Martí Biarnés
Laura Castro
Marc Fernández
Silvia Vidal-Acosta

# Turismo en España

## Análisis de los turistas que visitaron España entre 2016 y 2019

### 1.Introducción

El objetivo de este trabajo es realizar un análisis del turismo en España. Para poder realizarlo, hemos extraído los datos que nos proporciona el Instituto Nacional de Estadística, se trata de más de 700.000 observaciones tomadas a lo largo de los últimos 4 años (2016-2019), 355.664 para la tabla "GastoTotal" (Estudio del gasto de los turistas) y 423.893 para la tabla "pernoctas" (estudio de los habitos de duración de la estancia de los turistas) mediante entrevistas realizadas a turistas en los principales puntos de entrada al país (aeropuertos, estaciones, puertos, etcétera).

En concreto, nos centraremos en analizar las siguientes variables:

- Variables continuas: duración y gasto promedio por persona durante el viaje.
- Variables binarias: paquete turístico y procedencia (Europa o resto del mundo).
- Variables categóricas: tipo de alojamiento y medio de transporte.

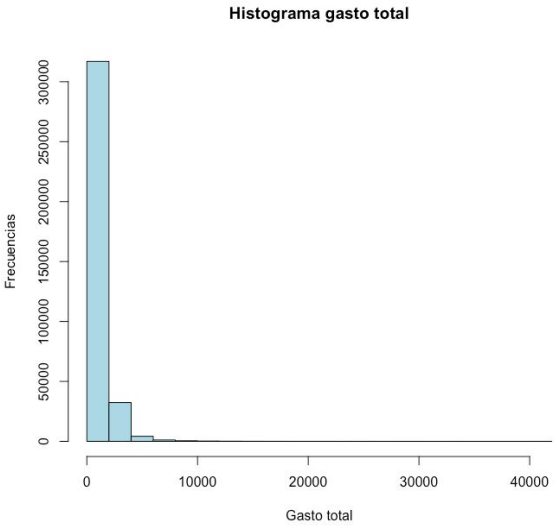
A continuación desarrollamos el trabajo realizando en primer lugar una presentación de las variables y, seguidamente, los cuatro ejercicios requeridos.

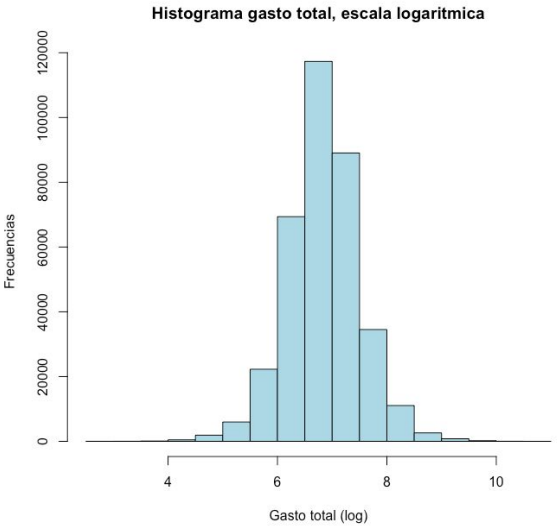
### 2. Tabla resumen de las estadísticas

Antes de empezar con el análisis, para tener claro los datos en los que basaremos nuestro estudio, proporcionamos una tabla resumen de las estadísticas de todas las variables, incluyendo gráficos. Además, en el apéndice se encuentra una tabla de explicación de todas las variables del estudio.

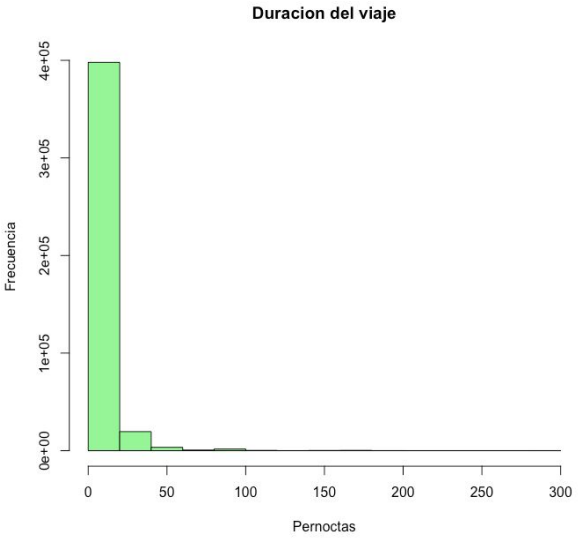
**Gasto total:** en esta variable modelamos la cantidad de dinero que los turistas se gastan en España durante su estancia vacacional. Adjuntamos, primeramente, el histograma de la variable. Debido a que nos salen únicamente 3 columnas (cosa que dificulta el análisis), también adjuntamos un histograma en escala logarítmica, para apreciar mejor las distribución de la variable.

