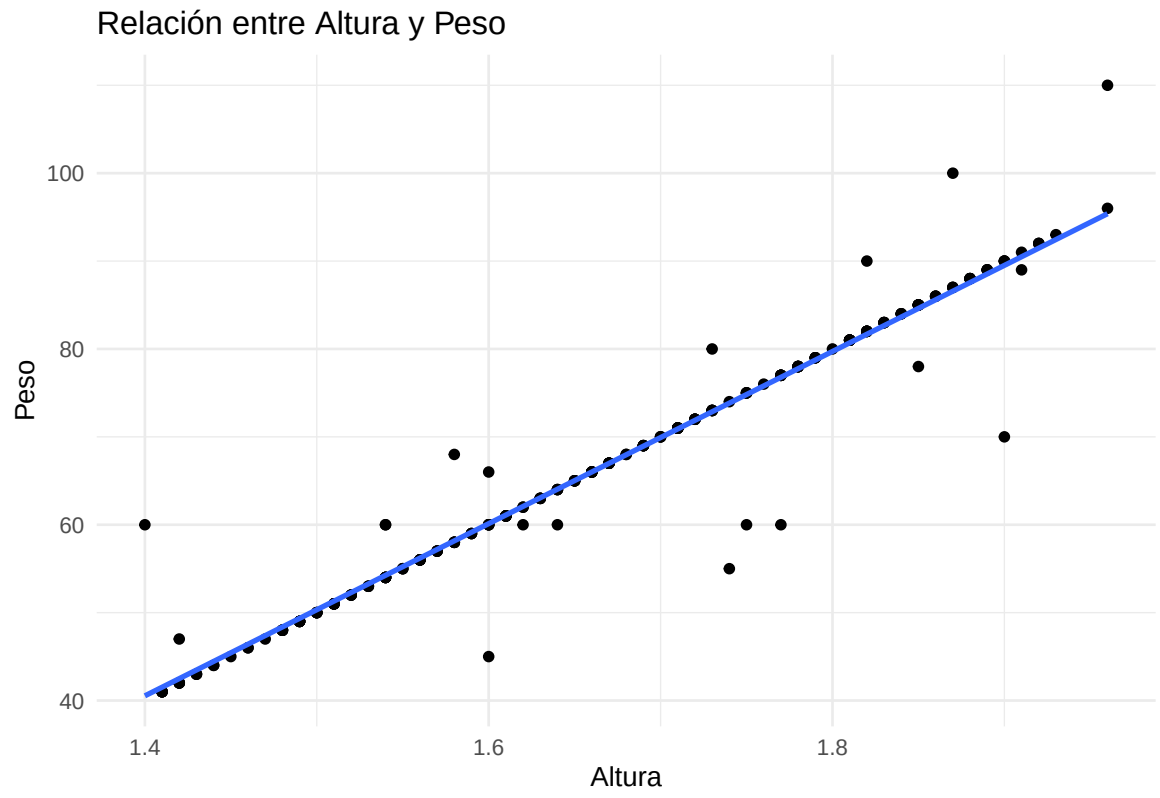
Uso: graf.cuan.cuan(X1, X2, nombrex1 ,nombrex2, título)

Argumentos:

- X1: vector de respuestas de la variable cuantitativa (tamaño n).
- X2: vector de respuestas de la variable cuantitativa (tamaño n).
- NombreX1: cadena de caracteres indicando la pregunta realizada para la variable x1.
- NombreX2: cadena de caracteres indicando la pregunta realizada para la variable x2.
- Título: cadena de caracteres indicando el nombre que se le quiere dar al gráfico.

```
graf.cuan.cuan=function(X1, X2, nombrex1 ,nombrex2, título) {

  print(ggplot(datos,aes(x=X1, y=X2)) +
  geom_point() +
  theme_minimal() +
  labs(title = título, x= nombrex1, y=nombrex2)+
  geom_smooth(method = "lm", se = F))

     }
```

```
nombrex1= "Altura"
nombrex2="Peso"
título= "Relación entre Altura y Peso"


graf.cuan.cuan(datos$V8, datos$V10, nombrex1, nombrex2, título)
```



3.3.2. Gráficos para más de tres variables.

3.3.2.1. Relación entre dos variables continuas clasificadas por una nominal

Esta función relaciona dos variables continuas según una variable nominal, muestra los gráficos de dispersión por separado.

Uso:

graf.2cont.nom (datos, X1, X2, X3, nombre1, nombre2, nombre3, etiquetas, título)

