



Universidad de Oviedo

FACULTAD DE CIENCIAS

INFERENCIA ESTADÍSTICA

## ANEXO

EQUIPO  $\varepsilon$

CORTE FERNÁNDEZ, CÉSAR

FERNÁNDEZ VEGA, NURIA

GARCÍA PESCADOR, ISABEL

LLORIÁN GONZÁLEZ, PABLO

RIVERA MENÉNDEZ, CARMEN ZIYI

RODRÍGUEZ CRISTÓBAL, MANUEL

RUBIO ROBLES, PABLO

SANTOS LÓPEZ, CANDELA DE LOS

30 de abril de 2025

# 1. Código de R

## 1.1. Filtrado de datos

```
1 # Librerías que se usan ----
2 library(tidyverse)
3 library(readr)
4 library(openxlsx)
5
6 # Cargamos el conjunto de datos
7 datos_completos <- read.csv("FU_A2024MM07V.csv", sep=";")
8
9 # DATOS PARA EL ESTUDIO ----
10 # Seleccionamos las variables de interés para el estudio
11 v_int <- c("PAISDEST", "NPERNOC", "VIAJA_SOLO", "VIAJA_PAREJA", "VIAJA_HIJOS",
12     "VIAJA_OTROSFAMILIARES", "VIAJA_AMIGOS", "VIAJA_COMPTRABAJO",
13     "GASTOFI_ALOJA", "RESERV_ALOJA", "GASTOFI_TOTAL", "INGR_HOG", "SEXO",
14     "MIEMV")
15
16 # Creamos un dataset que contega solo los datos de las variables de
17 # interes
18 d_int <- datos_completos[, v_int]
19
20 # VARIABLES ----
21 ## VARIABLE: Con quien viaja ----
22 # Definimos una variable para saber con quien viaja
23 # Definir una función que asigne el valor correcto según las variables
24 asignar_categoria <- function(row) {
25     if (!is.na(row["VIAJA_SOLO"])) && row["VIAJA_SOLO"] == 1) {
26         return("Solo")
27     } else if (!is.na(row["VIAJA_PAREJA"])) && row["VIAJA_PAREJA"] == 1) {
28         return("Enpareja")
29     } else if (!is.na(row["VIAJA_HIJOS"])) && row["VIAJA_HIJOS"] == 1) {
30         return("Conhijos")
31     } else if (!is.na(row["VIAJA_OTROSFAMILIARES"])) && row["VIAJA_OTROSFAMILIARES"] == 1) {
32         return("Conotrosfamiliares")
33     } else if (!is.na(row["VIAJA_AMIGOS"])) && row["VIAJA_AMIGOS"] == 1) {
34         return("Conamigos")
35     } else if (!is.na(row["VIAJA_COMPTRABAJO"])) && row["VIAJA_COMPTRABAJO"] == 1) {
```

```

32         return("Con_&compañeros_de_trabajo")
33     } else {
34         return("Desplazamiento_al_Centro_de_Estudios/trabajo")
35     }
36 }

37
38 # Aplicar la función a cada fila del dataframe
39 conquienviaja <- apply(d_int[, c("VIAJA_SOLO", "VIAJA_PAREJA", "VIAJA_HIJOS", "VIAJA_OTROSFAMILIARES", "VIAJA_AMIGOS", "VIAJA_COMPTRABAJO")], 1, asignar_categoria)
40
41 ## VARIABLE: Noches pernoctadas ----
42 npernoc <- d_int$NPERNOC
43
44 ## VARIABLE: Gasto total por día ----
45 # Adaptamos el formato de la variable gasto total
46 d_int$GASTOFI_TOTAL <- as.numeric(gsub(",",".", d_int$GASTOFI_TOTAL))
47
48 g_tot <- d_int$GASTOFI_TOTAL
49 g_tot_dia <- d_int$GASTOFI_TOTAL/d_int$NPERNOC
50
51 ## VARIABLE: Gasto alojamiento por día ----
52 # Adaptamos el formato de la variable gasto alojamiento
53 d_int$GASTOFI_ALOJA <- as.numeric(gsub(",",".", d_int$GASTOFI_ALOJA))
54
55 g_aloj <- d_int$GASTOFI_ALOJA
56 g_aloj_dia <- d_int$GASTOFI_ALOJA/d_int$NPERNOC
57
58 ## VARIABLE: Gasto alojamiento por día y persona ----
59 g_aloj_dia_persona <- g_aloj_dia/d_int$MIEMV
60
61 ## VARIABLE: Destino ----
62 asignar_destino <- function(pais) {
63     if (pais == 108) {
64         return("España")
65     } else if (pais == 1) {
66         return("Resto_de_Europa")
67     } else if (pais == 2) {
68         return("África")
69     } else if (pais == 3) {
70         return("América")

```

```

71     } else {
72         return("Resto del mundo")
73     }
74 }

75
76 destino <- sapply(d_int$PAISDEST, asignar_destino)
77
78 ## VARIABLE: Reserva alojamiento ----
79 asignar_modo_reserva <- function(modo) {
80     if (is.na(modo)) {
81         return("No procede") # Manejo de NA
82     } else if (modo %in% c(1, 2)) {
83         return("Con el hotel")
84     } else if (modo %in% c(3, 4)) {
85         return("Con una agencia de viajes")
86     } else if (modo == 5) {
87         return("En página web especializada")
88     } else if (modo %in% c(6, 7)) {
89         return("Directamente con el propietario")
90     } else if (modo %in% c(8, 9)) {
91         return("En una agencia inmobiliaria")
92     } else if (modo == 10) {
93         return("No sabe")
94     } else {
95         return("No procede")
96     }
97 }

98
99 # Aplicar la función a cada valor de RESERV_ALOJA
100 reserva <- sapply(d_int$RESERV_ALOJA, asignar_modo_reserva)
101
102 ## VARIABLE: Ingresos del hogar ----
103 asignar_ingresos <- function(ing) {
104     if (is.na(ing)) {
105         return("No procede") # Manejo de NA
106     } else if (ing == 1) {
107         return("Hasta 999 euros")
108     } else if (ing == 2) {
109         return("De 1000 a 1499 euros")
110     } else if (ing == 3) {
111         return("De 1500 a 2499 euros")

```

```

112     } else if (ing == 4) {
113         return("De 2500 a 3499 euros")
114     } else if (ing == 5) {
115         return("De 3500 a 4999 euros")
116     } else if (ing == 6) {
117         return("5000 euros o más")
118     } else {
119         return("No contesta")
120     }
121 }
122
123 ingr_hog <- sapply(d_int$INGR_HOG, asignar_ingeros, USE.NAMES = FALSE)
124
125 ## VARIABLE: Sexo ----
126 asignar_sexo <- function(sexo) {
127     if (sexo == 1) {
128         return("Hombre")
129     } else {
130         return("Mujer")
131     }
132 }
133
134 sexo <- sapply(d_int$SEXO, asignar_sexo, USE.NAMES = FALSE)
135
136 # DATAFRAME ----
137 datos <- data.frame(Destino = destino, Acompañantes = conquienviaja,
138                       NPERNOC = npernoc, GastoAloj = g_aloj, GastoAlojDía = g_aloj_dia,
139                       GastoAlojDíaPersona = g_aloj_dia_persona, GastoTotal = g_tot,
140                       GastoTotalDía = g_tot_dia, ModoReserva = reserva, IngresosHogar =
141                       ingr_hog, Sexo = sexo)
142
143 # Guardamos el conjunto de datos
144 write.xlsx(datos, file = "datosine.xlsx", rowNames=FALSE)

