PROYECTO INFERENCIA ESTADISTICA-NETFLIX

Diana Balanta-Danna Espinosa-Cristian Perafan-Geovanny Quintero

# **Objetivo 04.**

## **Identificar entre 6 géneros cuál tiene el mayor puntaje IMDb score promedio.**

**Generos:** Acción y aventura, niños y peliculas familiares, comedia,documentales,drama, horror.

## **Anova**

**Definición de variables:**

“score de los productos audiovisuales del género acción y aventura”

“score de los productos audiovisuales del género niños y peliculas familiares”

“score de los productos audiovisuales del género comedia”

“score de los productos audiovisuales del género documentales”

“score de los productos audiovisuales del género dramas”

“score de los productos audiovisuales del género peliculas de horror”

**Planteamiento de hipótesis**

: ““No existe diferencia entre la media de los score de los diferentes géneros”

: “Por lo menos un par de medias de los scores de los géneros son diferentes”

**Criterio de rechazo:**

Rechazo SII

**Implementación en R:**

library(readxl)  
library(rapportools)  
  
  
rutaExcel <-file.choose()  
  
anovaData <- read\_excel(rutaExcel,sheet = "Anova")  
  
attach(anovaData)  
  
anova <- aov(Score~Genres)  
summary(anova)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Genres 5 25.07 5.015 6.567 6.2e-05 \*\*\*  
## Residuals 60 45.82 0.764   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Análisis de resultados:**

Del anterior resumen generado por R podemos obtener , evaluamos el en nuestro criterio de rechazo:

Con un nivel de significancia de 5%, hay suficiente evidencia para rechazar la HO, por lo tanto existe por lo menos una diferencia entre el score de un par de generos.

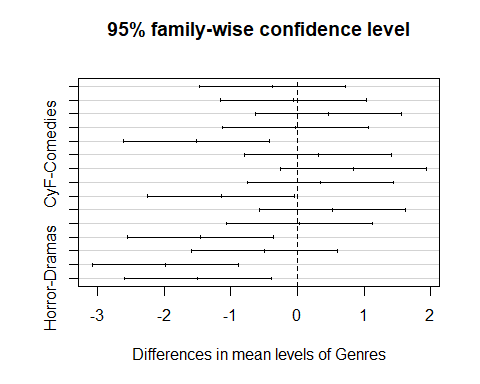
**Post Anova**

TukeyHSD(anova)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = Score ~ Genres)  
##   
## $Genres  
## diff lwr upr p adj  
## Comedies-AyA -0.37272727 -1.4696134 0.72415881 0.9160403  
## CyF-AyA -0.06363636 -1.1605225 1.03324972 0.9999787  
## Documentaries-AyA 0.46363636 -0.6332497 1.56052245 0.8132209  
## Dramas-AyA -0.02727273 -1.1241588 1.06961336 0.9999997  
## Horror-AyA -1.51818182 -2.6150679 -0.42129573 0.0018337  
## CyF-Comedies 0.30909091 -0.7877952 1.40597700 0.9608371  
## Documentaries-Comedies 0.83636364 -0.2605225 1.93324972 0.2332953  
## Dramas-Comedies 0.34545455 -0.7514315 1.44234063 0.9379424  
## Horror-Comedies -1.14545455 -2.2423406 -0.04856846 0.0357300  
## Documentaries-CyF 0.52727273 -0.5696134 1.62415881 0.7178359  
## Dramas-CyF 0.03636364 -1.0605225 1.13324972 0.9999987  
## Horror-CyF -1.45454545 -2.5514315 -0.35765937 0.0031753  
## Dramas-Documentaries -0.49090909 -1.5877952 0.60597700 0.7742730  
## Horror-Documentaries -1.98181818 -3.0787043 -0.88493209 0.0000234  
## Horror-Dramas -1.49090909 -2.5877952 -0.39402300 0.0023244

**Gráfico de diferencias**

plot(TukeyHSD(anova))



**Análisis PostAnova**

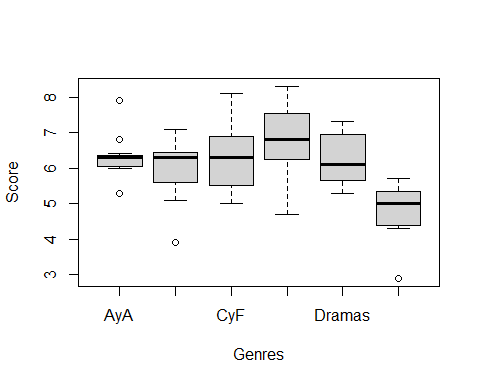
| Género 1 | Género 2 | Diff | Conclusión | Media Mayor |
| --- | --- | --- | --- | --- |
| Children & Family | Action & Adventure | -0.636 | Action & Adventure> Children & Family | Action & Adventure |
| Comedies | Action & Adventure | -0.372 | Action & Adventure>Comedies | Action & Adventure |
| Documentaries | Action & Adventure | 0.463 | Documentaries>Action & Adventure | Documentaries |
| Dramas | Action & Adventure | -0.27 | Action & Adventure> Dramas | Action & Adventure |
| Horror Movies | Action & Adventure | -1.518 | Action & Adventure> Horror Movies | Action & Adventure |
| Comedies | Children & Family | -0.30909091 | Children & Family>Comedies | Children & Family |
| Documentaries | Children & Family | 0.527 | Documentaries> Children & Family | Documentaries |
| Dramas | Children & Family | 0.036 | Dramas> Children & Family | Dramas |
| Horror Movies | Children & Family | -1.454 | Children & Family>Horror Movies | Children & Family |
| Documentaries | Comedies | 0.836 | Documentaries>Comedies | Documentaries |
| Dramas | Comedies | 0.345 | Dramas>Comedies | Dramas |
| Horror movies | Comedies | -1.145 | Comedies>Horror movies | Comedies |
| Dramas | Documentaries | -0.490 | Documentaries>Dramas | Documentaries |
| Horror movies | Documentaries | -1.981 | Documentaries>Horror movies | Documentaries |
| Horror movies | Dramas | -1.490 | Dramas>Horror movies | Dramas |

Con base a la tabla se deduce entonces que: *Documentaries>Action & Adventure>Dramas>Children & Family>Comedies>Horror Movies.*

**Concluyendo, se define que las producciones de tipo Documentaries tienen el mayor puntaje IMDb score promedio, mientras que las de Horror Movies el menor puntaje IMDb promedio.**

**Diagrama de cajas**

boxplot(Score~Genres)



Los datos coinciden con el diagrama de cajas, ya que se puede observar como la media del score de Documentaries es la mayor y la media de Horror movies es la menor. También gracias a los intervalos se deduce que la media con mayor diferencia entre todas es la de horror movies, mientras las demás tienen un comportamiento similar.

# **Objetivo 05**

## **Conocer si existe relación entre el género de la película y el continente donde se produjo. (Prueba de independencia)**

## **Chi-Cuadrado**

Se requiere analizar con un nivel de significancia del 5%, si existe una relacion entre el genero del filme y el continente donde se produjo este. Para ello se agruparon los datos de 3 continentes (Europa, Asia, America) con al menos 6 de los paises de los conforman en cada uno. Obteniendo de este modo la siguiente tabla.

**Importacion de los datos en R:**

library(readxl)  
  
##rutaExcel<-file.choose()  
##rutaExcel <- "C:/Users/Acer/Documents/GEOVANNY/Universidad/CuartoSemestre/Inferencia Estadistica/projecto/Statistical\_Inference\_Project/R\_project/dataSet/BaseDatosBuena.xlsx"  
  
chiData <- read\_excel(rutaExcel,sheet = "ChiCuadrado")  
attach(chiData)  
Europa <- c(chiData[,2])  
Asia <- c(chiData[,3])  
America <- c(chiData[,4])  
  
table <- data.frame(Europa,Asia,America)  
rownames(table) <- c("Action & Adventure","Anime","Children & Family","Classic Movies","Comedies","Cult Movies","Documentaries","Dramas","Horror","Independent Movies","International Movies","LGBTQ Movies","Music & Musicals","Romantic Movies","Sci-Fi & Fantasy","Stand-Up Comedy","Thrillers","Faith & Spirituality","Sports Movies")  
View(table)  
print(table)

