Práctica 2: Limpieza y validación de los datos

Javier Reina Gil

30 de diciembre 2017

Contents

1 Descripción del dataset

¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset que voy a utilizar para esta práctica es el construido en la práctica 1. Fue bautizado como Percepción_audiencia_cine_española. La descripción del dataset puede verse en https://github.com/ellibrorojo/stable.

En un principio la pregunta que quería responder era si la audiencia española premiaba a las películas de origen español por encima de otras nacionalidades como la estadounidense, sin embargo al llegar al punto de comprobar que se cumple la hipótesis de normalidad vi que no era así y por tanto no podría aplicar el método ANOVA. Así pues, la pregunta que finalmente trataré de responder es si la audiencia global (puesto que los datos salen de una página web internacional) valora mejor o peor las películas según su origen.

2 Limpieza de los datos

Comenzamos cargando el dataset.

Se ha visto que el campo Distribuidora tiene valores duplicados debido a pequeñas diferencias a la hora de escribir. Vamos a comenzar por corregir este punto. Aprovecharé para convertir los campos Largometraje, Distribuidora y Nacionalidad a valores numéricos. Esto no es necesario para el propósito de esta práctica, pero así simulo el tratamiento que podríamos tener que hacer en caso de tratar con datos personales. Además, si en lugar de tratar 375 registros estuviéramos tratando con cientos de miles o más, esta conversión de tipos mejoraría el rendimiento notablemente.

Destacar que hay tres nacionalidades con muy pocas películas en el dataset, por ese motivo se han agrupado Canadá, Francia, Nueva Zelanda y Suecia bajo el nombre Otros.

```
remove(list=ls())
library(plyr)
path <- "F:/Box Sync/MDS/S1 - Tipología y ciclo de vida de los datos/Práctica Limpieza y validación de l

data <- read.csv2(paste(path, "input.csv", sep = ''))

mapping_distribuidoras <- NULL
mapping_distribuidoras$nombre <- levels(data$Distribuidora)
mapping_distribuidoras <- as.data.frame(mapping_distribuidoras)
mapping_distribuidoras$id <- seq(1, nrow(mapping_distribuidoras))
mapping_distribuidoras$id2 <- mapping_distribuidoras$id
#Manualmente corrijo los valores duplicados, relacionándolos con un identificador único
mapping_distribuidoras$id2[6] <- 5
mapping_distribuidoras$id2[8] <- 10
mapping_distribuidoras$id2[9] <- 10
mapping_distribuidoras$id2[15] <- 16
mapping_distribuidoras$id2[18] <- 19
mapping_distribuidoras$id2[22] <- 21
#Transformo los nombrs en ids
data$Distribuidora <- mapvalues(data$Distribuidora, from=mapping_distribuidoras$nombre, to=mapping_distri
data$Distribuidora <- as.integer(as.character(data$Distribuidora))
remove(mapping_distribuidoras)

mapping_paises <- NULL
mapping_paises$nombre <- levels(data$Nacionalidad)
mapping_paises <- as.data.frame(mapping_paises)
mapping_paises$id <- seq(1, nrow(mapping_paises))
mapping_paises$id2 <- mapping_paises$id
#Asigno manualmente los orígenes poco populares al id 9
mapping_paises$id2[1] <- 9
mapping_paises$id2[4] <- 9
mapping_paises$id2[5] <- 9
mapping_paises$id2[7] <- 9
#Transformo nombre del país en id
data$Nacionalidad <- mapvalues(data$Nacionalidad, from=mapping_paises$nombre, to=mapping_paises$id2)
data$Nacionalidad <- as.factor(as.character(data$Nacionalidad))
remove(mapping_paises)

mapping_largometrajes <- NULL
mapping_largometrajes$nombre <- levels(data$Largometraje)
mapping_largometrajes <- as.data.frame(mapping_largometrajes)
mapping_largometrajes$id <- seq(1, nrow(mapping_largometrajes))
data$Largometraje <- mapvalues(data$Largometraje, from=mapping_largometrajes$nombre, to=mapping_largome
data$Largometraje <- as.integer(as.character(data$Largometraje))
remove(mapping_largometrajes)
```

2.1 Selección de los datos de interés a analizar

¿Cuáles son los campos más relevantes para responder al problema?