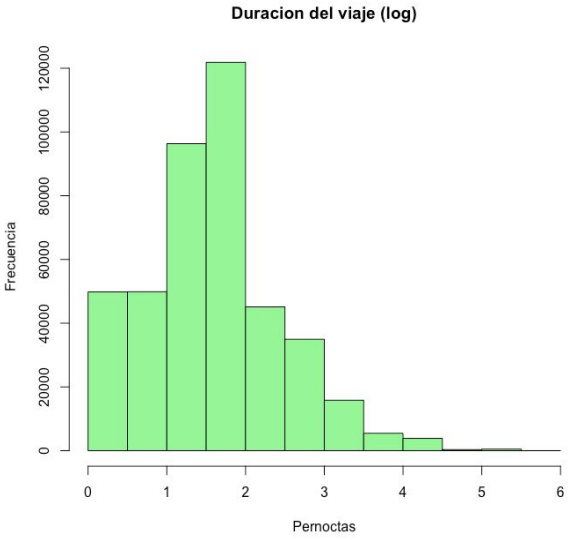
Martí Biarnés
Laura Castro
Marc Fernández
Silvia Vidal-Acosta



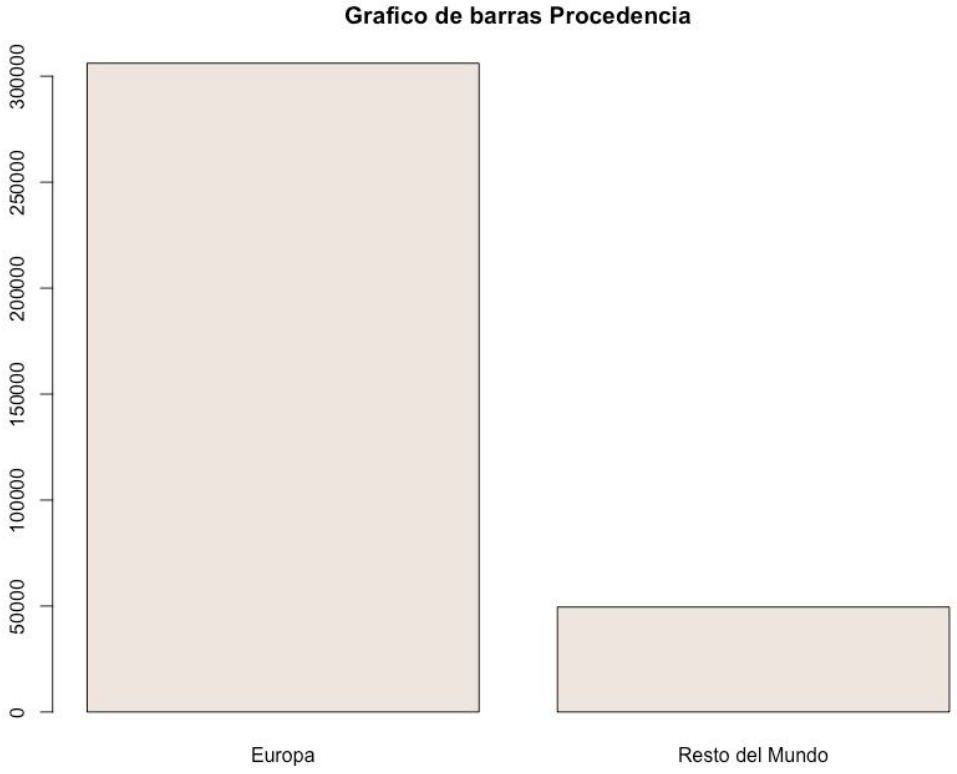

**Duración del viaje:** en esta variable modelamos el número de días que los turistas pasan en España. Igual que en la variable anterior, adjuntamos la versión aplicando logaritmos para poder apreciar el comportamiento de la variable.