```

## 1.2. Contraste acompañante vs gasto total día

```
1 library(ggplot2)
2 library(tidyverse)
3 tema <- theme_minimal() +
4   theme(
5     text = element_text(family = "serif", size = 14), # Fuente y tamaño de
6       texto
7     plot.background = element_rect(fill = "white", color = NA), # Fondo
8       blanco
9     panel.background = element_rect(fill = "white"), # Panel blanco
10    panel.grid = element_line(color = "gray90"), # Líneas de cuadrícula
11      suaves
12    axis.title = element_text(face = "bold"), # Texto en negrita para tí-
13      tulos de ejes
14    plot.title = element_text(face = "bold", hjust = 0.5) # Centrar título
15      en negrita
16  )
17
18 # Cargar los datos
19 datos<- read_csv("datos_finales.csv")
20
21 # Convertir la columna Acompañantes a factor
22 datos$Acompañantes <- as.factor(datos$Acompañantes)
23 levels(datos$Acompañantes)
24
25
26 # Calcular estadísticas descriptivas por grupo de acompañantes
27 # Número de observaciones (n) por grupo:
28 tabla_n <- table(datos$Acompañantes)
```

	Con amigos	Con compañeros de trabajo	Con hijos	Con otros familiares
623		72	414	562
Estudio/Trabajo		En pareja	Solo	
189		3249	1071	

```
1 media_por_grupo <- tapply(datos$GastoTotalDía, datos$Acompañantes, mean)
2 mediana_por_grupo <- tapply(datos$GastoTotalDía, datos$Acompañantes, median
3 )
4 sd_por_grupo <- tapply(datos$GastoTotalDía, datos$Acompañantes, sd)
```

```

4
5 estadisticas_por_grupo <- data.frame(
6   Acompañantes = names(tabla_n),
7   N = as.vector(tabla_n),
8   Media = as.vector(media_por_grupo),
9   Mediana = as.vector(mediana_por_grupo),
10  Desv_Tip = as.vector(sd_por_grupo)
11 )
12 estadisticas_por_grupo

```

	Acompañantes	N	Media	Mediana	Desv_Tip
	Con amigos	623	124.90181	109.65000	69.12204
Con compañeros de trabajo		72	222.21857	180.88750	144.36092
	Con hijos	414	85.95855	62.79937	66.56151
Con otros familiares		562	97.31153	77.85098	77.11130
	Estudio/Trabajo	189	91.55913	58.92333	80.01056
	En pareja	3249	89.61112	69.87000	63.71357
	Solo	1071	89.94068	57.15000	89.36226

```

1 # Diagrama de cajas del gasto diario por grupo de acompañantes
2 datos$Acompañantes <- factor(datos$Acompañantes,
3                               levels = levels(datos$Acompañantes),
4                               labels = c("Con_amigos",
5                                         "Con_compañeros_de_trabajo",
6                                         "Con_hijos",
7                                         "Con_otros_familiares",
8                                         "Estudio/Trabajo",
9                                         "En_pareja",
10                                        "Solo"))
11 par(mar = c(8, 4, 4, 2)) # margen inferior aumentado
12
13 boxplot(GastoTotalDía ~ Acompañantes, data = datos,
14           main = "Gasto_Total_por_Día_según_Acompañantes",
15           xlab = "", ylab = "Gasto_total_por_día_(euros)",
16           las = 2, cex.axis = 0.8,
17           col = c("lightblue", "lightgreen", "lightpink", "khaki",
18                  "lightgray", "plum", "salmon"))
19
20
21
22 # Prueba de normalidad Shapiro-Wilk para cada grupo de acompañantes

```

```

23 shapiro_por_grupo <- tapply(datos$GastoTotalDía, datos$Acompañantes,
24   function(x) shapiro.test(x)$p.value)
shapiro_por_grupo # mostrar los p-valores por grupo

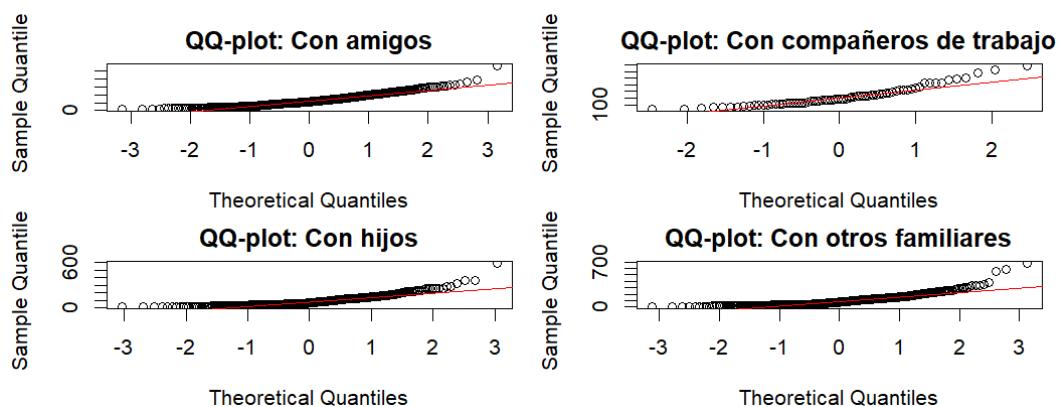
```

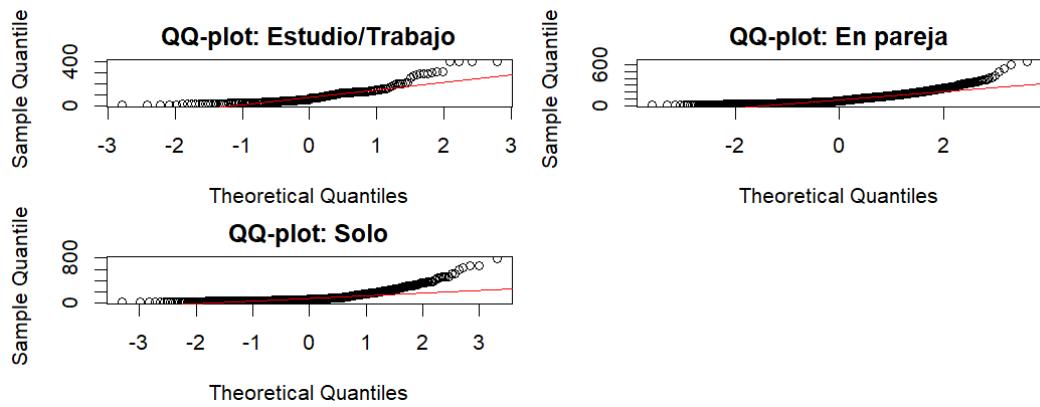
Con amigos 1.492781e-16	Con compañeros de trabajo 3.935399e-05	Con hijos 5.919280e-21	Con otros familiares 9.739784e-25
Estudio/Trabajo 1.152838e-14	En pareja 4.664967e-48	Solo 5.696209e-40	

```

1 # Gráficos Q-Q por grupo para evaluar visualmente la normalidad
2 par(mfrow = c(2,2), mar = c(4,4,2,1))
3 for(cat in levels(datos$Acompañantes)[1:4]) {
4   x <- datos$GastoTotalDía[datos$Acompañantes == cat]
5   qqnorm(x, main = paste("QQ-plot:", cat))
6   qqline(x, col = "red")
7 }
8 par(mfrow = c(1,1)) # Reset
9
10 # Luego los siguientes
11 par(mfrow = c(2,2), mar = c(4,4,2,1))
12 for(cat in levels(datos$Acompañantes)[5:8]) {
13   x <- datos$GastoTotalDía[datos$Acompañantes == cat]
14   qqnorm(x, main = paste("QQ-plot:", cat))
15   qqline(x, col = "red")
16 }
17 par(mfrow = c(1,1))

```





```

1 # Levene
2 library(car)
3 leveneTest(GastoTotalDía ~ Acompañantes, data = datos, center = "median")