Naturalmente el campo nacionalidad es fundamental para poder responder la pregunta. También lo son el campo Puntuacion y TasteInd.

2.2 Elementos vacíos

¿Los datos contienen ceros o elementos vacíos? ¿Y valores extremos? ¿Cómo gestionarías cada uno de estos casos?

No, los datos no contienen ceros ni elementos vacíos.

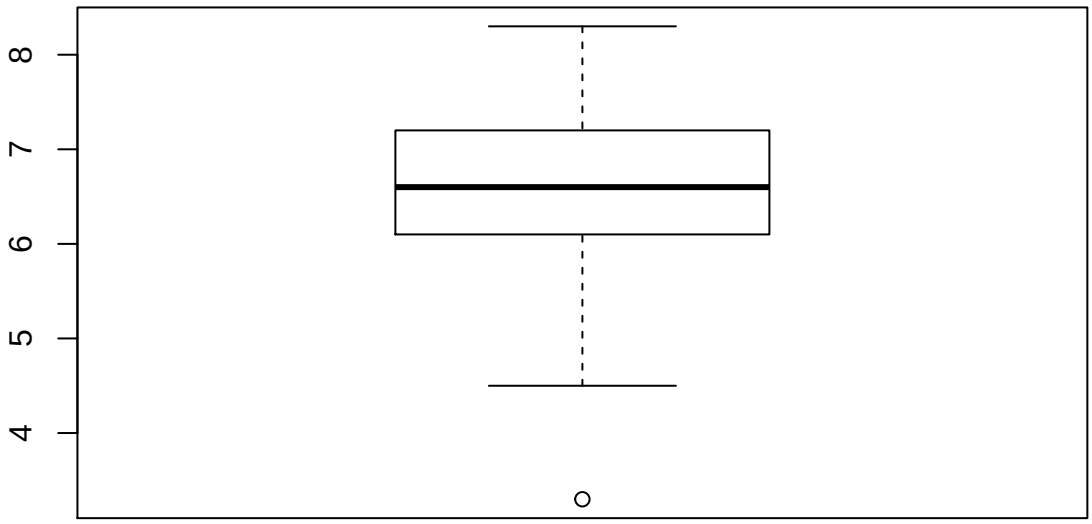
```
hayVacios <- 0

for (i in 1:length(colnames(data)))
{
  hayVacios <- hayVacios + length(which(is.na(data[colnames(data)[i]])))
  hayVacios <- hayVacios + length(which(is.null(data[colnames(data)[i]])))
  hayVacios <- hayVacios + length(which(trimws(as.character(data[colnames(data)[i]])) == ""))
  hayVacios <- hayVacios + length(which(trimws(as.character(data[colnames(data)[i]])) == "-"))
  hayVacios <- hayVacios + length(which(trimws(as.character(data[colnames(data)[i]])) == 0))
}
hayVacios <- as.logical(hayVacios)

hayVacios
```

```
## [1] FALSE
```

Comprobamos que el campo hayVacios vale FALSE, indicándonos que ningún elemento de la tabla está sin definir. Si los hubiera y fueran vacíos en campos relevantes debería descartar esos registros para no desvirtuar el estudio, ya que la única alternativa sería generar yo el valor faltante, y eso sería arriesgado. Sólo lo haría en caso de un valor determinista, que sea indiscutiblemente resultado de otros campos que sí están informados.

Para localizar valores extremos vamos a pintar un boxplot.

```
boxplot(data$Puntuacion)
```



Se observa que hay un único outlier:

```
length(boxplot.stats(data$Puntuacion)$out)
```

```
## [1] 1
```

Se trata de la película que tiene peor puntuación.

Conociendo el origen de los datos, y teniendo en cuenta que vamos a realizar un estudio y no un análisis predictivo, el tratamiento que voy a aplicar a los outliers es aceptarlos sin más, puesto que en principio no se deben a errores en los datos sino que son películas que tuvieron mala crítica.