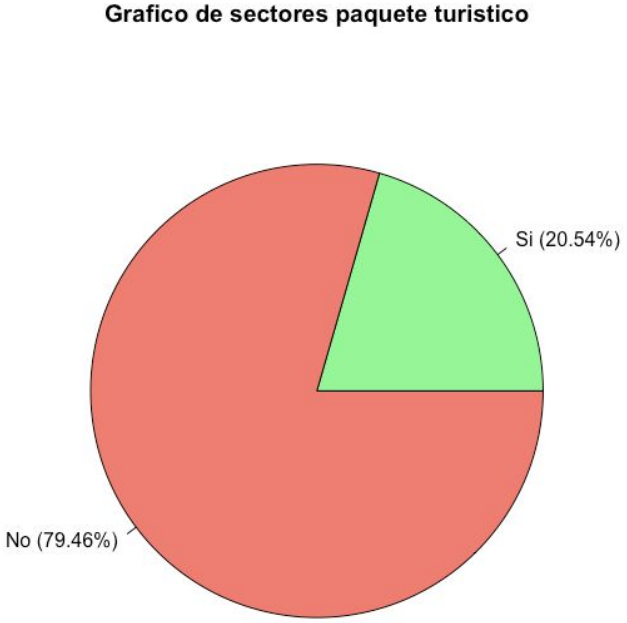




**Procedencia:** esta variable nos muestra el país de procedencia de los turistas que conforman nuestra población. Podemos ver como la mayoría de los turistas provienen de países europeos.