```

```

Levene's Test for Homogeneity of Variance (center = "median")
  Df F value    Pr(>F)
group      6 15.523 < 2.2e-16 ***
6173
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

1 # Contraste no paramétrico de Kruskal-Wallis
2 kruskal_result <- kruskal.test(GastoTotalDía ~ Acompañantes, data = datos)
3 kruskal_result

```

```

Kruskal-Wallis rank sum test

data: GastoTotalDía by Acompañantes
Kruskal-Wallis chi-squared = 324.01, df = 6, p-value < 2.2e-16

```

### 1.3. Contrastes ingresos hogar y destino; coste total y destino

```
1 # RESUTADOS Y CONCLUSIONES SOBRA LA INFLUENCIA DE LOS INGRESOS DEL HOGAR EN
2 # EL
3 # EL COSTE TOTAL Y EL DESTINO
4 #
5 # Nos podemos preguntar si existe una relacion entre los ingresos del hogar
6 # de las personas que realizan el viaje y sus gastos totales por dia en
7 # el viaje.
8 #
9 # VARIABLES DEL ESTUDIO----
10 #
11 # -----
12 #
13 ## VARIABLE GASTO TOTAL
14 gastototaldia<-datos$GastoTotalDía
15 # Se trata de una variable continua.
16 # Notar que se parece a una exponencial de parametro 1/ media muestral
17 hist(gastototaldia,prob=TRUE)
18 plot(function(t){dexp(t,1/mean(gastototaldia))},0,800,add=TRUE)
19 # Resumen de los datos de esta varaiable
20 summary(gastototaldia)
21 # El maximo gasto total al dia es de 789.880      y el minimo 4.372
22 #
23 # -----
24 #
25 ##
26 ## VARIABLE INGRESOS HOGAR
27 ingresoshogar<-datos$IngresosHogar;ingresoshogar
28 # Se trata de una variable discreta que toma 7 categorias de rangos de
29 # dinero
30 table(ingresoshogar)[c(6,2,3,4,5,1,7)]
31 length(table(ingresoshogar)) # toam 7 valores (hay 7 categorias)
32 # La ultima categoria estan los que no contestan que no intervienen en
33 # nuestro estudio luego los eliminamos de los datos
34 datos <- datos %>%
35   filter(IngresosHogar != "No\u00f1contesta")
36 # Renombrar las variables tras el filtrado
37 ingresoshogar<-datos$IngresosHogar;ingresoshogar
```

```

28 gastototaldia<-datos$GastoTotalDía
29 length(gastototaldia)==length(ingresoshogar) # bien
30
31 t<-table(ingresoshogar)[c(6,2,3,4,5,1)];t # 6 categorias por rango de dinero
32 # Visualmente se observa que la mayor parte de la gente que viaja se
   encuentra en las categorias 3,4
33
34 #
-----  

35 # GRAFICO VISUAL
36 #
-----  

37 boxplot(gastototaldia~ingresoshogar)
38 summarise(group_by(datos,IngresosHogar),
            Media=mean(GastoTotalDía),
            Mediana=median(GastoTotalDía),
            DesvTip=sd(GastoTotalDía),
            Mínimo=min(GastoTotalDía),
            Máximo=max(GastoTotalDía))
44 # A priori podemos observar que en todos los grupos hay valores atípicos ,
  lo cual es común en datos de gasto. Ademas, las cajas son más altas ,en
  los grupos de mayores ingresos , lo que podría indicar más variabilidad
  en el gasto diario entre personas con mayores ingresos.
45 #
-----  

46 # ANALISIS DE HOMOGENEIDAD
47 #
-----  

48 # Vamos ver si hay homogeneidad de los gastos totales del viaje entre las
  distintas categorias de ingresos del hogar que hemos discreto.
49 # Para ello primero resultaria interesante realizar un contraste de
  comparación de medias ANOVA para ver si hay diferencias significativas
  en el gasto promedio entre grupos de ingreso pero para ello debemos
  primero comprobar si las variables cumplen las condiciones de normalidad
  y homocedasticidad:
50
51 # Agrupamos en los datos de gasto total por los grupos de ingresos , dando

```

```

lugar a las 6 variable saleatorias continuas Xi: gasto/dia en el rango
de ingresos i

52 gastos_por_ingresos <- split(gastototaldia,ingresoshogar)
53 gastos_por_ingresos; #(lista, donde cada elemento es un vector de valores
    numéricos del
                                #gasto por día para cada grupo de ingresos)

55

56 # Visualmente ya apreciamos que no van seguir una distribucion normal
57 hist(gastos_por_ingresos[[1]], prob=TRUE)
58 plot(function(t){dexp(t,1/mean(gastos_por_ingresos[[1]]))},0,600,add=TRUE)
59 hist(gastos_por_ingresos[[2]],prob=TRUE)
60 plot(function(t){dexp(t,1/mean(gastos_por_ingresos[[2]]))},0,600,add=TRUE)
61 hist(gastos_por_ingresos[[3]],prob=TRUE)
62 plot(function(t){dexp(t,1/mean(gastos_por_ingresos[[3]]))},0,600,add=TRUE)
63 hist(gastos_por_ingresos[[4]], prob=TRUE)
64 plot(function(t){dexp(t,1/mean(gastos_por_ingresos[[4]]))},0,600,add=TRUE)
65 hist(gastos_por_ingresos[[5]],prob=TRUE)
66 plot(function(t){dexp(t,1/mean(gastos_por_ingresos[[5]]))},0,600,add=TRUE)
67 hist(gastos_por_ingresos[[6]],prob=TRUE)
68 plot(function(t){dexp(t,1/mean(gastos_por_ingresos[[6]]))},0,600,add=TRUE)

69

70

71 # si aplicamos el test grafico Q Q, la mayoria de los puntos aparecen lejos
    de la linea (no distrib normal)
72 qqnorm(gastos_por_ingresos[[1]],pch=16)
73 qqline(gastos_por_ingresos[[1]],pch=16)
74 # En efecto al realizar el test shapiro para comprobar la normalidad e cada
    grupo nos encontramos con p-valores muy bajos que rechazan la hip nula
    de deguir distribuciones normales.
75 by(gastototaldia,ingresoshogar,function(x)shapiro.test(x)$p.value)

76

77 # Aunque no hay ANOVA ya pue no hay normalidad veamos si hay
    homocedasticidad
78 library(car)
79 leveneTest(gastototaldia~ingresoshogar) # pvalor pequeño tambien se rechaza
80 # Por tanto recurrimos al test de Kruskal-Wallis
81 kruskal.test(gastototaldia~ingresoshogar)
82 # Con el test obtenemos un pvalor muy bajo menos que 2.2e-16 luego
    rechazamos la hipotesis nula(rechazamos la independencia), es decir, si
    hay difenrencia respecto a los gastos totales entre los distintos grupos
    de ingresos

```

```

83
84 # -----
85 # ANÁLISIS DE CORRELACION
86 #
87 # En el estudio anterior hemos visto que existe relación, vamos ahora a
88 # confirmar y cuantificar
89 # que esa relación es directa (es decir, a mayor ingreso, mayor gasto).
90
91 # Primero nos aseguramos de q la variable "Ingresos del Hogar" esté en el
92 # orden correcto
93 datos$ingresos_ordinal <- factor(datos$IngresosHogar, levels = c("Hasta 999
94 euros", "De 1000 a 1499 euros", "De 1500 a 2499 euros", "De 2500 a 3499
95 euros", "De 3500 a 4999 euros", "5000 euros o más"), ordered = TRUE)
96
97 # Para poder estudiar la correlación, convertimos este factor ordinal en
98 # valores numéricos (1, 2, 3, ...)
99 datos$ingresos_ordinal_num <- as.numeric(datos$ingresos_ordinal)
100 # Así podemos aplicar la correlación de Spearman entre el ingreso ordinal y
101 # el gasto total diario
102
103 # -----
104 ##      VARIABLE INGRESOS HOGAR

