

Diversidad lingüística en la música de Europa

INTRODUCCIÓN

Somos alumnos de la FIB (Facultad Informática de Barcelona) y nos hemos propuesto hacer un trabajo de análisis sobre las distintas lenguas en las canciones actuales de Europa.

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.

VARIABLES

Las variables escogidas son las siguientes:

- Variable binaria:

Zona de Europa (x), indica a qué zona de Europa pueden pertenecer las canciones. Sus valores son Este y Oeste.

- Variables numéricas (VADs):

Número de canciones propias (y), indica el número de canciones cantadas en la lengua propia del país al cual pertenece la lista.

Número de idiomas únicos (z), es la cantidad de idiomas únicos de las canciones de la lista del país.

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, Italia, 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 unir las 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 + +.

RESULTADOS

Análisis descriptivo :

Aquí se pueden ver los resúmenes de los datos de ambas zonas con los parámetros utilizados, es decir, el número de canciones en la lengua oficial y la cantidad de idiomas diferentes.

> Europa Oeste :

Canciones en lengua oficial								
Min	1sr Qu.	Median	Mean	3rd Qu.	Max	Desviación típica	Varianza	

6	15,25	21	25,22	36,50	48	13,5019	182,3007
---	-------	----	-------	-------	----	---------	----------

2	3	4	4,056	4	8	1,5894	2,5261
---	---	---	-------	---	---	--------	--------

Cantidad de idiomas

> Europa Este :

Canciones en lengua oficial

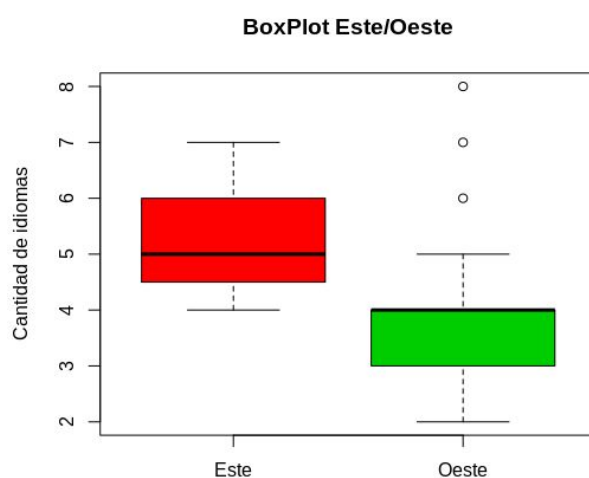
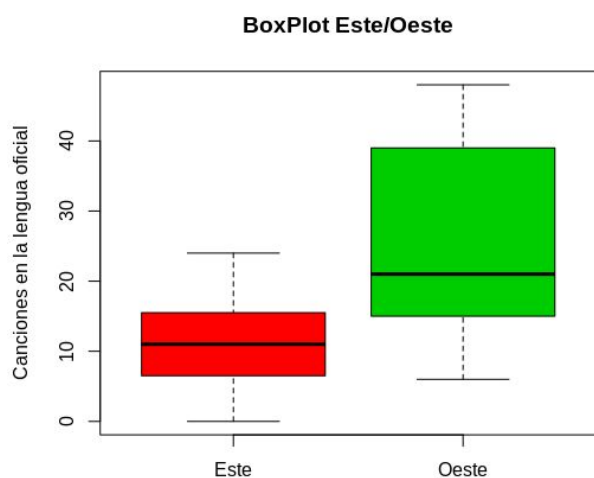
Min	1sr Qu.	Median	Mean	3rd Qu.	Max	Desviación típica	Varianza
0	6,5	11	10,91	15,50	24	7,3410	53,8909

4	4.5	5	5.273	6	7	1,1037	1,2181
---	-----	---	-------	---	---	--------	--------

Cantidad de idiomas

Estos datos, obtenidos mediante las comandas de R `summary()`, `var()` y `sd()`, muestran que la media de canciones en la lengua oficial es mucho mayor en la Europa Oeste que en la Este. Por contra, la cantidad de idiomas diferentes es ligeramente superior en la zona Este.

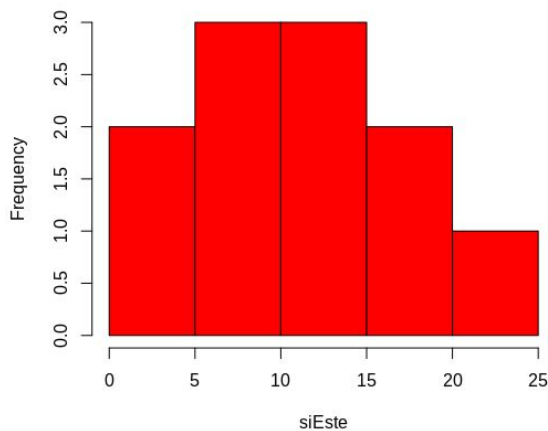
Con los boxplots y los histogramas podemos observar gráficamente estos resultados:



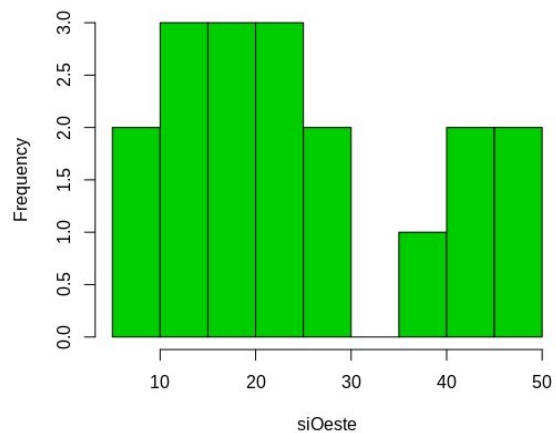
El boxplot de la izquierda compara el número de canciones en la lengua oficial de la zona Este y Oeste.

El de la derecha indica la cantidad de idiomas distintos a los que pertenecen las canciones de las respectivas zonas.

Histograma Canciones lengua oficial Europa Este

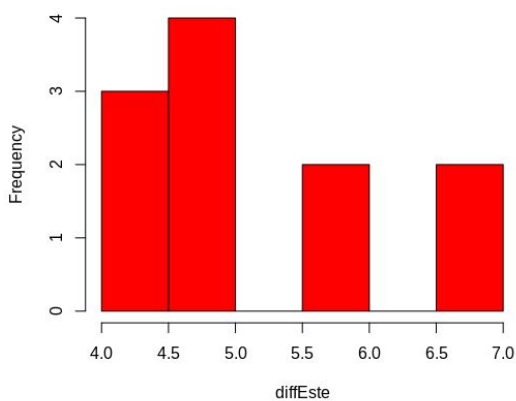


Histograma Canciones lengua oficial Europa Oeste

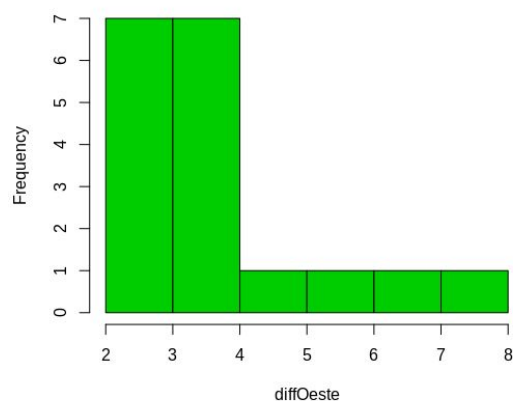


En el caso de los histogramas, podemos observar que en la zona Oeste la frecuencia de 20/30 canciones es muy constante y en cambio, esa misma frecuencia en la Europa Este solo la encontramos al reducir considerablemente la cantidad de canciones.

Histograma Idiomas diferentes Europa Este



Histograma Idiomas diferentes Europa Oeste

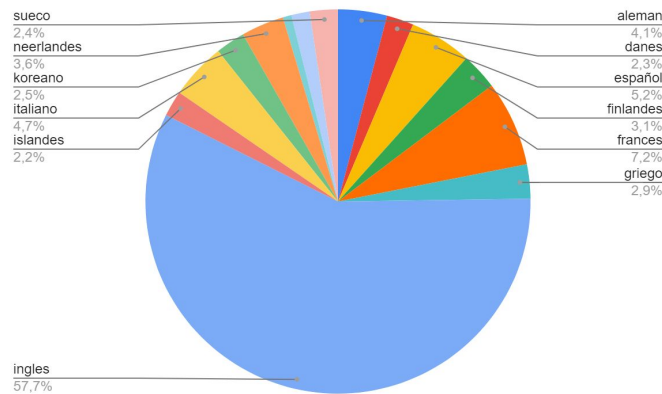


En el histograma de la derecha que representa la presencia de diferentes idiomas en un país del oeste, podemos apreciar la tendencia a tener poca cantidad de idiomas. En cambio en el histograma de la izquierda vemos una tendencia a tener 4 idiomas enfrente a las otras dos posibilidades.

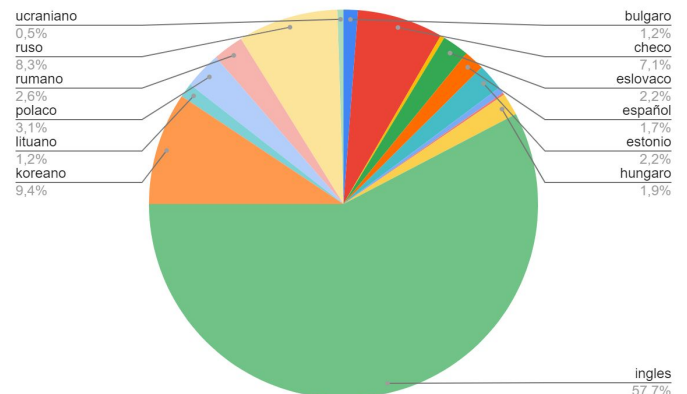
Adicionalmente hemos hecho estos gráficos con un script en C + + para poder ilustrar aún mejor los datos.

Porcentaje de aparición de los idiomas según la región

> Europa Oeste :



> Europa Este :



Como se puede observar, el porcentaje de canciones en inglés es muy superior a los demás (indiferentemente de la zona que se analize).

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.

$$\begin{cases} H_0 : \mu_{este} = \mu_{oeste} \\ H_1 : \mu_{este} < \mu_{oeste} \end{cases}$$

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:

- $\alpha = 0.05$. 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 $n_1 + n_2 - 2$ grados de libertad.

$$T \sim t_{n_1+n_2-2}$$

$$T = \frac{(\bar{Y}_1 - \bar{Y}_2)}{S\sqrt{1/n_1 + 1/n_2}}$$

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

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.

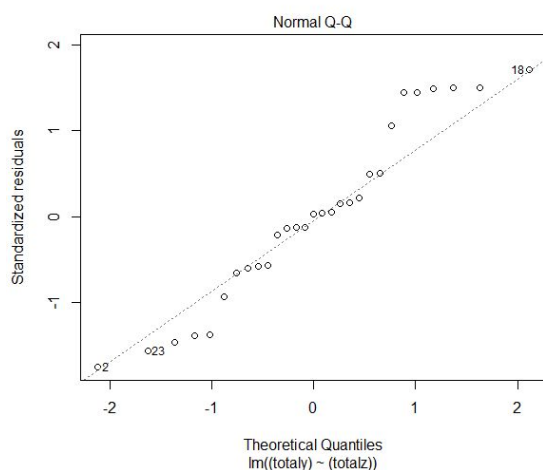
El intervalo de confianza calculado con un 5% de riesgo es **IC** = [6.75, +∞]. Dado el **p-value** = 0.001656, vemos que es inferior al valor de alpha (0.05), por lo que podemos rechazar la hipótesis **H₀**.

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

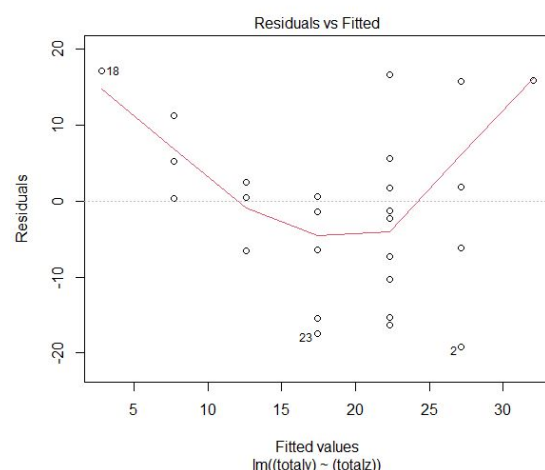
Con esto podemos concluir con un 95% de confianza que **H₁** se cumple, por lo que la cantidad de canciones en la lengua oficial del país es mayor en Europa Oeste que en Europa Este.

Estudio relación cantidad idiomas/canciones :

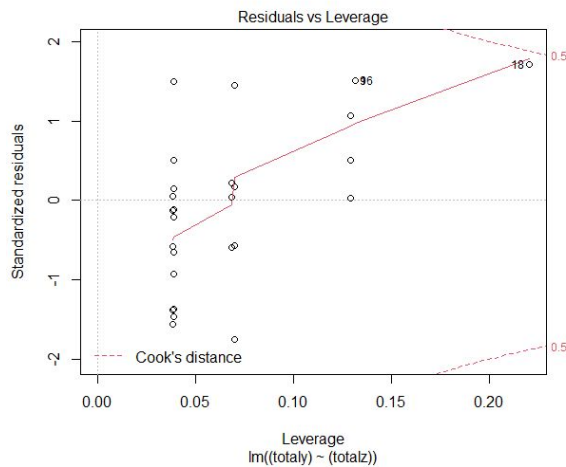
➤ Gráfico A. Normalidad :



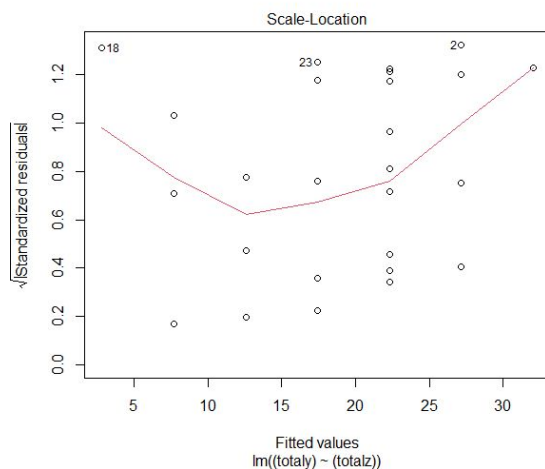
➤ Gráfico B. Homocedasticidad :



> 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.

Mediante RStudio obtenemos los valores de la recta de regresión.

```
Call:
lm(formula = CancionesLenguaOficial ~ CantidadIdiomas)
Coefficients:
  (Intercept)  CantidadIdiomas
       41.804        -4.873
```

Por lo que la función de regresión lineal es la siguiente:

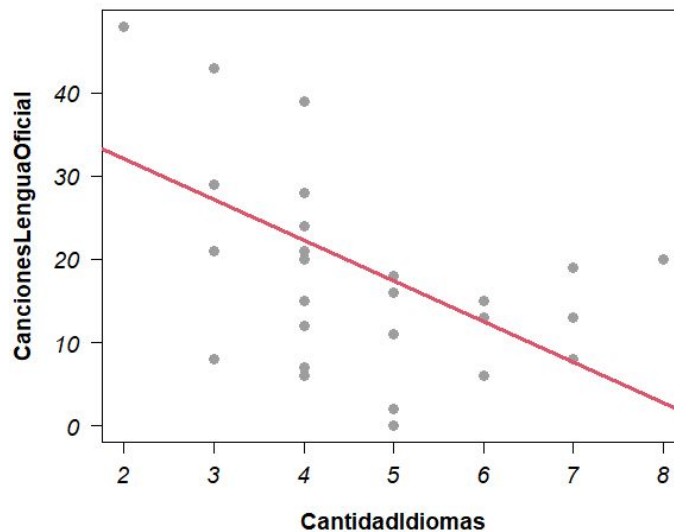
$$Y = 41,804 + (-4,873) * X$$

Además obtenemos también el resumen:

```
Call:
lm(formula = CancionesLenguaOficial ~ CantidadIdiomas)
Residuals:
    Min       1Q   Median       3Q      Max
-19.1860  -6.5682   0.3044   5.6866  17.1771
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	41.804	6.698	6.242	1.12e-06 ***
CantidadIdiomas	-4.873	1.407	-3.463	0.0018 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 11.37 on 27 degrees of freedom
Multiple R-squared: 0.3075, Adjusted R-squared: 0.2819
F-statistic: 11.99 on 1 and 27 DF, p-value: 0.001797



DISCUSIÓN

Con los resultados obtenidos tras hacer el summary de los datos, podemos observar los siguientes casos:

Aun y ser de la misma región, los diferentes países de cada zona no tienen el mismo número de canciones propias ni escuchan la misma cantidad de idiomas diferentes al suyo propio, pero se puede apreciar que en la zona oriental de media se escuchan más idiomas diferentes, en cambio en la zona occidental de media son más populares las canciones en el idioma del propio país.

La cantidad de idiomas escuchados en la zona Oeste es más amplia que en la zona Este.

Aunque la zona oeste tiene mayor cantidad de idiomas de media, como mínimo la zona este tiene 4 idiomas diferentes, mientras que la zona oeste como mínimo tiene 2 idiomas diferentes.

Podemos observar que el inglés es la lengua más escuchada a nivel europeo, esto concuerda con el hecho de que el inglés es la lengua internacional.

También podemos observar que el coreano tiene cierta influencia aún y es el único idioma que tiene varias canciones en el top de los países europeos y no pertenece a Europa.

ANEXO 1: Script de RStudio

```
Este <- read.csv("C:/Users/rguix/Desktop/Este.csv", header=FALSE)
Oeste <- read.csv("C:/Users/rguix/Desktop/Oeste.csv", header=FALSE)

#ANÁLISI DESCRIPTIU

#ESTE

Bulgaria <- data.frame(Este[Este$V3 == "Bulgaria", 2]); colnames(Bulgaria)<- c("Idioma")
siBulgaria <- length(Bulgaria[Bulgaria$Idioma == "s bulgaro", ]);
diffBulgaria = length(unique(Bulgaria)$Idioma)
Rusia <- data.frame(Este[Este$V4 == "Rusia", 2]); colnames(Rusia)<- c("Idioma")
siRusia <- length(Rusia[Rusia$Idioma == "s ruso", ])
diffRusia = length(unique(Rusia)$Idioma)
Ucrania <- data.frame(Este[Este$V5 == "Ucrania", 2]); colnames(Ucrania)<- c("Idioma")
siUcrania <- length(Ucrania[Ucrania$Idioma == "s ucraniano", ])
diffUcrania = length(unique(Ucrania)$Idioma)
Eslovaquia <- data.frame(Este[Este$V6 == "Eslovaquia", 2]); colnames(Eslovaquia)<- c("Idioma")
siEslovaquia <- length(Eslovaquia[Eslovaquia$Idioma == "s eslovaco", ])
diffEslovaquia = length(unique(Eslovaquia)$Idioma)
Rumania <- data.frame(Este[Este$V7 == "Rumania", 2]); colnames(Rumania)<- c("Idioma")
siRumania <- length(Rumania[Rumania$Idioma == "s rumano", ])
diffRumania = length(unique(Rumania)$Idioma)
Hungria <- data.frame(Este[Este$V8 == "Hungria", 2]); colnames(Hungria)<- c("Idioma")
siHungria <- length(Hungria[Hungria$Idioma == "s hungaro", ])
diffHungria = length(unique(Hungria)$Idioma)
Polonia <- data.frame(Este[Este$V9 == "Polonia", 2]); colnames(Polonia)<- c("Idioma")
siPolonia <- length(Polonia[Polonia$Idioma == "s polaco", ])
diffPolonia = length(unique(Polonia)$Idioma)
RepCheca <- data.frame(Este[Este$V10 == "Republica Checa", 2]); colnames(RepCheca)<- c("Idioma")
siRepCheca <- length(RepCheca[RepCheca$Idioma == "s checo", ])
diffRepCheca = length(unique(RepCheca)$Idioma)
Letonia <- data.frame(Este[Este$V11 == "Letonia", 2]); colnames(Letonia)<- c("Idioma")
siLetonia <- length(Letonia[Letonia$Idioma == "s leton", ])
diffLetonia = length(unique(Letonia)$Idioma)
Lituania <- data.frame(Este[Este$V12 == "Lituania", 2]); colnames(Lituania)<- c("Idioma")
siLituania <- length(Lituania[Lituania$Idioma == "s lituano", ])
diffLituania = length(unique(Lituania)$Idioma)
Estonia <- data.frame(Este[Este$V13 == "Estonia", 2]); colnames(Estonia)<- c("Idioma")
siEstonia <- length(Estonia[Estonia$Idioma == "s estonio", ])
diffEstonia = length(unique(Estonia)$Idioma)

#OESTE

Alemania <- data.frame(Oeste[Oeste$V4 == "Alemania", 2]); colnames(Alemania)<- c("Idioma")
siAlemania <- length(Alemania[Alemania$Idioma == "s aleman", ])
diffAlemania = length(unique(Alemania)$Idioma)
Austria <- data.frame(Oeste[Oeste$V5 == "Austria", 2]); colnames(Austria)<- c("Idioma")
siAustria <- length(Austria[Austria$Idioma == "s aleman", ])
diffAustria = length(unique(Austria)$Idioma)
Belgica <- data.frame(Oeste[Oeste$V6 == "Belgica", 2]); colnames(Belgica)<- c("Idioma")
siBelgica <- length(Belgica[Belgica$Idioma == "s frances", ]) + length(Belgica[Belgica$Idioma == "s neerlandes", ]) +
length(Belgica[Belgica$Idioma == "s aleman", ])
diffBelgica = length(unique(Belgica)$Idioma)
Dinamarca <- data.frame(Oeste[Oeste$V7 == "Dinamarca", 2]); colnames(Dinamarca)<- c("Idioma")
siDinamarca <- length(Dinamarca[Dinamarca$Idioma == "s danes", ])
diffDinamarca = length(unique(Dinamarca)$Idioma)
España <- data.frame(Oeste[Oeste$V8 == "esp", 2]); colnames(España)<- c("Idioma")
siEspaña <- length(España[España$Idioma == "s español", ])
diffEspaña = length(unique(España)$Idioma)
Finlandia <- data.frame(Oeste[Oeste$V9 == "Finlandia", 2]); colnames(Finlandia)<- c("Idioma")
```

```

siFinlandia <- length(Finlandia[Finlandia$Idioma == "s finlandes", ])
diffFinlandia = length(unique(Finlandia)$Idioma)
Francia <- data.frame(Oeste[Oeste$V10 == "Francia", 2]); colnames(Francia)<- c("Idioma")
siFrancia <- length(Francia[Francia$Idioma == "s frances", ])
diffFrancia = length(unique(Francia)$Idioma)
Grecia <- data.frame(Oeste[Oeste$V11 == "Grecia", 2]); colnames(Grecia)<- c("Idioma")
siGrecia <- length(Grecia[Grecia$Idioma == "s griego", ])
diffGrecia = length(unique(Grecia)$Idioma)
Irlanda <- data.frame(Oeste[Oeste$V12 == "Irlanda", 2]); colnames(Irlanda)<- c("Idioma")
siIrlanda <- length(Irlanda[Irlanda$Idioma == "s ingles", ])
diffIrlanda = length(unique(Irlanda)$Idioma)
Islandia <- data.frame(Oeste[Oeste$V13 == "Islandia", 2]); colnames(Islandia)<- c("Idioma")
siIslandia <- length(Islandia[Islandia$Idioma == "s islandes", ])
diffIslandia = length(unique(Islandia)$Idioma)
Italia <- data.frame(Oeste[Oeste$V14 == "Italia", 2]); colnames(Italia)<- c("Idioma")
siItalia <- length(Italia[Italia$Idioma == "s italiano", ])
diffItalia = length(unique(Italia)$Idioma)
Luxemburgo <- data.frame(Oeste[Oeste$V20 == "Luxemburgo", 2]); colnames(Luxemburgo)<- c("Idioma")
siLuxemburgo <- length(Luxemburgo[Luxemburgo$Idioma == "s aleman", ]) + length(Luxemburgo[Luxemburgo$Idioma == "s frances", ])
diffLuxemburgo = length(unique(Luxemburgo)$Idioma)
Noruega <- data.frame(Oeste[Oeste$V15 == "Noruega", 2]); colnames(Noruega)<- c("Idioma")
siNoruega <- length(Noruega[Noruega$Idioma == "s noruego", ])
diffNoruega = length(unique(Noruega)$Idioma)
PaísesBajos <- data.frame(Oeste[Oeste$V16 == "Países Bajos", 2]); colnames(PaísesBajos)<- c("Idioma")
siPaísesBajos <- length(PaísesBajos[PaísesBajos$Idioma == "s neerlandes", ])
diffPaísesBajos = length(unique(PaísesBajos)$Idioma)
Portugal <- data.frame(Oeste[Oeste$V3 == "Portugal", 2]); colnames(Portugal)<- c("Idioma")
siPortugal <- length(Portugal[Portugal$Idioma == "s portugueses", ])
diffPortugal = length(unique(Portugal)$Idioma)
Reino Unido <- data.frame(Oeste[Oeste$V17 == "Reino Unido", 2]); colnames(Reino Unido)<- c("Idioma")
siReino Unido <- length(Reino Unido[Reino Unido$Idioma == "s ingles", ])
diffReino Unido = length(unique(Reino Unido)$Idioma)
Suecia <- data.frame(Oeste[Oeste$V18 == "Suecia", 2]); colnames(Suecia)<- c("Idioma")
siSuecia <- length(Suecia[Suecia$Idioma == "s sueco", ])
diffSuecia = length(unique(Suecia)$Idioma)
Suiza <- data.frame(Oeste[Oeste$V19 == "Suiza", 2]); colnames(Suiza)<- c("Idioma")
siSuiza <- length(Suiza[Suiza$Idioma == "s aleman", ]) + length(Suiza[Suiza$Idioma == "s frances", ]) +
length(Suiza[Suiza$Idioma == "s italiano", ])
diffSuiza = length(unique(Suiza)$Idioma)
#RESULTADOS

#ESTE

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)

summary(siEste); sd(siEste) #desviación típica
summary(diffEste); sd(diffEste)

hist(siEste, main = "Histograma Canciones lengua oficial Europa Este", col = 2)
hist(diffEste, main = "Histograma Idiomas diferentes Europa Este", col = 2)

#OESTE

siOeste=c(siAlemania, siAustria, siBelgica, siDinamarca, siEspaña, siFinlandia, siFrancia, siGrecia, siIrlanda, siIslandia, siItalia, siLuxemburgo, siNoruega, siPaísesBajos, siPortugal, siReino Unido, siSuecia, siSuiza)

diffOeste=c(diffAlemania, diffAustria, diffBelgica, diffDinamarca, diffEspaña, diffFinlandia, diffFrancia, diffGrecia, diffIrlanda, diffIslandia, diffItalia, diffLuxemburgo, diffNoruega, diffPaísesBajos, diffPortugal, diffReino Unido, diffSuecia, diffSuiza)

summary(siOeste); sd(siOeste) #desviación típica
summary(diffOeste); sd(diffOeste)

hist(siOeste, main = "Histograma Canciones lengua oficial Europa Oeste", col = 3)
hist(diffOeste, main = "Histograma Idiomas diferentes Europa Oeste", col = 3)

#ESTE vs OESTE

boxplot(siEste, siOeste, main = "BoxPlot Este/Oeste", col = 2:3, names = c("Este", "Oeste"), ylab = "Canciones en la lengua oficial")

boxplot(diffEste, diffOeste, main = "BoxPlot Este/Oeste", col = 2:3, names = c("Este", "Oeste"), ylab = "Cantidad de idiomas")

#ANÁLISI INFERENCIAL
n1 = length(siOeste); n1
n2 = length(siEste); n2
sd1 = sd(siOeste); sd1

```

```

sd2 = sd(siEste);sd2
var1 = var(siOeste);var1
var2 = var(siEste);var2

s2 = (((n1-1)*sd1^2)+((n2-1)*sd2^2))/(n1+n2-2);s2
s = sqrt(s2);s
se = s*sqrt((1/n1)+(1/n2));se
t = (mean(siOeste)-mean(siEste))/se;t

#T.TEST

t.test(siOeste, siEste, alternative = "greater", var.equal = TRUE)
qt(0.975,27)

#RELACIÓN CANTIDAD IDIOMAS/CANCIONES

CancionesLenguaOficial = c(siOeste, siEste)
CantidadIdiomas = c(diffOeste, diffEste)
n = length(CancionesLenguaOficial)
muy = mean(CancionesLenguaOficial)
muz = mean(CantidadIdiomas)
sdy = sd(CancionesLenguaOficial)
sdz = sd(CantidadIdiomas)
syz=((sum(CancionesLenguaOficial*CantidadIdiomas)-sum(CancionesLenguaOficial)*sum(CantidadIdiomas))/n)/(n-1)
r = syz/(sdy*sdz);r
b1 = syz/sdy^2;b1
b0 = muy-b1*muz;b0
lm(CancionesLenguaOficial~CantidadIdiomas)
mod = lm(CancionesLenguaOficial~CantidadIdiomas)
summary(mod)

par(cex.lab=1.2,cex.axis=1.2,las=1,font.lab=2,font.axis=3)
plot(CancionesLenguaOficial~CantidadIdiomas,pch=19,col=8,cex=1.2)
abline(mod,col=2,lwd=3)

```

ANEXO 2: Script de C++

```

#include <iostream>
#include <map>
#include <string>
using namespace std;
int main(){
    string idioma;
    string original;
    map<string, int> total;
    map<string, int> si_no;
    map<string, int>:: iterator it;
    while(cin >> original and original != "end" ){
        it = si_no.find(original);
        if(it == si_no.end()) si_no[original] = 1;
        else it->second +=1; cin >> idioma; it = total.find(idioma);
        if(it == total.end()){
            total[idioma] = 0;
            it = total.find(idioma);
        }
        string res; while(cin>>res and res != "final") it->second +=1;
    }
    it = si_no.begin();
    while(it !=si_no.end()){
        cout << it->first << " " << it->second;
        ++it;
    }
    it = total.begin();
    while(it != total.end()){ cout << it->first << " " << it->second; ++it;}
}

```