Martí Biarnés
Laura Castro
Marc Fernández
Silvia Vidal-Acosta

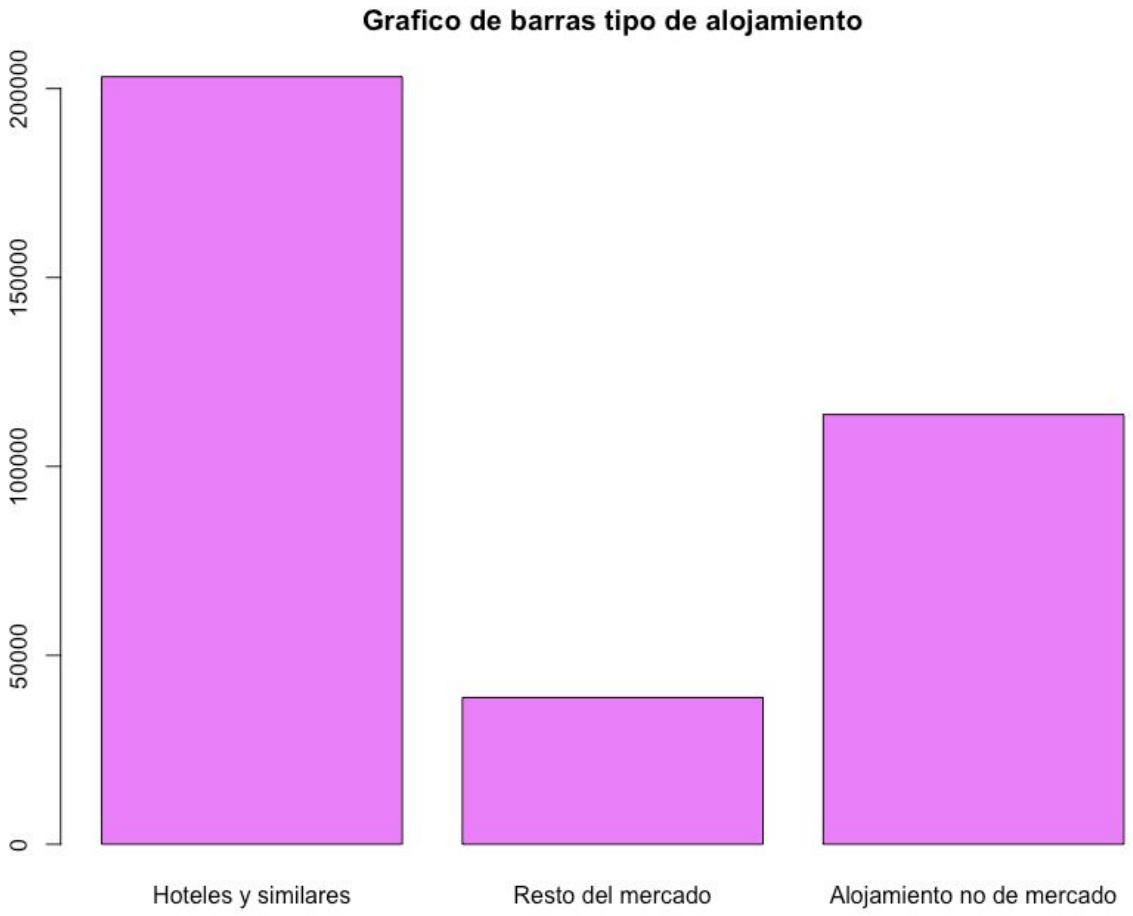



**Paquete turístico:** en esta variable observamos si los turistas han contratado sus vacaciones mediante un paquete turístico o no. Actualmente, los paquetes turísticos ya no son tan populares como en décadas anteriores y esto se puede ver reflejado en nuestra población.



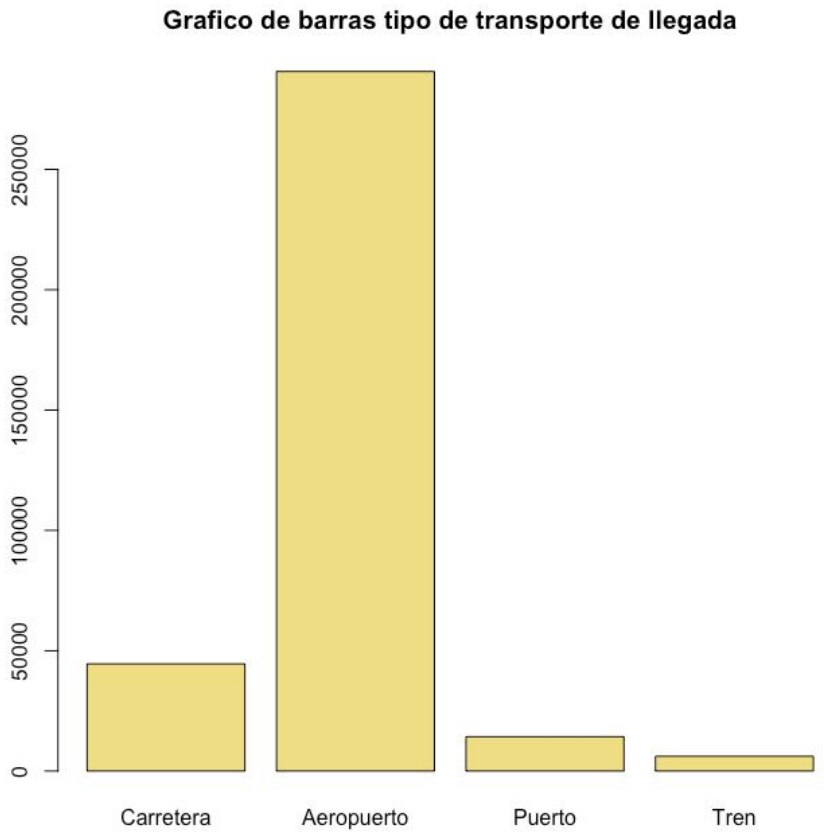

**Tipo de alojamiento:** en esta variable podemos observar que tipo de alojamiento han escogido los turistas durante su visita en España. Para nuestro estudio hemos agrupado los

datos en 3 categorías: “Hoteles y similares”, “Resto del mercado” y “Alojamiento no de mercado”. Vemos como la mayoría prefiere alojarse en hoteles o similares.




**Tipo de transporte:** en esta variable observamos con qué medio de transporte los turistas han llegado a España. En esta variable podemos ver, claramente, que la opción preferida ha sido el avión.

Martí Biarnés
Laura Castro
Marc Fernández
Silvia Vidal-Acosta




## 3. Intervalo de confianza

Para la realización de este ejercicio hemos escogido la variable “Duración del viaje”, estudiando el parámetro “Media de días de duración del viaje”, que nos muestra el número de pernoctaciones media que realizan los turistas en España.
Hemos calculado el intervalo de confianza al 95%, obteniendo el siguiente resultado: (7,611806; 7,679572).
Por lo tanto, con este resultado, si realizamos una muestra con el mismo número de observaciones (mismo valor de *n*), con un 95% de probabilidad la media de ésta muestra estará dentro del intervalo (7,611806; 7,679572).