```

```

105 ingresoshogar<-datos$IngresosHoga # de nuevo
106 #
107 ##      VARIABLE DESTINO
108 destino<-datos$Destino;
109
110 # Se trata de una variable discreta que toma 5 categorias.
111 table(destino)
112 length(table (destino)) # toma 5 valores (hay 5 categorias)
113 barplot(table(destino))
114 # A priori podemos ver que la mayor parte de las personas que viajan del
# estudio lo hacen dentro de españa
115 #

116 # GRAFICO VISUAL
117 #

118 library(ggplot2)
119 ggplot(datos, aes(x=Destino, fill=IngresosHogar)) +
120   geom_bar(position="fill")
121
122 #

123 # ANALISIS DE HOMOGENEIDAD
124 #

125 # Vamos a ver si hay homogenidad entre los ingresos del hogar y el destino
# escogido.
126 # Dado que menos ingresos llevan a intentar gastar menos en el viaje,
# parece razonable que menores ingresos nos lleven a viajes mas cercanos
# de españa puesto que en principio estos pueden ser mas baratos
127
128 # Para ello dado que estamos considerando dos variables aleatorias
# discretas usamos el test ji cuadrado. Construimos primero la tabla de
# contingencia
129 t<-table(ingresoshogar,destino)[c(6,2,3,4,5,1),];t

```

```

130 t1<-table(destino,ingresoshogar)
131 y<-chisq.test(t,simulate.p.value = TRUE) # la Eij >=5
132 # Se obtiene un pvalor bajo luego en efecto hay relacion,falta encontrarla.
133 round(prop.table(t,margin=1)*100)
134 round(prop.table(t1,margin=1)*100)
135 # RESIDUOS
136 y$residuals
137 #
-----#
139 # ANÁLISIS DE CORRELACION
140 #
-----#
141 # Podemos tambien ver si la relacion entre la lejanía del destino elegido y
142 # los ingresos del hogar es monótona creciente
143 # Ya teniamos la variable Ingresos del Hogar como un factor ordinal de
144 # valores numericos. Vamos a ordenar los distintos destinos de mas cercano
145 # a lejano españa con el siguiente criterio, y asociarles valores del 1
146 # al 5
147 datos$destino_ordinal <- factor(datos$Destino, levels = c("España", "Resto de Europa", "África", "América", "Resto del Mundo"), ordered = TRUE)
148 datos$destinos_ordinal_num <- as.numeric(datos$destino_ordinal)
149 # Asi podemos aplicar la correlación de Spearman entre el ingreso ordinal y
150 # el la lejania del destino (ordinal tb)
151 cor.test(datos$ingresos_ordinal_num,datos$destinos_ordinal_num,method = "spearman")

```

## 1.4. Influencia del destino en el gasto total del viaje por día

```
1 ##### Influencia del destino en el gasto total del viaje por día: #####
2
3 library(tidyverse)
4 library(ggplot2)
5 library(knitr)
6
7 datos <- read.csv("datos_filtrados.csv",
8                     fileEncoding = "UTF-8",
9                     sep = ",")
10
11 # Ver estructura de los datos
12 glimpse(datos)
13
14 tema <- theme_minimal() +
15   theme(
16     text = element_text(family = "serif", size = 14), # Fuente y tamaño de
17     # texto
18     plot.background = element_rect(fill = "white", color = NA), # Fondo
19     # blanco
20     panel.background = element_rect(fill = "white"), # Panel blanco
21     panel.grid = element_line(color = "gray90"), # Líneas de cuadrícula
22     # suaves
23     axis.title = element_text(face = "bold"), # Texto en negrita para títulos
24     # de ejes
25     plot.title = element_text(face = "bold", hjust = 0.5), # Centrar título
26     # en negrita
27     axis.text.x = element_text(angle = 45, hjust = 1) # Rotar etiquetas
28     # del eje X
29   )
30
31
32 ## ESTUDIO DESCRIPTIVO
33
34
35 # Convertir la columna Destino a factor:
36 # sirve principalmente para indicar que esa columna representa variables
37 # cualitativas
38 datos$Destino <- as.factor(datos$Destino)
39 levels(datos$Destino)
40
41
42 # Calcular estadísticas descriptivas por destino
43 # Número de observaciones (n) por grupo:
```

```

34 tabla_n <- table(datos$Destino)
35
36 media_por_destino <- tapply(datos$GastoTotalDía, datos$Destino, mean)
37 mediana_por_destino <- tapply(datos$GastoTotalDía, datos$Destino, median)
38 sd_por_destino <- tapply(datos$GastoTotalDía, datos$Destino, sd)
39
40 estadisticas_por_destino <- data.frame(
41   Destino = names(tabla_n),
42   N = as.vector(tabla_n),
43   Media = as.vector(media_por_destino),
44   Mediana = as.vector(mediana_por_destino),
45   Desv_Tip = as.vector(sd_por_destino)
46 )
47 estadisticas_por_destino
48
49 # Diagrama de cajas del gasto diario por destino
50 ggplot(datos, aes(x = Destino, y = GastoTotalDía, fill = Destino)) +
51   geom_boxplot() +
52   labs(title = "Distribución del Gasto por Día según Destino",
53       x = "Destino",
54       y = "Gasto Total por Día(") +
55   tema
56
57 ## CONTRASTES
58
59 # Antes de elegir y aplicar un contraste estadístico, es importante
60 # comprobar si se
61 # cumplen las hipótesis necesarias para aplicar los test más usuales (ANOVA
62 #).
63 # Es decir, comprobaremos la normalidad y la homocedasticidad.
64
65 # Prueba de normalidad K-S para cada destino
66 # H_0: la variable "GastoTotalDía" sigue una distribución normal
67 # H_1. la variable "GastoTotalDía" NO sigue una distribución normal
68 # Función para aplicar el test de Kolmogorov-Smirnov con corrección de
69 # Lilliefors
70
71 library(nortest)
72 # Aplicar el test por destino
73 ks_por_destino <- datos %>%
74   group_by(Destino) %>%

```

```

72 summarise(
73   ks_p_value = lillie.test(GastoTotalDía)$p.value,
74   ks_statistic = lillie.test(GastoTotalDía)$statistic
75 )
76 alfa<-0.05
77 ks_por_destino <- ks_por_destino %>%
78   mutate(rechazar_H0 = ks_p_value < alfa)
79 ks_por_destino
80 # Según el test de Kolmogorov-Smirnov con corrección de Lilliefors a nivel
81 # de
82 # significacion alfa=0.05:
83 # las distribuciones del gasto total por día en África, España y Resto de
84 # Europa
85 # no son normales.
86 # En cambio, las distribuciones en América y Resto del Mundo podrían ser
87 # normales,
88 # ya que no se rechaza la hipótesis nula en estos casos.

89 # Diagramas Q-Q por destino
90 ggplot(datos, aes(sample = GastoTotalDía)) +
91   stat_qq() +
92   stat_qq_line() +
93   facet_wrap(~ Destino) +
94   tema
95 # Las gráficas Q-Q confirman lo observado en el test de K-S con corrección
96 # de Lillefors.

97 # A continuación, analizaremos si se cumple la homogeneidad de varianzas
98 # con el
99 # test de Levene
100 # H_0: Las varianzas de la variable GastoTotalDía son iguales en todos los
101 # destinos
102 # H_1: Al menos una de las varianzas de la variable GastoTotalDía es
103 # diferente en los
104 # destinos
105 library(car)
106 leveneTest(GastoTotalDía ~ Destino, data = datos, center = "median")
107 # Rechazo de H_0 El valor p-valor es extremadamente pequeño (< 2.2e-16),
108 # muy por debajo de alfa=0.05
109
110 # Por tanto, no se cumplen las hipótesis necesarias para realizar un ANOVA