3 Análisis de los datos

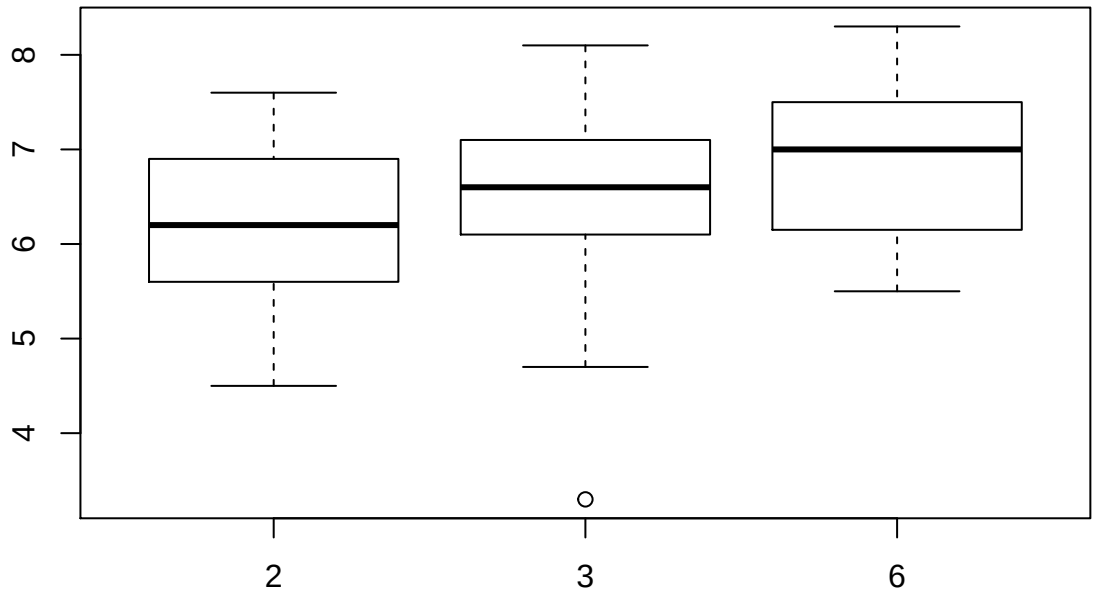
3.1 Grupos

Selección de los grupos que se quieren analizar/comparar.

El estudio a realizar consistirá en disgregar los registros por nacionalidad y ver cómo es la puntuación recibida para cada grupo.

Por nacionalidad: No es necesario realizar la creación de grupos explícitamente ya que el valor del campo Nacionalidad es suficiente para discernir. Los valores posibles son 2, 3, 6 y 9. Dado que el grupo 9, el formado por nacionalidades con poca penetración en España, es poco abundante, este lo voy a descartar para el estudio, y únicamente compararemos películas de origen español (2), estadounidense (3) e inglés (6).

```
data <- data[which(data$Nacionalidad != 9),]
data <- droplevels(data)
boxplot(data$Puntuacion ~ data$Nacionalidad)
```