Nota: el intervalo de confianza tiene una longitud muy pequeña debido al gran tamaño de la muestra, hecho que nos concreta el intervalo.

## 4. Prueba de hipótesis sobre una variable

Para la realización de este ejercicio hemos escogido la variable que nos muestra el gasto de los turistas según su lugar de origen. Esta variable la interpretaremos como una variable “Dummy” analizando el gasto de los turistas Europeos contra el gasto de los turistas del resto del mundo.

Martí Biarnés
Laura Castro
Marc Fernández
Silvia Vidal-Acosta

Planteamos las hipótesis:

- $H_0$ : $\mu_E - \mu_{RESTO} = 0$.
- $H_1$ : $\mu_E - \mu_{RESTO} < 0$.

Introduciendo el comando apropiado en R obtenemos el siguiente p-value: 2.2e^-16. Este valor, al ser muy cercano a 0, nos indica que debemos rechazar la hipótesis nula. Por lo tanto,aceptamos la hipótesis alternativa ( el gasto medio de los turistas europeos es menor al del resto del mundo).
Comparando las medias de gastos de ambos colectivos, que en el caso europeo es de 983,70€ y en el caso de el resto del mundo es de 2.329,71€, toma sentido el resultado obtenido que nos confirma la hipótesis alternativa.

## 5. Intervalo de confianza y prueba de hipótesis

Para este ejercicio hemos tomado como variable la diferencia en el gasto de los turistas entre 2019 y 2018, planteando la hipótesis nula que ambos años el gasto fue igual y, como alternativa, que el gasto en 2019 fue mayor que en 2018.

- $H_0$ : $\mu_{2019} - \mu_{2018} = 0$.
- $H_1$ : $\mu_{2019} - \mu_{2018} > 0$.

Introduciendo los datos en el R hemos creado la siguiente tabla, donde se aprecia el gasto por países en ambos períodos y su diferencia (en la cuarta columna titulada “X2019.1”).

| | Pais | X2018 | X2019 | X2019.1 |
|---|---|---|---|---|
| 1 | Alemania | 1062.5877 | 1099.8143 | 37.226548 |
| 2 | Belgica | 1042.7275 | 1089.0262 | 46.298671 |
| 3 | Francia | 726.0045 | 770.1435 | 44.138999 |
| 4 | Irlanda | 1127.9755 | 1176.8605 | 48.884927 |
| 5 | Italia | 837.2822 | 891.4994 | 54.217193 |
| 6 | Paises Bajos | 1136.7207 | 1147.8880 | 11.167335 |
| 7 | Portugal | 585.2659 | 618.7646 | 33.498677 |
| 8 | UK | 1020.9347 | 1062.1283 | 41.193629 |
| 9 | Suiza | 945.9425 | 979.5275 | 33.584990 |
| 10 | Rusia | 1583.9192 | 1618.4138 | 34.494579 |
| 11 | Paises Nordicos | 1280.5482 | 1286.7599 | 6.211761 |
| 12 | Resto de Europa | 1136.3846 | 1156.7858 | 20.401210 |
| 13 | EEUU | 2100.9892 | 2066.3426 | -34.646654 |
| 14 | Resto de America | 2378.2117 | 2314.5739 | -63.637843 |
| 15 | Resto del Mundo | 2464.0563 | 2440.9489 | -23.107411 |

Martí Biarnés
Laura Castro
Marc Fernández
Silvia Vidal-Acosta

A continuación, utilizando *t.test*, hemos obtenido el siguiente intervalo de confianza: (0.2000194; 38.4568621). Este resultado puede resultar sorprendente teniendo en cuenta que las medias de los gastos anuales superan los 2000 € però no debemos olvidar que nuestro estudio se basa en la diferencia entre ambos años y, en este caso, podemos incluso llegar a observar valores negativos (especialmente fuera de los turistas No-Europeos), factor que nos indica que el gasto medio por persona ha disminuido durante el periodo de estudio en algunas zonas.

Para la realización de la prueba de hipótesis hemos calculado el p-value, obteniendo como resultado: 0.02398. Si utilizamos un nivel de significancia del 5%, podemos rechazar la hipótesis nula llegando a la conclusión que los turistas se han gastado más a lo largo de 2019 que de 2018.