```

```

106
107 # La alternativa adecuada es el test de Kruskal-Wallis, que es un test no
108 # paramétrico
109 # para comparar más de dos grupos independientes y no requiere que los
110 # datos sean
111 # normales ni que las varianzas sean iguales
112 # H_0: Las distribuciones del gasto son iguales entre todos los destinos
113 # H_1: Existe al menos un destino cuya distribución difiere de las otras
114 kruskal_result <- kruskal.test(GastoTotalDía ~ Destino, data = datos)
115 kruskal_result
116
117
118 # Esto nos da p-valor extremadamente pequeño, que implica que se rechaza
119 # con claridad H_0 (igualdad de medias entre los grupos)
120
121 # Test de Dunn con la R.C. de Bonferroni:
122 library(FSA)
123 dunn_test_result <- dunnTest(GastoTotalDía ~ Destino, data = datos, method
124   = "bonferroni")
125 print(dunn_test_result)

```

## 1.5. Análisis de la relación entre el género del viajero y el gasto total

```
1 datos<- read.csv("datos_filtrados.csv")
2
3
4 # Convertir la columna Acompañantes a factor
5 datos$Sexo <- as.factor(datos$Sexo)
6 levels(datos$Sexo)
7
8
9 # Calcular estadísticas descriptivas por género
10 # Número de observaciones (n) por grupo:
11 tabla_n <- table(datos$Sexo)
12
13 media_por_grupo <- tapply(datos$GastoTotal, datos$Sexo, mean)
14 mediana_por_grupo <- tapply(datos$GastoTotal, datos$Sexo, median)
15 sd_por_grupo <- tapply(datos$GastoTotal, datos$Sexo, sd)
16
17 estadisticas_por_grupo <- data.frame(
18   Sexo = names(tabla_n),
19   N = as.vector(tabla_n),
20   Media = as.vector(media_por_grupo),
21   Mediana = as.vector(mediana_por_grupo),
22   Desv_Tip = as.vector(sd_por_grupo)
23 )
24 estadisticas_por_grupo
```

```
1 #####CONTRASTE DE PROPORCIONES
2
3 # 1) Tabla de conteos
4 tab_sexo <- table(datos$Sexo) # cuántos "Hombre" y "Mujer"
5 n_hombres <- tab_sexo["Hombre"] # número de hombres
6 n_total <- sum(tab_sexo) # tamaño total de la muestra
7 p_muestral <- n_hombres / n_total # estimador máximo verosímil de p
8
9 # 2) Mostrar resultados
10 tab_sexo
11 p_muestral
12
13 # Explicación:
14 # - tab_sexo nos dice cuántos Hombre y Mujer hay.
15 # - p_muestral = n_hombres / n_total es el estimador de máxima
```

```

verosimilitud para p.

16
17 # 3) Plantear hipótesis
18 #   H0: p = 0.5 (la proporción de hombres es 0.5)
19 #   H1: p ≠ 0.5 (la proporción de hombres difiere de 0.5)
20 # Contraste de proporciones con prop.test
21 test <- prop.test(
22   x           = n_hombres,      # éxitos = número de hombres
23   n           = n_total,       # tamaño de la muestra
24   p           = 0.5,          # proporción bajo H0
25   alternative = "two.sided",  # bilateral
26   correct     = FALSE         # sin corrección de continuidad (Z clásico)
27 )
28
29 # Explicación:
30 # prop.test realiza un test de proporciones basado en la aproximación
31 #   normal (o   ).  

32 # Al poner correct = FALSE obtenemos el estadístico Z clásico sin corrección
33 #   de Yates.
34
35 # 4) Extraer estadístico Z y p-valor
36 z_value <- (p_muestral - 0.5) / sqrt(0.5 * 0.5 / n_total)
37 p_valor <- test$p.value
38
39 # 5) Mostrar estadístico y p-valor
40 z_value
41 p_valor
42
43
44 #REPRESENTACIONES
45
46 library(ggplot2)
47 #Diagrama violin
48 ggplot(datos, aes(x = Sexo, y = GastoTotal, fill = Sexo)) +
49   geom_violin(trim = FALSE, alpha = 0.6) +
50   geom_boxplot(width = 0.1, fill = "white") +
51   theme_minimal() +
52   labs(title = "Gasto en Vacaciones según el Género",
53        y = "Gasto",

```

```

54         x = "Género")
55 # Diagrama de cajas del gasto total del viajes por genero
56 datos$Sexo <- factor(datos$Sexo,
57                         levels = levels(datos$Sexo),
58                         labels = c("HOMBRE", "MUJER"))
59
60
61 boxplot(GastoTotal ~ Sexo, data = datos,
62           main = "Gasto_Total",
63           xlab = "", ylab = "Gasto_total(euros)",
64           ylim=c(0,2000),
65           col = c("lightblue", "lightgreen"))
66
67
68
69 # Prueba de normalidad Shapiro-Wilk para cada grupo de acompañantes
70 shapiro_por_grupo <- tapply(datos$GastoTotal, datos$Sexo, function(x)
71   shapiro.test(x)$p.value)
72 shapiro_por_grupo # mostrar los p-valores por grupo
73
74 # Gráficos Q-Q por grupo para evaluar visualmente la normalidad
75 par(mfrow = c(1, 2)) # Mostrar 2 gráficos en una sola ventana
76
77 for(cat in levels(datos$Sexo)) {
78   x <- datos$GastoTotal[datos$Sexo == cat]
79   qqnorm(x,
80           main = paste("QQ-plot:", cat),
81           xlab = "Cuantiles teóricos",
82           ylab = "Cuantiles muestrales")
83   qqline(x, col = "red")
84 }
85
86 # Levene (más robusto) - requiere paquete car
87 # install.packages("car") # ejecutar si "car" no está instalado
88 install.packages("car")
89 library(car)
90 leveneTest(GastoTotal ~ Sexo, data = datos, center = "median")
91
92 # Test de Wilcoxon (Mann Whitney) para comparar 'gasto_total' según 'genero'

```

```

93 # Asumimos dos niveles en 'genero' (p. ej. "M" y "F")
94 res.wilcox <- wilcox.test(
95   GastoTotal ~ Sexo,
96   data = datos,
97   alternative = "two.sided",    # bilateral
98   exact = FALSE,                # FALSE si tienes muestras grandes o empates
99   correct = TRUE                 # corrección de continuidad
100 )
101
102 # Mostrar resultados
103 print(res.wilcox)
104 # Contraste no paramétrico de Kruskal-Wallis
105 kruskal_result <- kruskal.test(GastoTotal ~ Sexo, data = datos)
106 kruskal_result
107
108 hist(datos$GastoTotal)

```

## 1.6. Análisis de la relación entre el modo reserva y el gasto alojamieto por día y persona

```
1 # Librerías que se pueden ser necesarias ----
2 library(tidyverse)
3 library(readxl)
4 library(readr)
5 library(ggplot2)
6 library(nortest)
7 library(car)
8
9 # Cargamos el conjunto de datos
10 datosine <- read_excel("datosine.xlsx")
11
12 # Influencia del modo de reserva del alojamiento en el gasto por día y
13 # persona del mismo
14
15 # Para el estudio que se va a realizar, filtramos los datos, ya que no
16 # interesa tener en cuenta los valores "No sabe", "No procede". También
17 # aprovechamos para eliminar las variables que no intervienen en
18 # este análisis.
19 d_filtrado = datos[!(datos$ModoReserva %in% c("No_sabe", "No_procede"))
20   , c("ModoReserva", "GastoAlojDíaPersona")]
21
22 # Variables que intervienen en este análisis ----
23 ## Forma en que se reservó el alojamiento ----
24 ModoReserva <- datos$ModoReserva
25
26 # Es una variable discreta que toma los valores
27   # "Con el hotel"
28   # "Con una agencia de viajes"
29   # "Directamente con el propietario"
30   # "En página web especializada"
31   # "En una agencia inmobiliaria"
32   # "No sabe"
33   # "No procede"
34
35 # Frecuencias absolutas que toma cada uno de estos valores
36 table(ModoReserva)
```

```
1 # Las reservas de alojamientos no se distribuyen de forma uniforme
2 # entre todos los modos de reserva.
```

ModoReserva				
Con el hotel	896	Con una agencia de viajes	486	Directamente con el propietario
En página web especializada	503	En una agencia inmobiliaria	50	No procede
No sabe	237			3752

```

2
3 ## Gasto del alojamiento por día y persona ----
4 GastoAlojDíaPersona <- datos$GastoAlojDíaPersona
5 # Es una variable continua, que toma valores reales positivos
6 summary(GastoAlojDíaPersona)

```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.00	17.01	30.34	39.94	52.67	341.86

```

1 # Análisis de la influencia ----
2
3 ## Resumen estadístico básico ----
4 Resumen <- summarise(group_by(d_filtrado, ModoReserva), Minimo=min(
  GastoAlojDíaPersona), Primer_Q=quantile(GastoAlojDíaPersona, 0.25),
  Media=mean(GastoAlojDíaPersona), Tercer_Q=quantile(GastoAlojDía
  Persona, 0.75), Maximo=max(GastoAlojDíaPersona), DesvTip=sd(
  GastoAlojDíaPersona), ns=length(GastoAlojDíaPersona));
5 Resumen

```

ModoReserva	<chr>	Minimo	Primer_Q	Media	Tercer_Q	Maximo	DesvTip	ns
Con el hotel	<dbl>	3.79	18.9	43.2	61.3	342.	33.6	896
Con una agencia de viajes	<dbl>	6.42	25.5	52.9	68.7	316.	41.2	486
Directamente con el propietario	<dbl>	0	13.4	26.4	36.8	153.	18.9	256
En página web especializada	<dbl>	0	13.4	29.4	43.2	139.	20.6	503
En una agencia inmobiliaria	<dbl>	2.93	13.4	30.0	44.6	116.	25.1	50

```

1 ## Intervalos de confianza ----
2 # Vamos a construir un intervalo de confianza para la media de los
  gastos diarios en alojamientos agrupandolos en función del modo de
  reserva
3 alfa <- 0.05; # Nivel de confianza
4 z <- qnorm(1-alfa/2);
5 Extremo_izq = Resumen$Media - z*Resumen$DesvTip/Resumen$ns;
6 Extremo_dcho = Resumen$Media + z*Resumen$DesvTip/Resumen$ns;
7 IC <- cbind(Extremo_izq, Extremo_dcho);

```

```

8   rownames(IC) <- c("Con el hotel", "Con una agencia de viajes", "
9   Directamente con el propietario", "En página web especializada", "En
  una agencia inmobiliaria")
IC
```

	Extremo_izq	Extremo_dcho
Con el hotel	43.15841	43.30538
Con una agencia de viajes	52.72000	53.05219
Directamente con el propietario	26.29153	26.58066
En página web especializada	29.34745	29.50793
En una agencia inmobiliaria	28.99920	30.96919

```

1 ## Representación gráfica ----
2 # Para hacernos una idea inicial recurriremos a un diagrama de cajas
3 ggplot(d_filtrado,aes(x=ModoReserva,y=GastoAlojDiaPersona,fill=
4   ModoReserva)) +
5   # en ej eje x se representan los modos de reserva
6   # en el eje y los gastos en alojamiento por día
7   # especificamos que se coloree la caja en función del modo de reserva
8   geom_boxplot() +
9   stat_boxplot(geom="errorbar", width = 0.2) +
10  # agrega una barra horizontal al final de los "bigotes"
11  stat_summary(fun=mean, geom="point", size=2, shape=22, fill="white") +
12  scale_x_discrete(labels = c("Hotel","Agencia","Propietario","Web",
13    "Inmobiliaria")) +
14  theme_minimal()
15
16 ## NORMALIDAD ----
17 # Test de normalidad de GastoAlojDíaPersona agrupado por ModoReserva
18 summarise(group_by(d_filtrado,ModoReserva), shapiro.test(GastoAlojDí
19 aPersona)$p.value)
```

ModoReserva	shapiro.test(GastoAlojDíaPersona)\$p.value`
<chr>	<dbl>
Con el hotel	4.41e-29
Con una agencia de viajes	7.18e-24
Directamente con el propietario	4.83e-16
En página web especializada	1.94e-20
En una agencia inmobiliaria	1.24e- 6

```

1 # Como se puede comprobar mirando los p-valores obtenidos, ninguna de
2 las variables sigue una distribución normal a nivel de significacion
3 0.05
4
5 ## HOMOCEDASTICIDAD ----
6 leveneTest(GastoAlojDíaPersona~ModoReserva, data=d_filtrado)
```

```

Levene's Test for Homogeneity of Variance (center = median)
  Df F value    Pr(>F)
group     4   25.34 < 2.2e-16 ***
2186
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Aviso:
In leveneTest.default(y = y, group = group, ...) : group coerced to factor.

```

```

1 # Como se puede comprobar mirando los p-valores obtenidos, no hay
2   homocedasticidad entre las variables a nivel de significacion 0.05
3
4 ## HOMOGENEIDAD ----
5 # Representamos individualmente el gasto en alojamiento por día y por
6   persona agrupando los datos por categorías en función del modo de
7   reserva
8 ggplot(d_filtrado, aes(x = GastoAlojDíaPersona)) +
9   geom_density(fill = "lightblue", alpha = 0.5) +
10  facet_wrap(~ModoReserva, ncol = 2) +
11  labs(x = "Gasto por día y persona", y = "Densidad") +
12  theme_minimal()
13 # Este gráfico esta incluido en el trabajo, nos permite hacernos una
14   idea inicial sobre la homogeneidad de las variables ya que nos
   permite comparar sus distribuciones empíricas
15
16 # Como no hay normalidad -> usamos test no paramétrico:
17 # Test de homogeneidad de GastoAlojDíaPersona agrupado por ModoReserva
18 kruskal.test(GastoAlojDíaPersona~ModoReserva, data=d_filtrado)

```

```

Kruskal-Wallis rank sum test

data: GastoAlojDíaPersona by ModoReserva
Kruskal-Wallis chi-squared = 205.64, df = 4, p-value < 2.2e-16

```

```

1 # El p-valor es bajo, podemos concluir que no siguen la misma
2   distribución
3
4 # Test por parejas
5 # Comprobamos que parejas son las que dan problemas en el test de
   homogeneidad
6 pairwise.wilcox.test(d_filtrado$GastoAlojDíaPersona, d_filtrado$ModoReserva,
7   paired = FALSE)

```

```
Pairwise comparisons using Wilcoxon rank sum test with continuity correction

data: d_filtrado$GastoAlojDíaPersona and d_filtrado$ModoReserva

          Con el hotel Con una agencia de viajes
Con una agencia de viajes 4.6e-06      -
Directamente con el propietario 1.4e-14    < 2e-16
En página web especializada 7.6e-15    < 2e-16
En una agencia inmobiliaria  0.0032    1.1e-06
                                Directamente con el propietario En página web especializada
Con una agencia de viajes  -                  -
Directamente con el propietario -                  -
En página web especializada 0.3420      -
En una agencia inmobiliaria  1.0000      1.0000

P value adjustment method: holm
```

## 2. Intervalos de confianza

El objetivo es construir intervalos de confianza para la media del gasto en alojamiento por día y por persona en función del modo de reserva, es decir, un intervalo de confianza para la media de cada una de las siguientes variables aleatorias:

$X_1 \equiv$  gasto en alojamiento reservado con el hotel por día y por persona (en €)

$X_2 \equiv$  gasto en alojamiento reservado con una agencia de viajes por día y por persona (en €)

$X_3 \equiv$  gasto en alojamiento reservado directamente con el propietario por día y por persona (en €)

$X_4 \equiv$  gasto en alojamiento reservado en página web especializada por día y por persona (en €)

$X_5 \equiv$  gasto en alojamiento reservado en una agencia inmobiliaria por día y por persona (en €)

Se puede utilizar la media muestral como estadístico para conseguir la función pivote que permitirá construir los intervalos de confianza. Como no se conoce la distribución de ninguna de las variables aleatorias, no podemos concluir qué distribución siguen sus medias. Sin embargo, para tamaños de muestra suficientemente grandes sabemos que convergen a distribuciones completamente especificadas, basta aplicar el Teorema Central del Límite.

A continuación se relata de manera genérica como calcular el intervalo de confianza para la media de una variable aleatoria  $X$  con distribución desconocida. Por simplicidad se denota  $\mu = E(X)$  y  $\sigma^2 = Var(X)$ .

Sea  $(X_1, \dots, X_n)$  una muestra aleatoria simple procedente de la variable aleatoria  $X$ . Si el tamaño de muestra  $n$  es suficientemente grande ( $n \geq 30$ ), podemos aplicar el Teorema Central del Límite, de forma que

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

En este caso la varianza  $\sigma^2$  es desconocida, se recurrirá a un estimador, la varianza muestral. Es cierto que esto introduce más incertidumbre. Se trabajará con la función pivote

$$\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}},$$

donde  $s$  es la desviación típica muestral.

Ahora que ya se dispone de una función pivote cuya distribución asintótica es conocida, se puede calcular un intervalo de confianza para  $\mu$ . Fijado un nivel de confianza de  $1 - \alpha$ , se eligen dos probabilidades  $\alpha_1$  y  $\alpha_2$  tales que  $\alpha_1 + \alpha_2 = \alpha$ . En este caso se tomará  $\alpha_1 = \alpha_2 = \frac{\alpha}{2}$ .

Se buscan valores  $a$  y  $b$  tales que

$$P\left(\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \leq a\right) \leq \alpha_1,$$

$$P\left(\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} < b\right) \geq 1 - \alpha_2.$$

Aprovechando que se conoce la distribución asintótica de la función pivote, se puede calcular los valores de  $a$  y  $b$  de forma que

$$P(\mathcal{N}(0, 1) \leq a) = \alpha_1 = \frac{\alpha}{2},$$

$$P(\mathcal{N}(0, 1) < b) = 1 - \alpha_2 = 1 - \frac{\alpha}{2}.$$

Bastaría con tomar  $a = -z_{1-\frac{\alpha}{2}}$  y  $b = z_{1-\frac{\alpha}{2}}$ , siendo  $z_{1-\frac{\alpha}{2}}$  el cuantil de orden  $1 - \frac{\alpha}{2}$  de la normal  $\mathcal{N}(0, 1)$ .

Así se tiene que

$$\begin{aligned} & P\left(-z_{1-\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}}\right) = \\ &= P\left(\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}}\right) - P\left(\frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \leq -z_{1-\frac{\alpha}{2}}\right) \approx \\ &\approx P(\mathcal{N}(0, 1) < z_{1-\frac{\alpha}{2}}) - P(\mathcal{N}(0, 1) \leq -z_{1-\frac{\alpha}{2}}) = \\ &= \left(1 - \frac{\alpha}{2}\right) - \frac{\alpha}{2} = \alpha. \end{aligned}$$

Para obtener el intervalo de confianza a partir de esa expresión, se busca formular

$$\left(-z_{1-\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}}\right)$$

como un intervalo alrededor de  $\mu$ .

$$\begin{aligned} & -z_{1-\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}} \Leftrightarrow \\ & \Leftrightarrow -z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} < \bar{X} - \mu < z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \Leftrightarrow \\ & \Leftrightarrow -z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} < \mu - \bar{X} < z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \Leftrightarrow \\ & \Leftrightarrow \bar{X} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} < \mu < \bar{X} + z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \end{aligned}$$

Así, el intervalo de confianza será

$$\left( \bar{X} - z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} \right).$$

## 2.1. Código análisis univariable

```
1
2 library(readxl)
3 library(ggplot2)
4 library(dplyr)
5 library(tidyr)
6 #install.packages("gridExtra")
7 library(gridExtra)
8
9
10 datos <- read_excel("datosine.xlsx", sheet = "Sheet1")
11
12 #Creamos carpeta para guardar gráficos
13 dir.create("graficos_descriptivos", showWarnings = FALSE)
14
15 #Colores que vamos a usar (dobles para si nos pasamos que siga tirando)
16 colores<-c("peru","salmon","lightblue","lightgreen","lightpink","purple",
17   "steelblue","#11ff77","cyan","lightyellow","darkgreen","peru","salmon",
18   "lightblue","lightgreen","lightpink","purple","steelblue","#11ff77","cyan",
19   ", "yellow","darkgreen")
20 i<-0
21
22 # bucle análisis descriptivo
23 for (col in names(datos)) {
24   variable <- datos[[col]]
25   cat("\n\n===== Análisis de:", col, "=====\n")
26
27   if (is.numeric(variable)) {
28     i<-i+1
29     resumen <- summary(variable)
30     print(resumen)
31
32     # histograma
33     p1 <- ggplot(datos, aes_string(x = col)) +
34       geom_histogram(bins = 30, fill = colores[i], color = "black") +
35       ggtitle(paste("Distribución de", col))
36
37     #guardarlo para exportarlo
38     ggsave(filename = paste0("graficos_descriptivos/", col, "_numerico.png"))
```

```

),
plot = p1,
width = 8, height = 6)

40
41 } else { #variables discretas (cualitativas o cuantitativas)
42   i<-i+1
43   freq <- table(variable)
44   print(freq)

45
46
47 p <- ggplot(datos, aes_string(x = col)) +
48   geom_bar(fill = colores[i], color = "black") +
49   ggtitle(paste("Distribucion de", col)) +
50   theme(axis.text.x = element_text(angle = 45, hjust = 1))

51
52 ggsave(filename = paste0("graficos_descriptivos/", col, "_categorico.
png"),
53         plot = p,
54         width = 8, height = 6)
55 }

56
57 }
58 #filtrado datos Modo Reserva
59 Reservas<-datos[!(datos$ModoReserva %in% c("No_sabe", "No_procede")),
60                         c("ModoReserva", "GastoAlojDíaPersona")] #
61                         # ponemos ambas para poder repetir el codigo
62                         # mas adelante
63
64 Reserva<-Reservas$ModoReserva
65 pp<-ggplot(Reservas, aes_string(x = "ModoReserva")) +
66   geom_bar(fill = "red", color = "black") +
67   ggtitle(paste("Distribucion de", "Reserva","filtrado")) +
68   theme(axis.text.x = element_text(angle = 45, hjust = 1))
69 ggsave(filename = paste0("graficos_descriptivos/", col, "_categorico.png"),
70         plot = pp,
71         width = 8, height = 6)

```

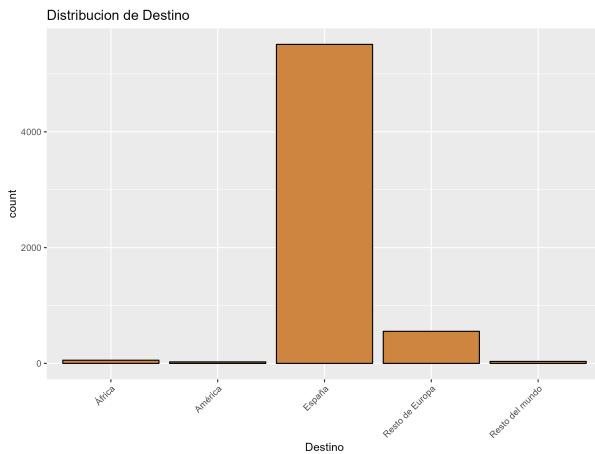
### 3. Análisis de las Variables

Realizaremos ahora un análisis unidimensional de cada una de las variables de la muestra que vamos a estudiar. Comenzamos con las variables discretas cualitativas.