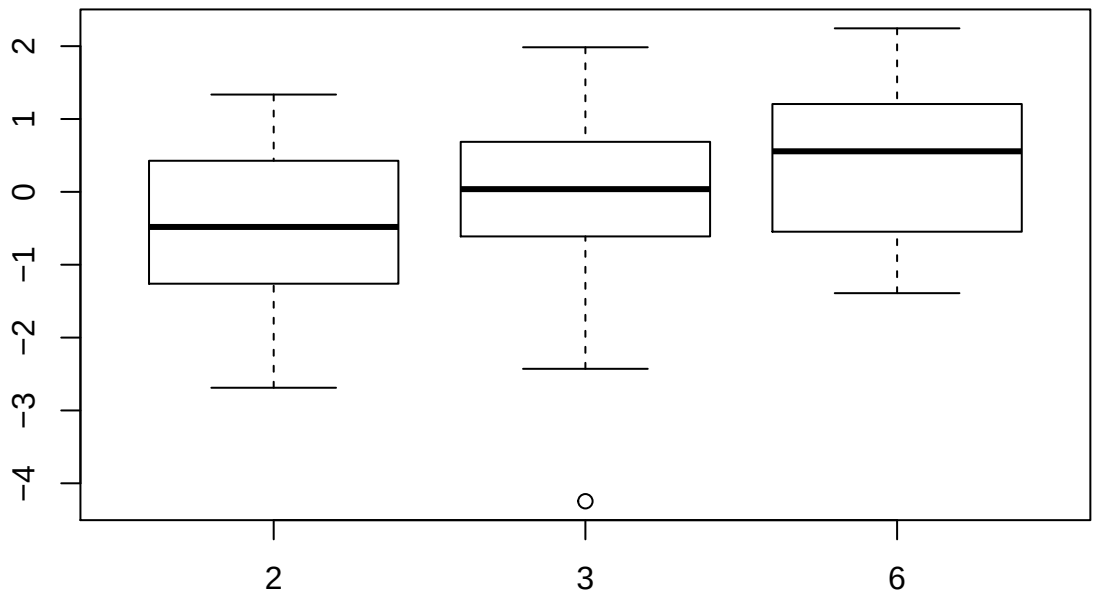


3.2 Transformaciones

Comprobación de la normalidad y homogeneidad de la varianza. Si es necesario (y posible), aplicar transformaciones que normalicen los datos.

Aplicaremos a cada grupo la función standardize para hacer más sencillos los valores.

```
data$Puntuacion <- standardize(data$Puntuacion)
data$TasteInd <- standardize(data$TasteInd)
data$Puntuacion <- as.numeric(data$Puntuacion)

boxplot(data$Puntuacion ~ data$Nacionalidad)
```



La primera condición que se debe satisfacer para aplicar ANOVA es la independencia de los grupos. Esta condición se cumple puesto que todas las películas pertenecen únicamente a una nacionalidad y en principio no hay relación entre la nacionalidad de una película con la del resto.

Comprobamos su normalidad mediante dos tests:

- **el test de Lilliefors (Kolmogorov-Smirnov)**

```
lillie.test(data[which(data$Nacionalidad==2),]$Puntuacion)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data[which(data$Nacionalidad == 2), ]$Puntuacion
## D = 0.11395, p-value = 0.08323
```

```
lillie.test(data[which(data$Nacionalidad==3),]$Puntuacion)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data[which(data$Nacionalidad == 3), ]$Puntuacion
## D = 0.054696, p-value = 0.05155
```

```
lillie.test(data[which(data$Nacionalidad==6),]$Puntuacion)
```

```
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  data[which(data$Nacionalidad == 6), ]$Puntuacion
```

```
## D = 0.13335, p-value = 0.07082
```

- **El test de Shapiro**

```
shapiro.test(data[which(data$Nacionalidad==2),]$Puntuacion)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[which(data$Nacionalidad == 2), ]$Puntuacion
## W = 0.95324, p-value = 0.03715
```

```
shapiro.test(data[which(data$Nacionalidad==6),]$Puntuacion)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data[which(data$Nacionalidad == 6), ]$Puntuacion
## W = 0.9438, p-value = 0.04652
```

Vemos que según el test de Lilliefors los tres grupos pasan el test al ser el p-valor mayor que 0.05, mientras que el test de Shapiro no lo pasan. El caso de nacionalidad = 3 no se ha pasado por Shapiro al ser un grupo demasiado numeroso para esta prueba. Como conclusión, continuaré usando estos grupos asumiendo normalidad si bien ya hemos visto que ésta no es muy holgada.

A continuación vamos a verificar la homogeneidad de la varianza (homocedasticidad) para acabar de estar seguros de que podemos aplicar ANOVA, aplicando el

- **Test de Levene**

```
leveneTest(data$Puntuacion, data$Nacionalidad, center = "mean")
```

```
## Levene's Test for Homogeneity of Variance (center = "mean")
##        Df F value Pr(>F)
## group   2  1.7565 0.1741
##       357
```

El valor de $Pr(> F)$ nos indica que no se detectan diferencias en las varianzas de los 4 grupos.

Replicamos con el

- **Test de Barlett**

```
bartlett.test(Puntuacion ~ Nacionalidad, data=data)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Puntuacion by Nacionalidad
## Bartlett's K-squared = 0.78198, df = 2, p-value = 0.6764
```

y obtenemos lo mismo (un p-valor alto que nos obliga a quedarnos con la hipótesis nula).