## 6. Independencia

Para realizar una prueba de independencia entre dos variables categóricas, hemos utilizado las variables:"Tipo de alojamiento" y "Medio de Transporte para llegar a España" (de ahora en adelante las nombraremos "Alojamiento" y "Tipo de Transporte").

Para empezar hemos planteado las siguientes hipótesis:

- $H_0$ : no hay relación entre las variables (las variables son independientes).
- $H_1$ : hay relación entre las variables (las variables son dependientes).

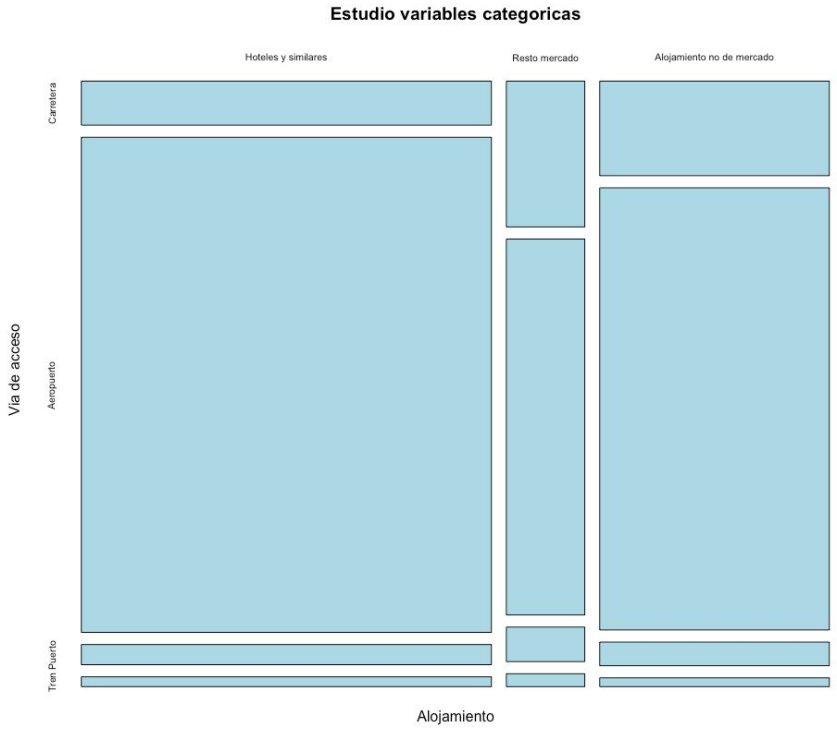
Las siguientes tablas muestran los valores observados y los valores esperados:

```
> chi$observed
          Hoteles y similares Resto mercado Alojamiento no de mercado  Total
Carretera               15713          9947                     18897  44557
Aeropuerto             176689         25651                     88333 290673
Puerto                   7205          2356                      4726  14287
Tren                     3504           882                      1761   6147
Total                  203111         38836                    113717 355664
> chi$expected
          Hoteles y similares Resto mercado Alojamiento no de mercado  Total
Carretera           25445.411      4865.310                 14246.278  44557
Aeropuerto         165996.232     31739.441                 92937.327 290673
Puerto               8158.956      1560.040                  4568.005  14287
Tren                 3510.401       671.209                  1965.390   6147
Total              203111.000     38836.000                113717.000 355664
```

Una vez definidas las hipótesis y teniendo claros los valores observados que tiene nuestra muestra y los valores que debería tener nuestra muestra si las variables fueran

independientes, utilizamos el comando *chisq.test* para calcular el valor del estadístico de contraste, que en este caso es 13244, y el p-value, que en este caso es 2.2e^-16.

Con estos resultados llegamos a la conclusión que debemos rechazar la hipótesis nula. Es decir, las variables son dependientes. Ésta dependencia se puede observar en el *mosaic.plot* de las variables, que adjuntamos a continuación.




## 7. Prueba del buen ajuste

Para realizar la prueba del buen ajuste hemos seleccionado la variable numérica gasto medio por persona, planteando como hipótesis nula que el gasto medio es igual para toda la población. Consecuentemente, la hipótesis alternativa será que el gasto medio por persona difiere a lo largo de la población.

Una vez establecidas las hipótesis, mediante el comando chisq.test hemos obtenido un p-value igual a 2.2e^-16, lo que nos lleva a rechazar la hipótesis nula y aceptar la hipótesis alternativa. De esta manera podemos confirmar que el gasto medio por persona de la muestra es diferente.