Primero vemos la variable *Destino* mediante su tabla de frecuencias y diagrama de barras (como haremos con el resto de cualitativas).

Destino	Frecuencia
África	55
América	26
España	5511
Resto de Europa	554
Resto del mundo	34

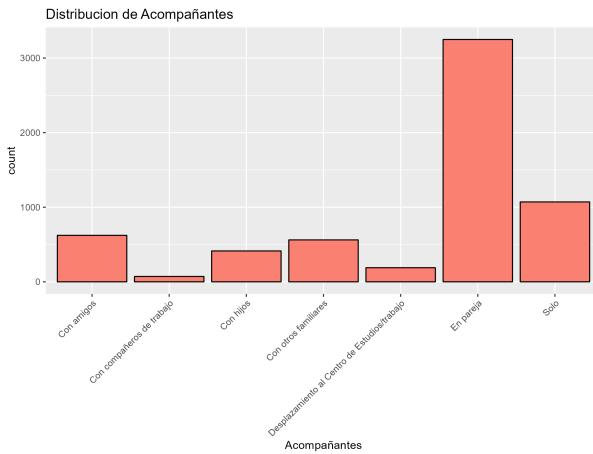
Tabla 1: Tabla frecuencias *Destino*



Se ve como la mayoría de personas optaron por viajar por España  
De forma similar vemos la de *Acompañantes* y como un gran número de personas viajaron en pareja.

Acompañantes	Frecuencia
Con amigos	623
Con compañeros de trabajo	72
Con hijos	414
Con otros familiares	562
Desplazamiento al Centro de Estudios/trabajo	189
En pareja	3249
Solo	1071

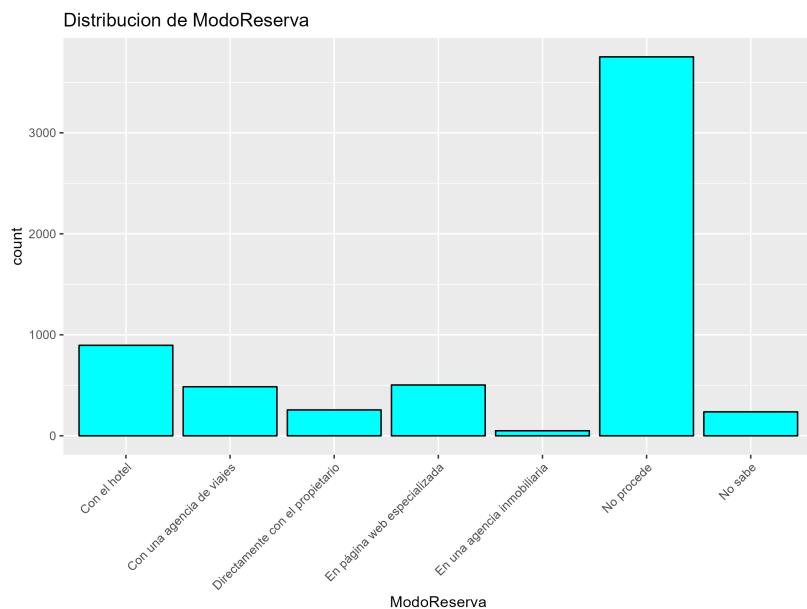
Tabla 2: Frecuencia por tipo de acompañante



Con el *Modo de reserva* al haber un gran número de respuestas de *No procede*, *No sabe* realizamos un nuevo filtrado de esta variable, eliminando estas observaciones, pues para los futuros estudios que vamos a realizar en futuras secciones.

Modo de Reserva	Frecuencia
Con el hotel	896
Con una agencia de viajes	486
Directamente con el propietario	256
En página web especializada	503
En una agencia inmobiliaria	50
No procede	3752
No sabe	237

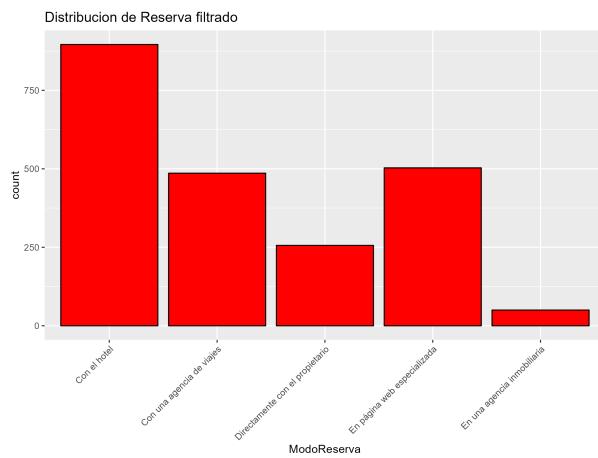
Tabla 3: Frecuencia por Modo de Reserva



Finalmente el *Sexo*, que pese a que parecía que podía ser un número equitativo, pero mediante un contraste <sup>1</sup> Luego trabajamos las discretas cuantitativas, es decir con los *Ingresos del hogar*.

<b>Sexo</b>	<b>Frecuencia</b>
Hombre	2849
Mujer	3331

Tabla 4: Frecuencia por Sexo



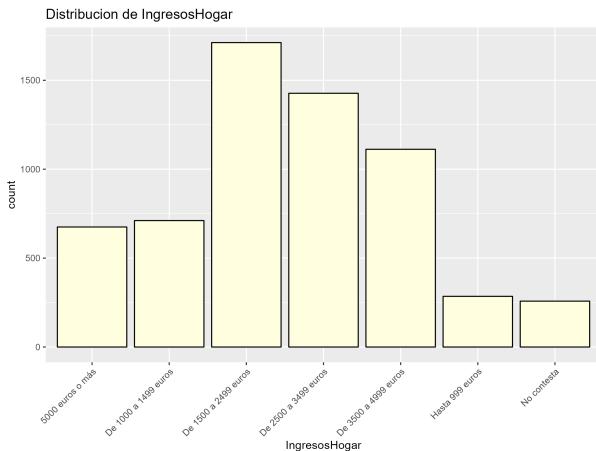
Representando también una tabla de frecuencias y diagrama de barras.

<b>Ingresos del Hogar</b>	<b>Frecuencia</b>
5000 euros o más	675
De 1000 a 1499 euros	711
De 1500 a 2499 euros	1712
De 2500 a 3499 euros	1427
De 3500 a 4999 euros	1112
Hasta 999 euros	285
No contesta	258

Tabla 5: Frecuencia por Ingresos del Hogar

---

<sup>1</sup>se puede ver en 1.5



Y finalmente las variables continuas, que agrupamos todas en una misma tabla indicamos valores minimos y maximo los primer y tercer cuartil además de media y mediana.

Variable	Mín	Q1	Mediana	Media	Q3	Máx
NPERNOC	1.00	2.00	3.00	4.91	6.00	99.00
GastoAlojDíaPersona	0.00	0.00	0.00	17.01	25.22	341.86
GastoTotal	20.17	102.36	242.75	407.12	536.91	9721.60
GastoTotalDía	4.37	42.23	72.87	95.29	127.93	789.88

Tabla 6: Estadísticos descriptivos de variables numéricas