La asunción de homocedasticidad parece razonable. Era de esperar al ver los boxplots pintados anteriormente, en los que se percibe que las características de las cajas son similares.

3.3 Pruebas estadísticas

Aplicación de pruebas estadísticas (tantas como sea posible) para comparar los grupos de datos.

En primer lugar vamos a aplicar la prueba de

- **ANOVA**

```
p1t1 <- anova(aov(Puntuacion~ Nacionalidad, data=data))
p1t1
```

```
## Analysis of Variance Table
##
## Response: Puntuacion
##               Df Sum Sq Mean Sq F value     Pr(>F)
## Nacionalidad   2   18.7  9.3482  9.8068 0.00007142 ***
## Residuals    357  340.3  0.9532
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Un valor bajo de $Pr(> F)$ indica que podemos descartar la hipótesis nula y por tanto hay, al menos uno de los grupos, cuya media poblacional para el campo puntuación es distinta a los otros dos.

Ahora aplicaremos varios métodos no paramétricos, es decir, métodos que no requieren asunciones sobre la distribución de los datos a estudiar.

Vamos a realizar el

- **Test de Mann-Whitney-Wilcoxon**

Este test podemos aplicarlo sin asumir normalidad en los datos (http://www.r-tutor.com/elementary-statistics/non-parametric-methods/mann-whitney-wilcoxon-test), algo muy valioso en nuestro caso al tener grupos de normalidad dudosa.

Se ha de aplicar a parejas de grupos, por lo que debemos practicarlo en 3 ocasiones dadas las nacionalidades que consideramos (2-3, 2-6, 3-6)

2-3

```
p1t2.2_3 <- wilcox.test(Puntuacion ~ Nacionalidad, data=data[which(data$Nacionalidad != 6), ])
p1t2.2_3
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Puntuacion by Nacionalidad
## W = 5183, p-value = 0.002085
## alternative hypothesis: true location shift is not equal to 0
```

2-6

```
p1t2.2_6 <- wilcox.test(Puntuacion ~ Nacionalidad, data=data[which(data$Nacionalidad != 3), ])
p1t2.2_6
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Puntuacion by Nacionalidad
## W = 561.5, p-value = 0.0001095
## alternative hypothesis: true location shift is not equal to 0
```

3-6

```
p1t2.3_6 <- wilcox.test(Puntuacion ~ Nacionalidad, data=data[which(data$Nacionalidad != 2), ])
p1t2.3_6
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Puntuacion by Nacionalidad
## W = 4196, p-value = 0.02881
## alternative hypothesis: true location shift is not equal to 0
```

Las tres instancias ejecutadas del test de Mann-Whitney-Wilcoxon dejan como resultado un p-valor inferior a 0.05, lo cual nos permite confirmar que las medias de puntuación son distintas para cada pareja de nacionalidades.

Otro test que podemos aplicar es el

- **Test de Kruskal-Wallis**

, que funciona del mismo modo que el anterior (http://www.r-tutor.com/elementary-statistics/non-parametric-methods/kruskal-wallis-test).

2-3

```
p1t3.2_3 <- kruskal.test(Puntuacion ~ Nacionalidad, data=data[which(data$Nacionalidad != 6), ])
p1t3.2_3
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Puntuacion by Nacionalidad
## Kruskal-Wallis chi-squared = 9.4783, df = 1, p-value = 0.002079
```

2-6

```
p1t3.2_6 <- kruskal.test(Puntuacion ~ Nacionalidad, data=data[which(data$Nacionalidad != 3), ])
p1t3.2_6
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Puntuacion by Nacionalidad
## Kruskal-Wallis chi-squared = 14.995, df = 1, p-value = 0.0001078
```

3-6