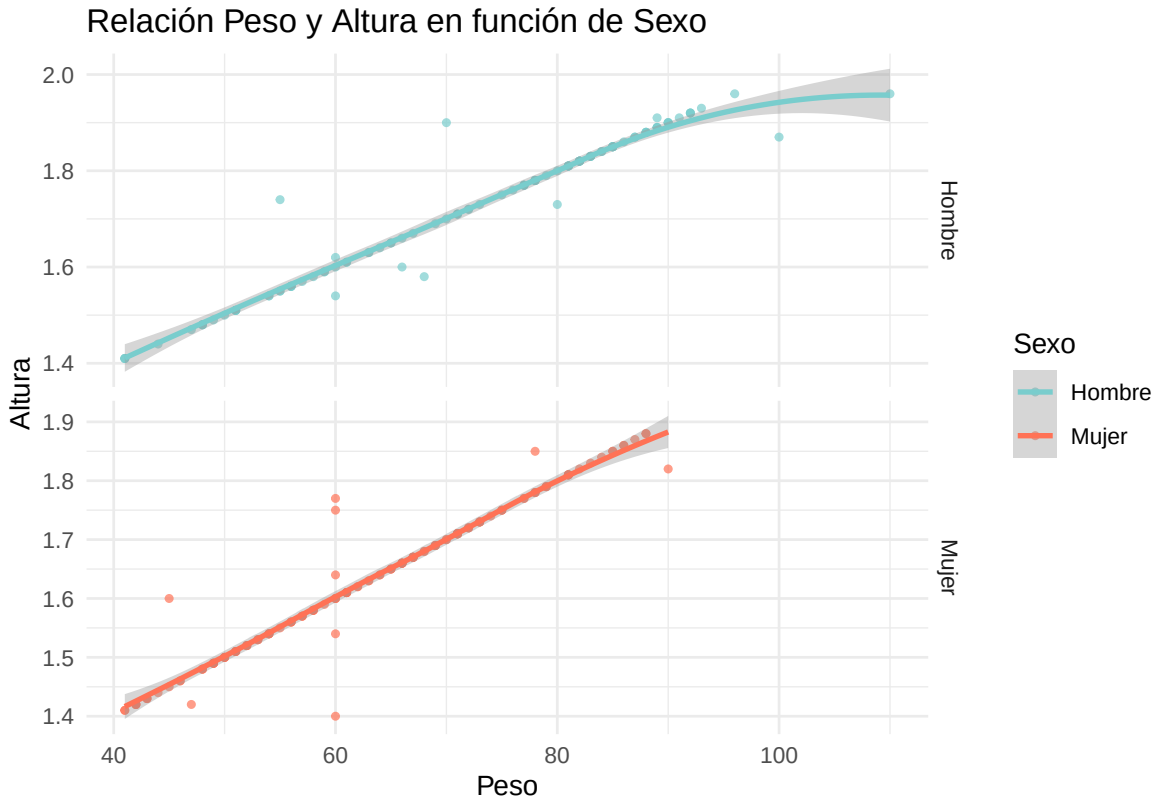
Argumentos:

- Datos: data.frame donde se encuentra X.
- X1: vector de respuestas de la variable continua 1 (eje x) (tamaño n).
- X2: vector de respuestas de la variable continua 2 (eje y) (tamaño n).
- X3: vector de respuestas de la variable nominal que definirá los grupos (tamaño n).
- Etiquetas: vector de caracteres con las posibles respuestas de la variable 3.
- Nombre1: cadena de caracteres con el nombre de la variable 1.
- Nombre2: cadena de caracteres con el nombre de la variable 2.
- Nombre3: cadena de caracteres con el nombre de la variable 3.
- Título: cadena de caracteres indicando el nombre que se le quiere dar al gráfico.

```
graf.2cont.nom= function(datos, X1, X2, X3, nombre1, nombre2,
                         nombre3, etiquetas, título) {

  datos <- mutate(datos, factor= factor(X3, labels = etiquetas))


  mis.colores <- colorRampPalette(c("darkslategray3", "coral1"))
  netiq=length(etiquetas)

  #Gráfico de dispersión
  ggplot(data = datos, aes(x = X1, y = X2)) +
    geom_point(aes(color = factor), size = 1, alpha = 0.7) +
    geom_smooth(aes(color = factor)) +
    facet_grid(factor~., scales = 'free') +
    scale_color_manual(name= nombre3 ,values = mis.colores(netiq))+
    xlab(nombre1) +
    ylab(nombre2) +
    ggtitle(título) +
    theme_minimal()


}
```

```
etiquetas=c("Hombre", "Mujer")
nombre2="Altura"
nombre1="Peso"
nombre3= "Sexo "
título= "Relación Peso y Altura en función de Sexo"
```

```
graf.2cont.nom(datos, V10, V8, V1,
               nombre1, nombre2, nombre3, etiquetas, título)
```



3.3.2.2. Relación entre dos variables continuas clasificadas por dos variables nominales

Esta función relaciona dos variables continuas según una variable nominal, cada una por separado.

Uso:

graf.2cont.2nom= function(datos, X1, X2, X3, X4, nombre1, nombre2, nombre3, nombre4, etiquetas1, etiquetas2, título)

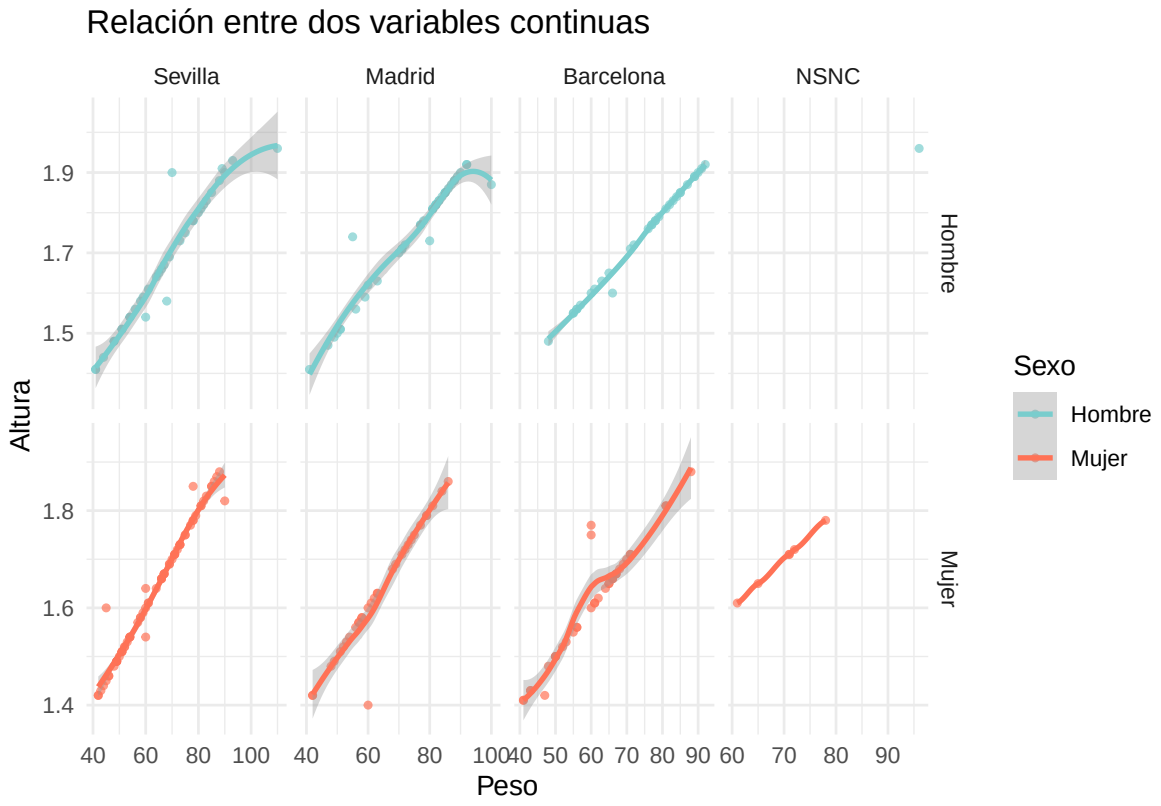
Argumentos:

- Datos: data.frame donde se encuentran los datos de la encuesta.
- X1: vector de respuestas de la variable cuantitativa 1 (tamaño n).
- X2: vector de respuestas de la variable cuantitativa 2 (tamaño n).
- X3: vector de respuestas de la variable cualitativa 3 (tamaño n).
- X4: vector de respuestas de la variable cualitativa 4 (tamaño n).
- EtiquetasX3: cadena de caracteres con las posibles respuestas no numéricas de la variable 3.
- EtiquetasX4: cadena de caracteres con las posibles respuestas no numéricas de la variable 4.
- Nombre1: cadena de caracteres con el nombre de la variable 1.
- Nombre2: cadena de caracteres con el nombre de la variable 2.
- Nombre3: cadena de caracteres con el nombre de la variable 3.

- Nombre4: cadena de caracteres con el nombre de la variable 4.
- Título: cadena de caracteres indicando el nombre que se le quiere dar a los gráficos.

```
graf.2cont.2nom= function(datos, X1, X2, X3, X4, nombre1, nombre2,
                  nombre3, nombre4, etiquetas1, etiquetas2, título) {

  datos <- mutate(datos, factor1= factor(X3, labels = etiquetas1))
  datos <- mutate(datos, factor2= factor(X4, labels = etiquetas2))

  mis.colores <- colorRampPalette(c("darkslategray3", "coral1"))
  netiq=length(etiquetas)

  #Gráfico de dispersión
  ggplot(data = datos, aes(x = X1, y = X2)) +
    geom_point(aes(color = factor1), size = 1, alpha = 0.7) +
    geom_smooth(aes(color = factor1)) +
    facet_grid(factor1~factor2, scales = 'free') +
    scale_color_manual(name= nombre3 ,values = mis.colores(netiq))+
    xlab(nombre1) +
    ylab(nombre2) +
    ggtitle(título) +
    theme_minimal()


}
```

```
etiquetas1=c("Hombre", "Mujer")
etiquetas2=c("Sevilla", "Madrid", "Barcelona","NSNC")
nombre1="Peso"
nombre2="Altura"
nombre3="Sexo "
nombre4="Ciudad"
título= 'Relación entre dos variables continuas'
graf.2cont.2nom(datos, V10, V8, V1, V4,
              nombre1, nombre2, nombre3, nombre4,
              etiquetas1, etiquetas2, título)
```


Relación entre dos variables continuas
Sevilla
Madrid
Barcelona
NSNC
Hombre
Mujer
Altura
Peso
1.9
1.7
1.5
1.8
1.6
1.4
40
60
80
100
40
50
60
70
80
90
Sexo
Hombre
Mujer

Capítulo 4

Medidas de asociación

4.1. Introducción

Este capítulo se centra en el estudio de la relación que existe entre dos o más variables, su grado de asociación, el aumento o disminución conjunto o inverso, la influencia de una sobre otra. El tipo de correlación y la intensidad que existe entre pares de variables se pueden medir a través de distintos coeficientes o medidas de asociación. Se incluyen aquí tres de las medidas más utilizadas: coeficiente de correlación de Pearson, coeficiente de correlación entre rangos de Spearman y el coeficiente de correlación Tau de Kendall. Además de la medida de asociación lambda de Kruskal Goodman. Todas ellas son medidas basadas en el criterio de reducción proporcional del error en el análisis de tabulación cruzada. En la segunda parte de este capítulo se muestran las funciones creadas para analizar estas medidas de asociación tanto analítica como gráficamente.

Sheskin (2003) expresa que las medidas de asociación no son pruebas estadísticas inferenciales, por el contrario, son medidas estadísticas descriptivas que demuestran la dirección, fuerza o grado de relación entre variables. Existen dos estrategias básicas para obtener medidas de asociación:

- Basándose en el estadístico chi-cuadrado, comparando las frecuencias observadas y esperadas en las tablas de contingencia.
- Medidas basadas en el criterio de reducción proporcional del error (R.P.E), que consiste en la razón de la cantidad de error cometido al predecir la variable objetivo en dos situaciones distintas. Por ejemplo, el coeficiente de correlación lineal.

Consideramos como método de predicción la esperanza y como medida del error el error cuadrático medio:

- Predicción de Y sin información de X : $E(Y) = \mu_y$; Error: $E[(Y - \mu_y)^2] = \sigma_y^2$
- Predicción de Y con información de X : $E(Y \mid x) = \mu_y + \beta_1(X - \mu_x)$; Error: $\sigma_y^2(1 - \rho^2)$

Por tanto, la medida de la reducción proporcional del error será:

$$\rho^2 = \frac{\sigma_y^2 - \sigma_y^2(1 - \rho^2)}{\sigma_y^2}$$

Se consideran los siguientes errores:

- **Error sin información**. Error de predicción cuando no se dispone de información adicional alguna, salvo la distribución propia de la variable dependiente: $ERROR_S$
- **Error con información**. Error de predicción cuando ésta se realiza con la información adicional de la variable independiente o explictiva: $ERROR_C$

$$MEDIDA(RPE) = \frac{ERROR_S - ERROR_C}{ERROR_S}$$

Propiedades:

1. $0 \leq RPE \leq 1$
2. $RPE = 0 \Rightarrow$ reducción del error nula $\Rightarrow$ Asociación nula
3. $RPE = 0 \Rightarrow$ reducción del error completa $\Rightarrow$ Asociación perfecta

4.2. Algunas medidas de asociación

4.2.1. Correlación de Pearson

Se trata del método más usado para medir el grado de relación entre dos variables continuas. Se define como la razón de la covarianza de las dos variables entre el producto de sus respectivas desviaciones estándar. El coeficiente de Pearson toma valores en [-1,1]. Se denota por la letra griega ρ (para mas detalles, véase Chok 2010)

$$\rho = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

Su versión muestral es:

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

donde:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad ; \quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

4.2.2. Correlación de Spearman

Es la correlación entre los rangos de los datos ordenados de las dos variables, por ello es aplicable a variables ordinales o a variables cuantitativas siempre que existan suficientes evidencias de que no siguen un comportamiento normal. Es, por tanto, una medida no paramétrica de la dependencia estadística entre las ordenaciones o rankings de las dos variables.

Sea (x_i, y_i), $i = 1, \ldots, n$, una muestra de dos variables ordinales, y consideremos los rangos: $u_i = rango\ de\ x_i\ en\ \{x_1, ..., x_n\}$, $v_i = rango\ de\ y_i\ en\ \{y_1, ..., y_n\}$

Así, se consideran los n puntos: (u_i, v_i). El **coeficiente de correlación entre rangos de Spearman** , r_s, para la muestra (x_i, y_i), $i = 1, \ldots, n$, se define como el coeficiente de correlación lineal para los datos bidimensionales (u_i, v_i), $i = 1, \ldots, n$.

$$r_s = 1 - \frac{6}{n(n^2-1)} \sum_{i=i}^{n} d_i^2, \quad donde \;\; d_i = u_i - v_i$$

Por tanto, es una medida simétrica que toma valores en [-1,1]

4.2.3. Correlación de Kendall

El coeficiente de correlación de Kendall, generalmente conocido como el Coeficiente Tau de Kendall, es una medida de la asociación entre dos variables ordinales. Básicamente, es una medida de la capacidad de predecir el orden entre dos observaciones de una de las variables conocido el orden de la otra variable. Fue propuesta por (Kendall 1938)

Consideremos dos variables ordinales (X,Y). Si dadas dos observaciones de las mismas (X1,Y1), (X2,Y2) pretendemos predecir la ordenación de la variable Y:

$$Y1 > Y2 \quad \text{ó} \quad Y2 > Y1$$

Se puede hacer:

1. Sin información adicional alguna.
2. Con la información adicional de la ordenación en la variable X.

Consideremos las funciones indicadores:

$$I_x = \begin{cases} 0 & \text{si } X1 < X2 \\ 1 & \text{si } X1 > X2 \end{cases} \qquad I_y = \begin{cases} 0 & \text{si } Y1 < Y2 \\ 1 & \text{si } Y1 > Y2 \end{cases}$$

En el primer caso, la predicción entre $I_y = 1$ ó $I_y = 1$ se realizará de forma aleatoria, es decir

$$E_S = P[error_S] = 1/2$$

En el segundo caso, podemos predecir de la siguiente forma:

$$Si \; I_x = 0 \; entonces \; Iy = 0 \quad ; \; Si \; I_x = 1 \; entonces \; Iy = 1$$

y por tanto,

$$EC = P[error_C] = P[I_x = 0, I_y = 1] + P[I_x = 1, I_y = 0] = \pi_d$$

donde π_d es la probabilidad de discordancia, que será igual a 1- π_c, siendo π_c la probabilidad de concordancia.

En consecuencia, la medida RPE vendrá dada por

$$\frac{\frac{1}{2} - \pi_d}{\frac{1}{2}} = 1 - 2\pi_d = \pi_c + \pi_d - 2\pi_d = \pi_c - \pi_d$$

La segunda igualdad es cierta para variables continuas, pues $P[X1 = X2)] = p[Y1 = Y2] = 0$

Definición 1 Dada una variable aleatoria bidimensional (X,Y), ambas al menos en la escala ordinal. Se define la Tau de Kendall entre ambas como: $\tau = \pi_c - \pi_d$

Cuando las variables no son absolutamente continuas, se ha de buscar una solución al problema $P[(X1 - X2)(Y1 - Y2) = 0] \neq 0$

Se definen las variables aleatorias $A_{ij} = sign(X_j - X_i)(Y_j - Y_i)$.

