**OBJETIVO**

Nuestro objetivo es determinar en qué zona de Europa (Este u Oeste) se escucha más música en la lengua propia de la región, o si en consecuencia son más populares las canciones extranjeras

Esto será posible gracias a que la gran mayoría de países de Europa tiene una playlist en Spotify llamada TOP 50, que recoge las cincuenta canciones más populares del momento.

Además de esto, queremos averiguar donde hay más diversidad lingüística en el ámbito musical comparando el número de idiomas diferentes de las canciones.

## MÉTODO

En este apartado explicaremos cómo hemos llevado a cabo nuestro estudio.

### **Plan de recogida :**

En nuestra recogida de datos, usaremos una web que permite acceder a las listas TOP 50 de Spotify por países y fecha. De todos los países que tienen lista, seleccionaremos solo los que pertenezcan a Europa y los clasificaremos en dos bloques:

* **Europa Este** incluye: Bulgaria, Eslovaquia, Estonia, Hungría, Letonia, Lituania, Polonia, República Checa,  Rumanía, Rusia y  Ucrania.
* **Europa Oeste** incluye: Alemania, Austria, Bélgica, Dinamarca, España, Finlandia, Francia, Grecia, Irlanda, Islandia, Itália, Luxemburgo, Noruega, Países Bajos, Portugal, Reino Unido, Suecia y Suiza.

Por lo tanto, al haber 29 países (11 en Europa Este y 18 en Europa Oeste), habrá como máximo 1450 canciones diferentes. Pero ese no es el valor real debido a que hay canciones que se escuchan en varios países. Por lo que después de analizar las listas y unirlas en dos bloques (Este, Oeste), el número de canciones únicas pasa a ser 608.

Seguidamente escribimos el idioma de cada canción, trabajo que ha sido relativamente sencillo en las canciones en inglés, español, francés o portugués, pero no en esos casos donde hay dudas. Entonces buscamos el nombre de la canción seguido de “lyrics”, y teniendo la letra de la canción usamos la herramienta detectar idioma de Google para saber el idioma de la canción.

Una vez conseguidos todos los idiomas se bifurca en dos ramas. Una para hacer el summary, histogramas y boxplot con RStudio y la otra para conseguir la presencia de los idiomas, con C + +.

Todos los datos pertenecen a las listas TOP 50 del día 20/11/2020, y se han obtenido mediante el siguiente enlace: <https://spotifycharts.com/regional>.

### **RStudio :**

El estudio estadístico lo hemos realizado mediante RStudio, y el procedimiento es el siguiente.

Primero de todo dividimos los datos de ambas zonas en sus respectivos países, para así poderlos analizar con más profundidad. Seguidamente buscamos dentro de cada país las canciones que sí están en el idioma oficial de la región en cuestión.

Además, contamos también la cantidad de lenguas diferentes que hay en cada lista.

Alemania <- data.frame(Oeste[Oeste$V4 == "Alemania", 2]); colnames(Alemania)<- c("Idioma")

siAlemania <- length(Alemania[Alemania$Idioma == "s aleman", ])

diffAlemania = length(unique(Alemania)$Idioma)

Este fragmento de código busca en la tabla *Oeste* los datos que pertenecen a un país (en este caso Alemania) y los asigna a la variable de la región con el header igual a “Idioma”, para así poderlos consultar más fácil posteriormente.

Seguidamente calcula el número de canciones en su idioma y lo almacena en la variable *siAlemania*. Lo mismo sucede con *diffAlemania*, que guarda los diferentes idiomas.

Este proceso se repite para todos los países de las dos zonas (11 en el caso de Europa Este y 18 en Oeste).

Una vez hecho esto, se agrupan los datos de cada región en las siguientes variables:

siEste=c(siBulgaria,siEslovaquia,siEstonia,siHungria,siLetonia,siLituania,

siPolonia,siRepCheca,siRumania,siRusia,siUcrania)

diffEste=c(diffBulgaria,diffEslovaquia,diffEstonia,diffHungria,diffLetonia,

diffLituania,diffPolonia,diffRepCheca,diffRumania,diffRusia,diffUcrania)

A partir de aquí solo queda extraer la información que aportan estos datos mediante las diferentes funciones que incorpora RStudio (mean(), sd(), summary(), var()...) y representarlos con ya sean Boxplots o Histogramas.