```
p1t3.3_6 <- kruskal.test(Puntuacion ~ Nacionalidad, data=data[which(data$Nacionalidad != 2), ])
p1t3.3_6
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Puntuacion by Nacionalidad
## Kruskal-Wallis chi-squared = 4.7834, df = 1, p-value = 0.02874
```

Con este test reproducimos casi exactamente los mismos resultados que con el anterior, es decir que también nos indica que los tres grupos son poblaciones no idénticas.

Adicionalmente voy a tratar de responder la pregunta que inicialmente quería responder y luego descarté por no poder responderla con ANOVA. Voy a aplicar el test de Wilcoxon a los tres grupos para ver si la audiencia española distingue entre las nacionalidades de las películas.

2-3

```
p2t1.2_3 <- wilcox.test(TasteInd ~ Nacionalidad, data=data[which(data$Nacionalidad != 6), ])
p2t1.2_3
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  TasteInd by Nacionalidad
## W = 5639, p-value = 0.0196
## alternative hypothesis: true location shift is not equal to 0
```

2-6

```
p2t1.2_6 <- wilcox.test(TasteInd ~ Nacionalidad, data=data[which(data$Nacionalidad != 3), ])
p2t1.2_6
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  TasteInd by Nacionalidad
## W = 781.5, p-value = 0.03098
## alternative hypothesis: true location shift is not equal to 0
```

3-6

```
p2t1.3_6 <- wilcox.test(TasteInd ~ Nacionalidad, data=data[which(data$Nacionalidad != 2), ])
p2t1.3_6
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  TasteInd by Nacionalidad
## W = 4844, p-value = 0.3439
## alternative hypothesis: true location shift is not equal to 0
```

Los únicos grupos cuya no-identicidad podemos garantizar son los grupos 3-6. Curiosamente, estas dos nacionalidades son Estados Unidos y Reino Unido. Por tanto según los datos de esta tabla, podemos confirmar que sí hay distinción entre la percepción que tiene la audiencia española respecto a películas de origen español y películas de otras nacionalidades.

4 Representación

Representación de los resultados a partir de tablas y gráficas.

Voy a resumir en una tabla las pruebas que se han realizado.

En esta tabla lo referente a la primera pregunta:

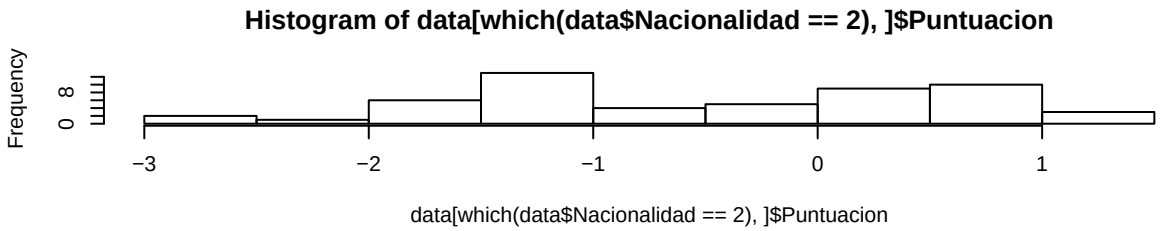
```
summ1 <- NULL
summ1$Test <- c('P1 - ANOVA - Todos',
                'P1 - Wilcox  - 2-3',
                'P1 - Wilcox  - 2-6',
                'P1 - Wilcox  - 3-6',
                'P1 - Kruskal - 2-3',
                'P1 - Kruskal - 2-6',
                'P1 - Kruskal - 3-6'
              )
summ1 <- as.data.frame(summ1)
```

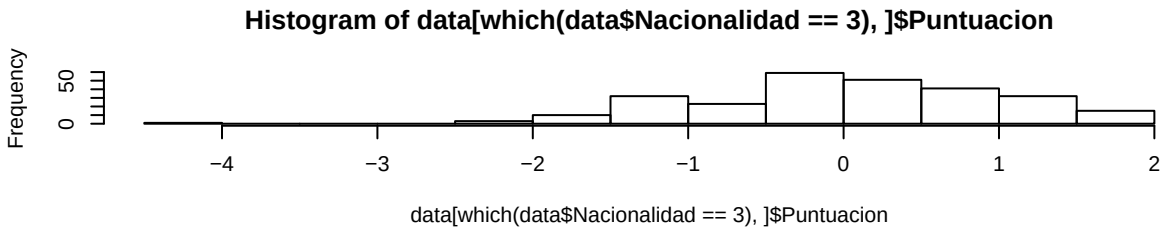
```
summ1$Resultado <- c(p1t1$'Pr(>F)'[1],
                     p1t2.2_3$p.value,
                     p1t2.2_6$p.value,
                     p1t2.3_6$p.value,
                     p1t3.2_3$p.value,
                     p1t3.2_6$p.value,
                     p1t3.3_6$p.value
                     )