Definición 2 Dada una variable aleatoria bidimensional (X,Y), al menos en la escala ordinal, y una muestra aleatoria de la misma, (X_i, Y_i) $i = 1, \ldots, n$, al estadístico

$$\frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} A_{ij}$$

se le denomina **coeficiente Tau de Kendall muestral** y se denota por $\hat{\tau}$

En la práctica se recurre a:

$$\left.\begin{array}{l} P = n\acute{u}m.\ de\ A_{ij}\ positivos \\ N = n\acute{u}m\ de\ A_{ij}\ negativos \end{array}\right\} \Rightarrow \hat{\tau} = \frac{P - N}{\binom{n}{2}}$$

El número total de parejas es T = $\binom{n}{2}$, de las cuales pueden ser:

Concordantes: $P = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_{ij} [\sum_{h=i+1}^{r} \sum_{k=j+1}^{c} n_{hk}]$

Discordantes: $N = \sum_{i=1}^{r-1} \sum_{j=1}^{c-1} n_i (c - j + 1) [\sum_{h=i+1}^{r} \sum_{k=c-j}^{c} n_{hk}]$

Conclusión:

Es importante estudiar la intensidad de una relación lineal entre dos variables. El coeficiente de correlación de Spearman es pertinente si se presenta uno de los siguientes casos: el primero, supongamos que se estudia la asociación lineal entre variables cuantitativas con escalas de medidas al menos de intervalos, y bajo esta condición sería conveniente el uso del coeficiente de Pearson, pero si estas variables no siguen un comportamiento normal en sus datos, necesariamente se debe estimar Spearman; el segundo, cuando ambas variables originales presentan escalas de medidas ordinales y su determinación es directa. Por último, el coeficiente de Kendall es adecuado cuando ambas variables presentan escalas de medidas ordinales. (Morales & Rodríguez 2016)

4.2.4. Lambda de Goodman Kruskal

Para estudiar la asociación entre dos variables nominales, es decir, categóricas sin un orden establecido, la mejor medida es la propuesta por Goodman y Kruskal. Es conocida como el **Coeficiente Lambda de Y sobre X**.

El método de predicción usado es la **moda** de la distribución o marginal o condicionada.

Sean X e Y dos variables nominales con modalidades $A_1..A_r$ y $B_1..B_c$, respectivamente.La distribución de probabilidad conjunta de X e Y es

$$P[X = A_i; Y = B_j] = p_{ij}$$

Las probabilidades marginales son:

$$p_j = P[Y = B_j] = \sum_{j=1}^{c} p_{ij} \qquad q_i = P[X = A_i] = \sum_{i=1}^{r} q_{ij}$$

Si se predice Y con la moda, sin información adicional de X, la probabilidad de error cometido es:

$$E_S^Y = P[error_S] = 1 - p_{max}, \ \ donde \ \ p_{max} = max_j \{p_j\}$$

Si se dispone de la información adicional de la X, $X = A_i$, la probabilidad de error será:

$$1 - pi^*, \ donde \ pi^* = max_j \ P[Y = B_j \mid X = a_i] = \frac{1}{q_i} max_j \{p_{ij}\} = \frac{1}{q_i} p_{i,max}$$

$$E_C^Y = P[error_C] = \sum_{i=1}^{r} q_i(1 - p_i^*) = 1 - \sum_{i=1}^{r} p_{i,max}$$

Luego, una medida de asociación basada en el criterio RPE viene dada por la expresión:

$$\lambda_{Y|X} = \frac{\sum_{i=i}^{r} p_{i,max} - p_{max}}{1 - p_{max}} = \frac{\sum_{i=1}^{r}[max_j \{p_{ij}\}] - max_j \{\sum_{i=1}^{r} p_{ij}\}}{1 - max_j \{\sum_{i=1}^{r} p_{ij}\}}$$

Dado que es asimétrica, $\lambda_{X|Y}$ no necesariamente coincide con $\lambda_{Y|X}$.

Para obtener una medida simétrica a partir de este esquema se procede de la siguiente manera:

$$\lambda = \frac{(E_S^Y + E_S^X) - (E_C^Y + E_C^X)}{E_S^Y + E_S^X}$$

denominada **Coeficiente Lambda simétrica**.

VERSIÓN MUESTRAl: Se sustituye las probabilidades por frecuencias relativas $f_{ij} = n_{ij}/n$. O bien, trabajando con frecuencias absolutas:

$$\hat{\lambda}_{Y|X} = \frac{\sum_{i=1}^{r} m_i^Y - M_Y}{n - M_Y} \ ; \quad \hat{\lambda} = \frac{\sum_{i=1}^{r} m_j^X - (M_Y + M_X)}{2n - (M_Y + M_X)}$$

$$donde \begin{cases} m_i^Y = frecuencia\, absoluta\, modal\, de\, Y \mid X = A_i \\ m_j^X = frecuencia\, absoluta\, modal\, de\, X \mid Y = B_j \\ M_Y = frecuencia\, absoluta\, modal\, de\, Y \\ M_X = frecuencia\, absoluta\, modal\, de\, X \end{cases}$$

El intervalo modal es el de mayor frecuencia absoluta. La frecuencia absoluta del intervalo modal son las frecuencias absolutas de los intervalos anterior y posterior respectivamente, al intervalo modal.

4.3. Ejemplos:

Para ilustrar algunas de las siguientes funciones se usan un conjunto de medidas corporales tomadas a un grupo de estudiantes en una universidad de Medellín. Las variables de la base de datos son (Hernández & Usuga 2019a)

Cuadro 4.1: Descripción de los datos

Variables	**Tipo**	**Descripción**
Edad	Cuantitativa	Edad del estudiante en años.
Peso	Cuantitativa	Peso del estudiante en kilogramos.
Altura	Cuantitativa	Estatura del estudiante en centímetros.
Sexo	Nominal	Género del estudiante.
Muneca	Cuantitativa	Diametro de la muñeca derecha en centímetros.
Biceps	Cuantitativa	Diametro del biceps derecho en centímetros.

4.3.1. Coeficiente de correlación

Devuelve el valor del coeficiente de correlación lineal en función del método indicado. Pearson, Kendall o Spearman.

Uso

correlacion(x, y , nombres, método)

Argumentos:

- X: vector numérico.
- Y: vector numérico.
- Nombres: vector de caracteres con las preguntas o afirmaciones de las variables (el número de columnas de data debe ser igual a p).
- Método: cadena de caracteres indicando qué coeficiente de correlación se debe calcular: “pearson”, “kendall” o “spearman”.

```
correlacion= function(X, Y , nombres, método){
  a= cor(X,Y , method = método)
  cat("El valor del coeficiente de correlación de", nombres[1],
    "y", nombres[2], "por el método \n", método,  "es",  round(a,4))
}
```

```
correlacion(datoscor1$edad, datoscor1$peso, c("edad","peso"), "pearson")
```

El valor del coeficiente de correlación de edad y peso por el método pearson es 0.5154

4.3.2. Matriz de correlaciones

La siguiente función genera una matriz con las correlaciones dos a dos de hasta p variables en función del método indicado.

Uso

matriz.cor(data, nombres, título, método)

Argumentos:

- Datos: data.frame formada por las p variables que se desea estudiar su correlación.
- Nombres: vector de caracteres con las preguntas o afirmaciones de las variables (el número de columnas de data debe ser igual a p).
- Título: cadena de caracteres indicando el nombre que se le quiere dar a la salida.
- Método: cadena de caracteres indicando qué coeficiente de correlación se debe calcular. "pearson", "kendall" o "spearman".

```
matriz.cor= function( data, nombres, título, método){
  colnames(data) <- nombres
  kable(cor(data, method = método) ,
        caption = título , booktabs = TRUE,escape=FALSE)
}
```

Para ilustrar la función se usan las variables ordinales de valoración de los datos simulados con el método Spearman.

```
data=cbind.data.frame(datos$V5,datos$V6,datos$V7)
nombres= c("Comida","Servicio","Limpieza")
título= "Relación entre las valoraciones "
matriz.cor(data, nombres , título, "spearman")
```

Cuadro 4.2: Relación entre las valoraciones

	Comida	Servicio	Limpieza
Comida	1.0000000	-0.0289983	-0.0082606
Servicio	-0.0289983	1.0000000	-0.0215291
Limpieza	-0.0082606	-0.0215291	1.0000000

4.3.3. Medida de asociación Lambda de Goodman Kruskal.

La siguiente función devuelve el coeficiente Lambda simétrico para variables nominales, además de un intervalo con un nivel de confianza dado.

Uso

asocnom(X, Y, conf, nombrex1, nombrex2)

Argumentos:

- X: Vector numérico de la variable 1.
- Y: Vector numérico de la variable 2.
- Conf: Valor que indica el nivel de confianza para crear el intervalo (0<conf<1).

- Nombrex: cadena de caracteres que indica nombre de la variable 1.
- Nombrey: cadena de caracteres que indica nombre de la variable 2.

```
asocnom= function(X, Y, conf, nombrex, nombrey) {
  a=Lambda(X, Y, conf.level = conf)

cat("El valor de Lambda de Goodman Kruskal para", nombrex,
    "y", nombrey, "es \n",  round(a[1],4) ,
    "con un intervalo de confianza al", conf*100,
    "% de: (",round(a[2],4),",",round(a[3],4),")")

 }
```

Para ilustrarlo hemos usado las variables V1 y V4 de los datos simulados

```
nombrex1="Sexo"
nombrex2="Ciudad de Origen"
conf=0.95
asocnom(datos$V1, datos$V4, conf, nombrex1, nombrex2)
```

El valor de Lambda de Goodman Kruskal para Sexo y Ciudad de Origen es 0.0586 con un intervalo de confianza al 95 % de: (0 , 0.1724)

4.3.4. Visualización gráfica matriz de correlaciones

La función graf.cor genera un gráfico muy visual de las correlaciones entre varias variables. Se muestra la matriz con los coeficientes de correlación. En la diagonal están las variables, por encima se sitúan círculos, entre más intensidad del color, ya sea azul o rojo, mayor es la correlación, colores tenúes significan correlación baja; el tamaño de los círculos está asociado al valor absoluto de correlación. Por debajo de la diagonal se observan los valores exactos de correlación. (Hernández & Usuga 2019b)

Uso

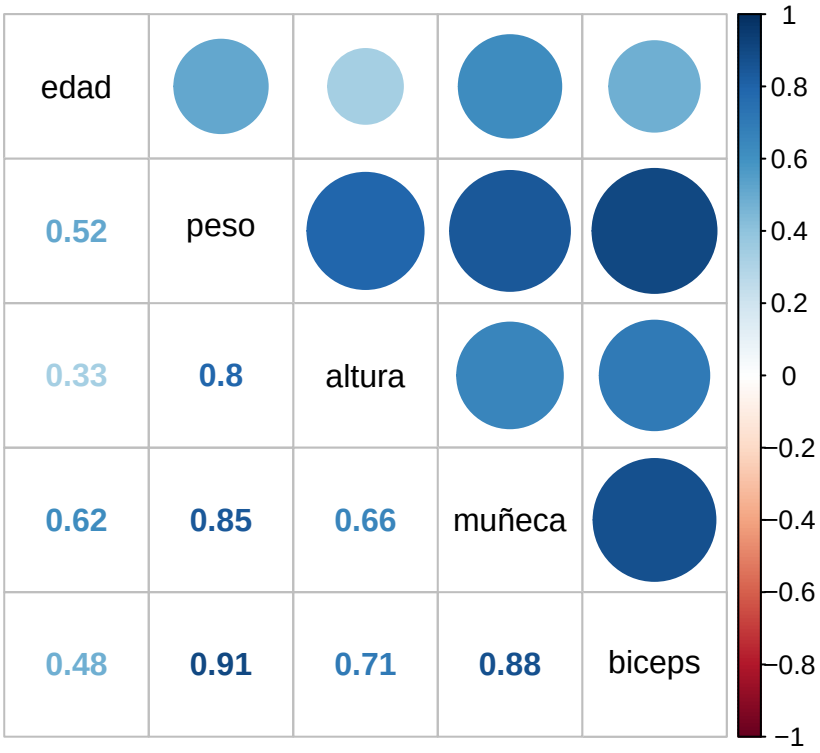
graf.cor(data, nombres, título, método)

Argumentos:

- Data: data.frame formada por las p variables que se desea estudiar su correlación.
- Nombres: vector de caracteres con los nombres de las variables (su longitud igual a p).
- Título: cadena de caracteres indicando el nombre que se le quiere dar a la salida.
- Método: cadena de caracteres indicando qué coeficiente de correlación se debe calcular: “pearson”, “kendall” o “spearman”.

```
graf.cor=function(data, nombres, título, método) {
  m=cor(data, method = método)
  corrplot.mixed(m, tl.pos= "d", tl.col="black")
  }
```

```
título= "Gráfico de correlación  "
graf.cor(datoscor1, nombres , título, "pearson")
```


edad
0.52
peso
0.33
0.8
altura
0.62
0.85
0.66
muñeca
0.48
0.91
0.71
0.88
biceps
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Capítulo 5

Creación del paquete en R

5.1. Introducción

Para completar esta memoria se ha creado un paquete en R, esto obliga a pulir las funciones y, sobre todo, a documentar todo el trabajo. Podremos usar el comando *?()* o help() para ver detalles de los parámetros, resultados y usos de cada función.
Además, es un método elegante de compartir el trabajo realizado. El paquete contendrá todas las funciones para el análisis de datos obtenidos a través de cuestionarios.

5.2. Proceso de creación

Para crear el paquete, se necesita tener instalado los paquetes *devtools* (herramientas de desarrollo de paquetes) y *roxygen2* (que permite generar muy fácilmente la documentación de ayuda de nuestro paquete).
El primer paso es crear el proyecto (.Rproj) como un R Package, se creará un directorio que contendrá una serie de archivos, los que hay que modificar manualmente serían:

- *DESCRIPTION*: en él se especifica la información general del paquete.

DESCRIPTION

```
Package: Cuestionarios
Type: Package
Title: Análisis de cuestionarios.
Version: 1.0
Date: 12-05-2020
Author: Celia Romero Gustos
Maintainer: <celiaromerogustos@gmail.com>
Description: Análisis de cuestionarios por tipo de variables.
License: GNU General
Encoding: UTF-8
RoxygenNote: 7.1.0
```

Figura 5.1: Descripción del paquete

- *R*: donde se almacenan los ficheros con cada una de las funciones. Para que R pueda generar automáticamente el archivo de ayuda incluimos la información antes del comienzo de la función en cada fichero.

```
#' @import dplyr
#' @import knitr
#' @title Tabla de frecuencias para variables binarias o dicotómicas.
#' @description Función que genera una tabla de frecuencias absolutas, relativas y acumuladas
#' para la variable x.
#' @param Datos data.frame con los datos a analizar.
#' @param X vector de respuestas de la variable dicotómica a analizar (tamaño n)
#' @param Etiquetas vector de cadena de caracteres con las posibles respuestas.
#' @param Título cadena de caracteres indicando el nombre que se le quiere dar a la tabla
#' @export

freq.dic= function(datos, X, etiquetas, titulo)
 {
  datos1 <- mutate(datos, v1 = factor(X, labels = etiquetas))
  ##transforma los datos numéricos a factor

  Frec.abs=table(datos1$v1)
  Frec.rel=round(prop.table(Frec.abs),2)

  tabla=cbind(Frec.abs,Frec.rel )
  tabla=rbind(tabla, Total = colSums(tabla))
  kable( tabla, caption = título)
}
```

Figura 5.2: Script documentado

- Los archivos *man* y *NAMESPACE* se generan a través de la documentación en los scripts gracias a devtools y roxygen.

El segundo paso es construir la librería, en la pestaña Built, se ajusta la configuración y se carga.

5.3. Alojar el paquete en Github

Aunque no es específico de R, Github es probablemente el repositorio más popular para proyectos de código abierto. Su popularidad proviene del espacio ilimitado, la integración con git, un software de control de versiones y su facilidad para compartir y colaborar con otros.
Aunque hay que tener en cuenta que no hay un proceso de revisión asociado.

Lo primero es crear un nuevo repositorio en una cuenta de Github, una vez creado el repositorio se "clona" en nuestro equipo y se trasladan los archivos que componen el paquete.

Por último, usando la terminal del equipo, la aplicación previamente descargada *Git* y una serie de comandos, se añaden todos los archivos al control de versiones y se sincroniza la versión de la nube.

5.4. Resultados

Para instalar el paquete desde Github se usa la orden *install_github("celiaromerogustos/Cuestionarios")* de la librería *devtools*.

Una de las mayores ventajas de crear un paquete es la ayuda de R, como se muestra en la siguiente imagen, usando el comando *?* seguido del nombre de la función, R nos facilitará la descripción, uso y argumentos.

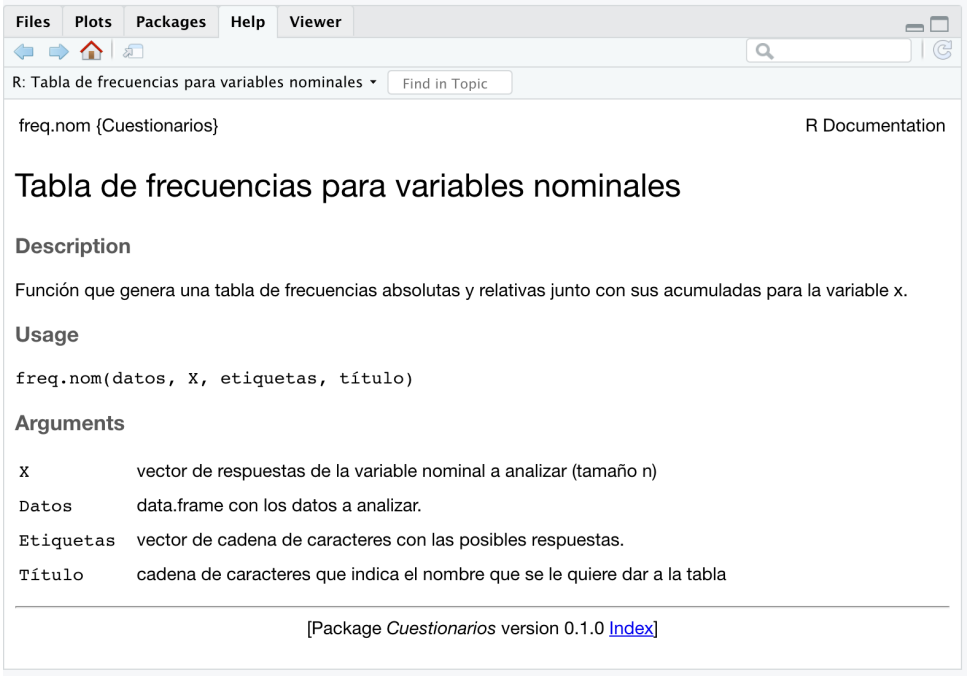


Figura 5.3: Ayuda en R

Bibliografía

Carmona Bolaños, M.J. (2007). *Estadística descriptiva de una variable*. Disponible en https://www.ugr.es/~rruizb/cognosfera/index.htm.

Chok, N.S. (2010). *Pearson's versus Spearman's and Kendall's Correlation Coefficients for Continuous data. Master's thesis, University of Pittsburgh.* Disponible en http://d-scholarship.pitt.edu/8056/1/Chokns_etd2010.pdf.

Devlin, M. (2016). *Using likert on summary results*. Disponible en https://rpubs.com/m_dev/likert_summary.

Hernández, F. & Usuga, O. (2019a). Datos medidas cuerpo.

Hernández, F. & Usuga, O. (2019b). *Manual de* R. Disponible en https://fhernanb.github.io/Manual-de-R/.

Janssenswillen, G. (2019). *Univariate and bivariate descriptive analysis: A not so short introduction to dplyr*. Disponible en https://beta.rstudioconnect.com/content/3350/dplyr_tutorial.html.

Kendall, M. (1938). *A new measure of rank correlation.* Biometrika, *Vol. 30, No. 1/2, pp. 81-93*. https://doi.org/10.1093/biomet/30.1-2.81.

López-Roldán, P. & Fachelli, S. (2015). *Metodología de la investigación social cuantitativa*. Disponible en https://ddd.uab.cat/record/129382; Universitat Autònoma de Barcelona,

Luque-Calvo, P. (2019). *Cómo crear tablas de información en* R *Markdown*. Disponible en http://destio.us.es/calvo.

Luque-Calvo, P. (2017). *Escribir un trabajo fin de estudios con* R *markdown*. Disponible en http://destio.us.es/calvo.

Lüdecke, D. (2020). *SjPlot: Data visualization for statistics in social science. R package version 2.8.4*. https://CRAN.R-project.org/package=sjPlot.

Morales, P. & Rodríguez, L. (2016). *Aplicación de los coeficientes de correlación de Kendall y Spearman*. Disponible en http://www.postgradovipi.50webs.com/archivos/agrollania/2016/agro8.pdf.

Murdoch, D. (2020). *Tables: Formula-driven table generation. R package version 0.9.3*. https://CRAN.R-project.org/package=tables.

Puigde, J. *Ggplot2 .The easiest path to graphics*. Disponible en https://rpubs.com/jpuigde/Ggplot2.

Servicio Gallego de Salud. (2014). *Epidat 4: Ayuda de análisis descriptivo*. Disponible en https://www.sergas.es/Saude-publica/Documents/1891/Ayuda_Epidat_4_Analisis_descriptivo_Octubre2014.pdf.

Sheskin, D.J. (2003). *Handbook of parametric and nonparametric statistical procedures. Chapman and* H*all/* CRC.

Signorell, A. & mult.al. (2020). *DescTools: Tools for descriptive statistics. R package version 0.99.36*. https://CRAN.R-project.org/package=DescTools.

STHDA. (2019). *Quick start guide-* R *software and data visualization*. Disponible en http://www.sthda.com/english/wiki/ggplot2-essentials.

Wei, T. & Simko, V. (2017). *R package "corrplot": Visualization of a correlation matrix. R package version 0.84*. https://github.com/taiyun/corrplot.

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. https://ggplot2.tidyverse.org; Springer-Verlag New York.

Wickham, H. & Bryan, J. (2019). *Readxl: Read excel files. R package version 1.3.1*. https://CRAN.R-project.org/package=readxl.

Wickham, H., François, R., Henry, L. & Müller, K. (2020). *Dplyr: A grammar of data manipulation. R package version 1.0.0*. https://CRAN.R-project.org/package=dplyr.

Xie, Y. (2020). *Knitr: A general-purpose package for dynamic report generation in r. R package version 1.28*. https://yihui.org/knitr/.