## Europa Asia America  
## Action & Adventure 65 194 215  
## Anime 0 51 5  
## Children & Family 54 59 266  
## Classic Movies 6 4 15  
## Comedies 129 288 473  
## Cult Movies 8 5 21  
## Documentaries 93 19 319  
## Dramas 241 535 568  
## Horror 30 39 119  
## Independent Movies 71 113 252  
## International Movies 301 760 272  
## LGBTQ Movies 6 3 48  
## Music & Musicals 33 70 111  
## Romantic Movies 40 114 158  
## Sci-Fi & Fantasy 31 30 103  
## Stand-Up Comedy 32 7 240  
## Thrillers 84 80 184  
## Faith & Spirituality 3 5 20  
## Sports Movies 17 14 67

Con ellos es posible plantear las hipotesis que responden a la pregunta: **¿El continente donde se produce un filme, está relacionado con su genero?**

: “No hay relacion entre el genero de un filme y el continente donde se produce”

: “Hay relacion entre el genero de un filme y el pais continente se produce”

**Criterio de rechazo:**

Rechazo Si

**Implementacion en R:**

chiTest <- chisq.test(table,correct=FALSE)  
chiTest

##   
## Pearson's Chi-squared test  
##   
## data: table  
## X-squared = 1138.4, df = 36, p-value < 2.2e-16

**Análisis de resultados:**

Del anterior resumen generado por R se obtiene , ahora se evalua en el criterio de rechazo:

Con un nivel de significancia de 5%, hay suficiente evidencia para rechazar , por lo tanto, hay relacion entre el genero de un filme y el continente que lo produce.

# **Objetivo 06**

## **Establecer el tipo de relación que hay entre la duración de las peliculas y el IMDb Score**

## **Regresión lineal**

Se desea estudiar la relación que existe entre la duración y el score del producto audiovisual en la plataforma Netflix, específicamente para el género horror, para ello se selecciona una muestra de 118 productos audiovisuales del género horror.

**Definición modelo de regresión**

Modelo regresión lineal simple:

Donde

“Intercepto de la recta”

“Pendiente de la recta”

**Definición de variables**

La variable dependendiente es el score y la variable independiente es la duración.

**Implementación en R**

library(readxl)  
library(QuantPsyc)  
  
regresionData<- read\_excel(rutaExcel,sheet = "regresion")  
  
model = lm(SCORE~DURACION,data = regresionData)  
  
summary(model)

##   
## Call:  
## lm(formula = SCORE ~ DURACION, data = regresionData)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.82057 -0.65423 0.00661 0.72021 2.06972   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.385943 0.496240 4.808 3.86e-06 \*\*\*  
## DURACION 0.028803 0.004926 5.847 3.33e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.006 on 141 degrees of freedom  
## Multiple R-squared: 0.1951, Adjusted R-squared: 0.1894   
## F-statistic: 34.19 on 1 and 141 DF, p-value: 3.331e-08

**Modelo de estimación**

A partir del dataframe obtenido con ayuda de R, podemos establecer el modelo de estimación:

**Interpretación**

Si no hay un incremento en la duración de las películas el score en promedio sera 2,3.

**Interpretación**

Cuando la duración (en minutos) de las películas del género documental aumenta en una unidad (minutos), el score se incrementa en aproximadamente en unidad.

**Interpretación**

Para tenemos **tres interpretaciones**, nuestra primera interpretación es respecto al porcentaje de y la segunda y tercera es respecto a la siguiente tabla:

**Tabla : Para interpretar coeficiente de determinación**

| **Coeficiente** | **Bondad de Ajuste** | **Para estimar** |
| --- | --- | --- |
| 0.6 | MALA | NO ES CONFIABLE PARA HACER ESTIMACIONES |
| 0.6 0.7 | REGULAR | NO ES CONFIABLE PARA HACER ESTIMACIONES |
| 0.7 0.8 | BUENA | ES CONFIABLE PARA HACER ESTIMACIONES |
| 0.8 0.9 | MUY BUENA | ES CONFIABLE PARA HACER ESTIMACIONES |
| 0.9 1.0 | EXCELENTE | ES CONFIABLE PARA HACER ESTIMACIONES |

**Primera Interpretación** El modelo propuesto explica el 19.51% de la variabilidad del score que toma las películas del género documental.

**Segunda Interpretación** El modelo propuesto presenta una bondad de ajuste considerada mala

**Tercera Interpretación** Por lo tanto, este modelo no es confiable para hacer estimaciones.

**Test de correlación**

with(regresionData, cor.test(DURACION, SCORE, alternative="two.sided", method="pearson"))

##   
## Pearson's product-moment correlation  
##   
## data: DURACION and SCORE  
## t = 5.847, df = 141, p-value = 3.331e-08  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.2993120 0.5649391  
## sample estimates:  
## cor   
## 0.4417563

**Interpretación**

Ahora encontramos que el coeficiente de correlación de Pearson es . Por lo tanto podemos definir dos interpretaciones: Como el coeficiente de correlación es positivo decimos que la relación entre Duración y Score IMDb del género horror es directamente proporcional.

La segunda interpretación es respecto a la siguiente tabla:

| **Coeficiente de correlación ()** | **Relación** |
| --- | --- |
| 0.3 | DEBIL O NULA |
| 0.3 0.5 | MODERADA |
| 0.5 0.75 | FUERTE |
| 0.75 1.0 | MUY FUERTE |

De la tabla se infiere que la relación entre Duración y Score IMDb es MODERADA ya que el coeficiente de correlación se encuentra dentro del rango 0.3 0.5 .

**Validación de supuestos**

Para realizar la validación de los supuesto de normalidad y varianza constante (homocedasticidad) primero obtenemos los residuos en R:

library(zoo)  
library(lmtest)  
library(nortest)  
  
residuals<- within(regresionData, {  
 residuals.model <- residuals(model)   
})

***Normalidad***

Definición de hipótesis:

“Los residuos se distribuyen de forma normal” “Los residuos NO se distribuyen de formal normal”

Criterio de rechazo:

Rechazo sii

Realizamos el Lilliefors (Kolmogorov-Smirnov) normality test:

residuos <- residuals(model)  
  
lillie.test(residuos)

##   
## Lilliefors (Kolmogorov-Smirnov) normality test  
##   
## data: residuos  
## D = 0.049669, p-value = 0.5253

Del anterior resultado VrP que obtuvimos en R evaluamos en nuestro criterio de rechazo:

A un nivel de significancia del 5% hay suficiente evidencia para rechazar Ho por lo tanto se cumple que los residuos se comportan de forma normal y se valida el supuesto de normalidad.

***Varianza constante***

Definición de hipótesis:

“Los residuos presentan varianza constante (Homocedasticidad)”

“Los residuos NO presentan varianza constante (Heterocedasticidad)”

Criterio de rechazo:

Rechazo sii

Realizamos el Breusch-Pagan test con R e imprimimos resultado :

bptest(SCORE ~ DURACION, varformula = ~ fitted.values(model), studentize=FALSE, data=residuals)

##   
## Breusch-Pagan test  
##   
## data: SCORE ~ DURACION  
## BP = 0.26203, df = 1, p-value = 0.6087

Evaluamos los resultados:

A un nivel de significancia del 5% no hay suficiente evidencia para rechazar Ho por lo tanto se cumple que los residuos presentan una varianza constante(Homocedasticidad) y se valida el supuesto de homocedasticidad.

***Linealidad***

Definición de hipótesis:

Los residuos NO se comportan de forma lineal

Los residuos se comportan de forma lineal

fm = aov(lm(SCORE~DURACION,data = regresionData) )  
  
summary(fm)

## Df Sum Sq Mean Sq F value Pr(>F)   
## DURACION 1 34.61 34.61 34.19 3.33e-08 \*\*\*  
## Residuals 141 142.76 1.01   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Criterio de rechazo:

Rechazo sii

Evaluamos los resultados: 1.09e-05

se cumple la desigualdad, por lo tanto, a un nivel de significancia del 5% hay suficiente evidencia para rechazar Ho, es decir que los residuos se comportan de forma lineal y se valida el supuesto de linealidad.