summ1$Resultado <- round(summ1$Resultado, digits=4)
summ1$'H0 descartable (Grupos son no-idénticos)' <- summ1$Resultado < 0.05
summ1
```

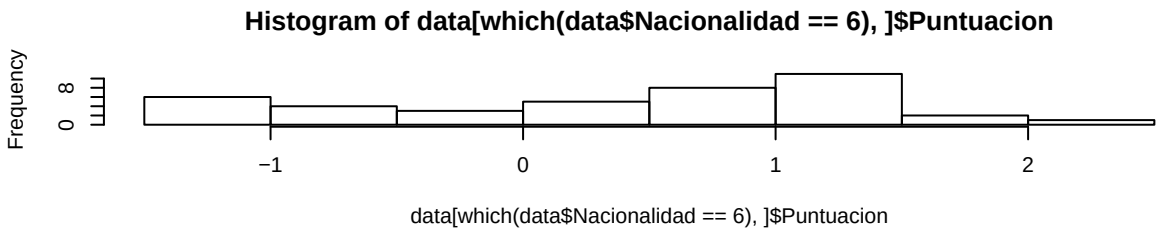
```
##                 Test Resultado H0 descartable (Grupos son no-idénticos)
## 1 P1 - ANOVA - Todos    0.0001                                      TRUE
## 2 P1 - Wilcox  - 2-3    0.0021                                      TRUE
## 3 P1 - Wilcox  - 2-6    0.0001                                      TRUE
## 4 P1 - Wilcox  - 3-6    0.0288                                      TRUE
## 5 P1 - Kruskal - 2-3    0.0021                                      TRUE
## 6 P1 - Kruskal - 2-6    0.0001                                      TRUE
## 7 P1 - Kruskal - 3-6    0.0287                                      TRUE
```

A continuación unos histogramas que muestran las diferentes distribuciones. En ellos se intuye lo que los análisis confirman: que no son grupos confundibles.

```
par(mfrow=c(3,1))
hist(data[which(data$Nacionalidad==2),]$Puntuacion)
hist(data[which(data$Nacionalidad==3),]$Puntuacion)
hist(data[which(data$Nacionalidad==6),]$Puntuacion)
```







Hacemos lo mismo con la segunda pregunta:

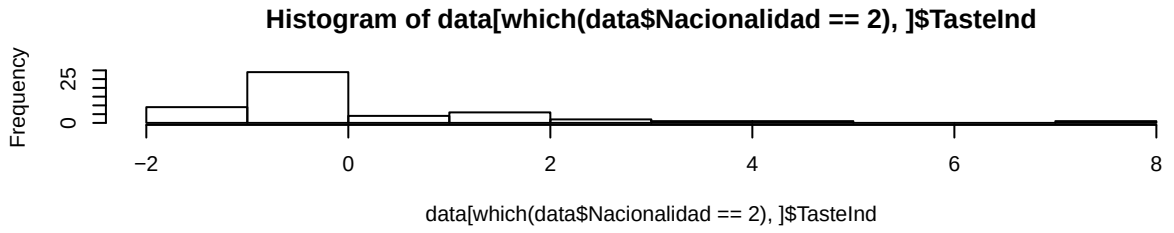
```
summ2 <- NULL
summ2$Test <- c('P2 - Wilcox  - 2-3',
                'P2 - Wilcox  - 2-6',
                'P2 - Wilcox  - 3-6'
              )
summ2 <- as.data.frame(summ2)

summ2$Resultado <- c( p2t1.2_3$p.value,
                      p2t1.2_6$p.value,
                      p2t1.3_6$p.value
                    )