### **C++ :**

Para poder obtener los gráficos del *“Porcentaje de aparición de los idiomas según la región”* hemos usado un script en C + + que se puede consultar en el Anexo 2: Script de C + +.

### **Estudio inferencial :**

#### **Prueba de hipótesis :**

La hipótesis que queremos probar es la siguiente: en la Europa Oeste se escucha más música en la lengua oficial de esa región que en la Europa Este.

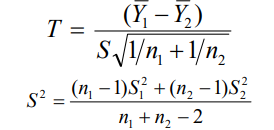
|  |
| --- |
| ***H*0** : *µeste* = *µoeste* |
| ***H*1** : *µeste* < *µoeste* |

En el caso de rechazar ***H*0** estaríamos probando que nuestra suposición inicial es correcta. Se trata de una prueba de hipótesis unilateral debido a que si se rechaza ***H*0**, veremos que la cantidad de canciones en lengua propia de las dos zonas no es igual y entonces, solo tendremos que buscar que ***H*1** sea cierta.

Los datos obtenidos son mayoritariamente independientes (en algunos casos se pueden repetir canciones).

Debido a que ambas variables son discretas, no siguen una distribución normal.

Variables:

* α = 0.05. Con el valor por defecto podremos rechazar o no la hipótesis con una confianza del  95%.
* Para nuestro estadístico usaremos la ‘t’ dado el tamaño de nuestras muestras. Por lo que la distribución será la t-student con n1+n2 - 2 grados de libertad.

https://lh4.googleusercontent.com/FkQXkwkXhOYSV5a04il_hdva-b2zw8mvxtX8MI5sCPQdyIJDNoUvoEfx8LWzD9nEESldr6f7bYwQhMIuazQha1nX7WnODFK_gsmSQh7UNkoTv3lCNqXdckHOwUBuaerh8faCJi7w

Para probar esta hipótesis realizamos un t.test con RStudio (*t.test(siOeste, siEste, alternative = "greater", var.equal = TRUE*) y obtenemos el valor del estadístico, el p-valor y el intervalo de confianza con un riesgo del 5%.

*Two Sample t-test*

*data:  siOeste and siEste*

*t = 3.2219, df = 27, p-value = 0.001656*

*alternative hypothesis: true difference in means is greater than 0*

*95 percent confidence interval:*

*6.746453      Inf*

*sample estimates:*

*mean of x mean of y*

*25.22222  10.90909*

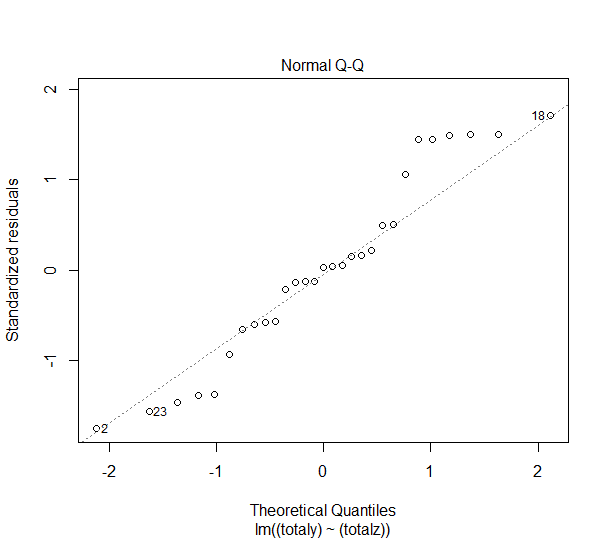
Según nuestros cálculos y los obtenidos mediante RStudio el valor del estadístico es **t** =  3,221941 con 27 grados de libertad.

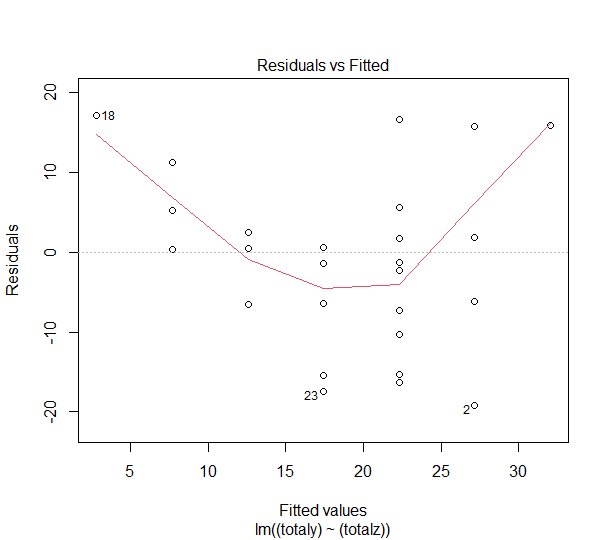
El intervalo de confianza calculado con un 5% de riesgo es **IC** = [6.75, +∞]. Dado el  **p-value** = 0.001656, vemos que es inferior al valor de alpha (0.05), por lo que podemos rechazar la hipótesis ***H0***.

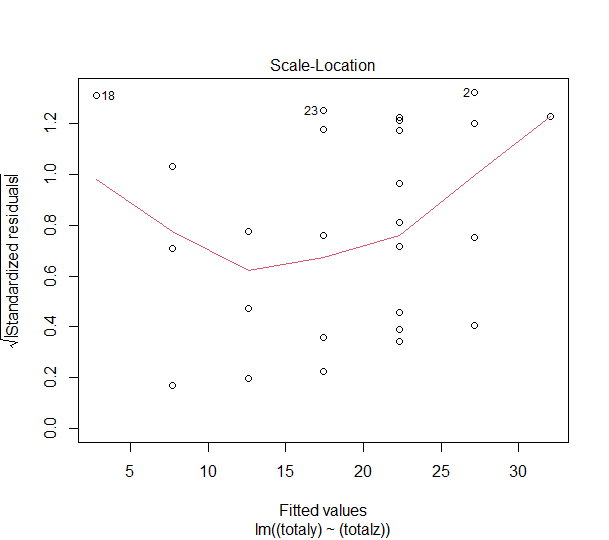
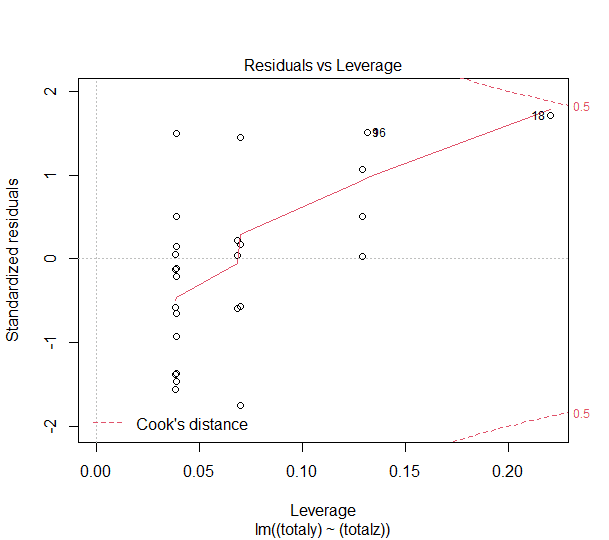
Además, calculando el punto crítico (qt(0.975,27)= 2.051831) vemos que el estadístico está en la zona de rechazo.

Con esto podemos concluir con un 95% de confianza que ***H1*** se cumple, por lo que la cantidad de canciones en la lengua oficial del país es mayor en Europa Oeste que en Europa Este.

#### **Estudio relación cantidad idiomas/canciones :**

 **>** Gráfico A. *Normalidad* **:** **>** Gráfico B. *Homocedasticidad* **:**



 **>** Gráfico C. *Independencia* **:** **>** Gráfico D. *Linealidad* **:**

En el gráfico A podemos ver como las variables no siguen del todo la normalidad debido a que hay cierta dispersión entre los valores y estos no acaban de formar una línea recta.

Del gráfico B y D se puede observar que no se cumple la linealidad ya que los puntos no se mantienen en la misma altura. Además, vemos como la variabilidad de los residuos no se mantiene constante, al principio no se altera demasiado, pero a medida que va avanzando aumenta la dispersión y con ello la variabilidad.

Con el gráfico C vemos que podría haber una cierta tendencia al alza, pero no está demasiado claro, así que no podemos confirmar la independencia.