summ2$Resultado <- round(summ2$Resultado, digits=4)
summ2$'H0 descartable (Grupos son no-idénticos)' <- summ2$Resultado < 0.05
summ2
```

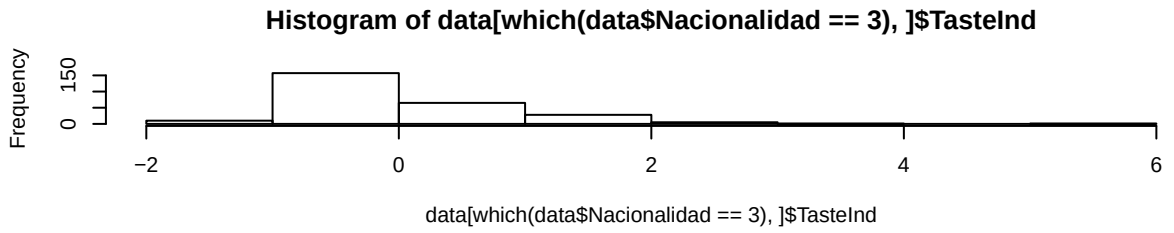
```
##                 Test Resultado H0 descartable (Grupos son no-idénticos)
## 1 P2 - Wilcox  - 2-3    0.0196                                      TRUE
## 2 P2 - Wilcox  - 2-6    0.0310                                      TRUE
## 3 P2 - Wilcox  - 3-6    0.3439                                     FALSE
```

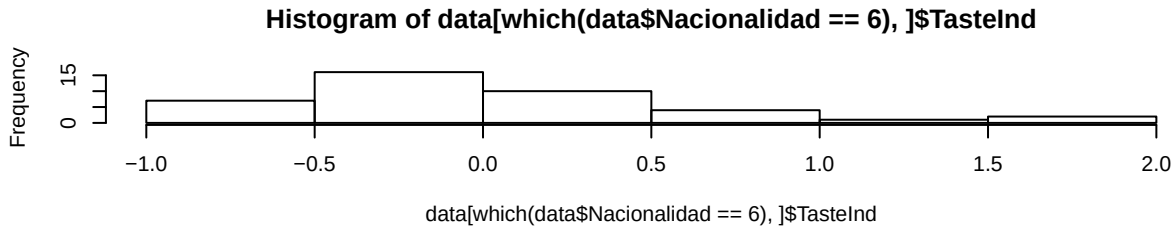
Pintamos los histogramas de TasteInd. Para este campo no se intuye lo que obtenemos en los tests, posiblemente debido a la falta de normalidad o la presencia de outliers.

```
par(mfrow=c(3,1))
hist(data[which(data$Nacionalidad==2),]$TasteInd)
```

```
hist(data[which(data$Nacionalidad==3),]$TasteInd)
hist(data[which(data$Nacionalidad==6),]$TasteInd)
```







5 Resolución del problema

A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

La pregunta de si hay relación entre la puntuación que la comunidad internacional otorga a las películas y su origen se puede responder afirmativamente. En la tabla summ1 vemos que se puede descartar la hipótesis nula de que los grupos pertenecen a las mismas poblaciones. Por tanto, poder descartarse esta hipótesis implica que los grupos son distintos y sí que hay una relación entre el origen de una película y su puntuación.

Respecto a la pregunta de si la audiencia española hace distinciones a la hora de apoyar cine de un país u otro, también podemos dar respuesta: los españoles perciben distinto el cine español y el extranjero. Se justifica porque hay diferencia entre los grupos 2 y 3 y 2 y 6 pero no entre 3 y 6. Esto indica que la audiencia española trata distinto al cine español y el extranjero pero no distingue entre los orígenes si éstos son extranjeros.

Finalmente guardamos la tabla que hemos utilizado en el análisis y las dos tablas resumen.

```
write.csv2(data, paste(path, "/datosAnalizados.csv", sep=""), quote=FALSE, row.names = FALSE)
write.csv2(summ1, paste(path, "/summ1.csv", sep=""), quote=FALSE, row.names = FALSE)
write.csv2(summ2, paste(path, "/summ2.csv", sep=""), quote=FALSE, row.names = FALSE)
```

6 Código

Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos.