jmoralesfucPRA2

jose morales

26 December 2018

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El Dataset escogido para esta practica es Global Terrrorism Dataset , descargado de Kaggle en la dirección https://www.kaggle.com/START-UMD/gtd (https://www.kaggle.com/START-UMD/gtd). Este dataset representa los actos terroristas acaecidos entre 1970 y 2017 , considero que es importante porque nos permitirá visualizar la evolución de estos hechos , la evolución en el tipo de actos así como la tendencia actual, asi como demostrar o no que el miedo al terrorismo en los paises occidentales esta sobredimensionado, espeero demostrar que tanto en numero ce incidentes como en numero de muertes los paises occidentales son mucho mas seguros. Este Dataset contiene 135 columnas y 181691 entradas. De todas estas variables eliminaremos la gran mayoría ya que muchas de ellas son redundantes o carecen de interés para nuestro estudio.

2. Integración y selección de los datos de interés a analizar.

No mostrare un head del dataframe hasta realizar una eliminación del número de columnas ya que sería poco útil. Enumerare las columnas que utilizaremos para nuestro estudio:

Iyear, country, country_txt,region , region_txt, attactcktipe1,attacttype1_ txt, ,nkill,nwound

Creamos un Nuevo dataset gtd2 con la function select dentro de la librería dplyr Otro detalle que se observa en este dataset es que no hay datos para el anyo 1993,aunque desconocemos el porque de esta ausencia.

```
library( dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
gtd2=select(gtd,iyear,country,country_txt,region,region_txt,attacktype1,attacktype1_txt,nkill,nwound)
```

Mostramos ahora un head y un summary de nuestro nuevo dataset.

```
head(gtd2)
```

```
##   iyear country        country_txt region                 region_txt
## 1  1970      58 Dominican Republic      2 Central America & Caribbean
## 2  1970     130             Mexico      1               North America
## 3  1970     160        Philippines      5             Southeast Asia
## 4  1970      78             Greece      8             Western Europe
## 5  1970     101              Japan      4                  East Asia
## 6  1970     217      United States      1               North America
##   attacktype1                attacktype1_txt nkill nwound
## 1           1                  Assassination     1      0
## 2           6    Hostage Taking (Kidnapping)     0      0
## 3           1                  Assassination     1      0
## 4           3              Bombing/Explosion    NA     NA
## 5           7 Facility/Infrastructure Attack    NA     NA
## 6           2                  Armed Assault     0      0
```

```
summary(gtd2)
```

```
##      iyear         country          country_txt         region
##  Min.   :1970   Min.   :   4   Iraq       : 24636   Min.   : 1.000
##  1st Qu.:1991   1st Qu.:  78   Pakistan   : 14368   1st Qu.: 5.000
##  Median :2009   Median :  98   Afghanistan: 12731   Median : 6.000
##  Mean   :2003   Mean   : 132   India      : 11960   Mean   : 7.161
##  3rd Qu.:2014   3rd Qu.: 160   Colombia   :  8306   3rd Qu.:10.000
##  Max.   :2017   Max.   :1004   Philippines:  6908   Max.   :12.000
##                                (Other)    :102782
##                       region_txt     attacktype1
##  Middle East & North Africa:50474   Min.   :1.000
##  South Asia                :44974   1st Qu.:2.000
##  South America             :18978   Median :3.000
##  Sub-Saharan Africa        :17550   Mean   :3.248
##  Western Europe            :16639   3rd Qu.:3.000
##  Southeast Asia            :12485   Max.   :9.000
##  (Other)                   :20591
##                        attacktype1_txt      nkill
##  Bombing/Explosion             :88255   Min.   :   0.000
##  Armed Assault                 :42669   1st Qu.:   0.000
##  Assassination                 :19312   Median :   0.000
##  Hostage Taking (Kidnapping)   :11158   Mean   :   2.403
##  Facility/Infrastructure Attack:10356   3rd Qu.:   2.000
##  Unknown                       : 7276   Max.   :1570.000
##  (Other)                       : 2665   NA's   :10313
##      nwound
##  Min.   :   0.000
##  1st Qu.:   0.000
##  Median :   0.000
##  Mean   :   3.168
##  3rd Qu.:   2.000
##  Max.   :8191.000
##  NA's   :16311
```

y por utlimo , revisamos las assignaciones de tipo realizadas por R para ello utilizamos el comando :

```
sapply(gtd2,class)
```

```
##           iyear         country     country_txt          region
##       "integer"       "integer"        "factor"       "integer"
##      region_txt     attacktype1 attacktype1_txt           nkill
##        "factor"       "integer"        "factor"       "integer"
##          nwound
##       "numeric"
```

3. Limpieza de los datos.

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

Pasamos ahora a verificar los elementos , vacios o nulos en nuesto dataset. Los valores de cero por definicion de nuestros datos son posibles asi que no se trataran

```
sapply(gtd2, function(x) sum(is.na(x)))
```

```
##           iyear         country     country_txt          region
##               0               0               0               0
##      region_txt     attacktype1 attacktype1_txt           nkill
##               0               0               0           10313
##          nwound
##           16311
```

Observamos que tenemos na en las variables , numero de muertos y numero de heridos, trataremos cada uno de ellos de la siguiente manera: Las variables nkill y nwound que corresponden al numero de muertos y numero de heridos,en el incidente.Corresponden a hechos de los que no se tiene conocimiento de esos numeros por que no han sido reportados o por que se han perdido , en este caso se puede optar por dos estrategias, la primera seria la eliminacion de los datos inconpletos o una segunda en la que se sustituye este dato. Para optar por la primera estrategia debemos evaluar si esta eliminacion de datos afecta a nuestra muestra, teniendo en cuenta que el tamanyo de la muestra es de 181691 observaciones , elegimos aplicar esta tecnica para el numero de heridos(nwounds) , ya que una gran parte de los datos que se observa este valor corresponde a secuestros . En el caso de nkill utilizaremos una tecnica de sustitucion de valores,para esta sustitucion queriamos utilizar la funcion kNN, que se basa en la imputacion de valores por la cercania de los vecinos,pero dado el tamanyo del dataset obtenemos un error de memoria , se podria escoger los datos cercanos a cada ausencia pero dado que tenemos 10313 na , se deberian de seleccionar el mismo numero de vecinos, por ello considero que la mejor opcion en estos casos es la de sustitucion , en este caso utilizando un estidistico robusto como es la mediana.

Pasamos ahora a ejecutar los cambios descritos anteriormente.

Aplicaremos la sustitucion por la mediana en nkill, para ello primero la obtenemos con el comando median e indicando que existen na

```
med=median (gtd2$nkill,na.rm=TRUE)
temp=which(is.na(gtd2$nkill))


gtd2$nkill[temp]=med
```

Y por ultimo eliminaremos los datos incompletos de nwound

```
gtd2=na.omit(gtd2)
```

Una vez realizado esto revisamos que en niuestro dataset no hayvalores con na

```
sapply(gtd2, function(x) sum(is.na(x)))
```

```
##           iyear         country     country_txt          region
##               0               0               0               0
##      region_txt     attacktype1 attacktype1_txt           nkill
##               0               0               0               0
##          nwound
##               0
```

3.2. Identificación y tratamiento de valores extremos.

Pasamos ahora a realizar el tratamiento de valores extremos. para ellos utilizaremos las propiedades del boxplot, antes de ello haremos un pequenyo resumen de los datos numericos en este caso nkill y nwound

```
summary(gtd2$nkill)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    0.000    0.000    2.124    2.000 1384.000
```

```
summary(gtd2$nwound)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    0.000    0.000    3.168    2.000 8191.000
```

Dado que tenemos muchos valores con valor cero obtenemos una distribucion de nuestros datos biasados a la derecha. Aunque encontremos estos valores extremos en estas variables no los trataremos ya que estas variables no forman parte directa del estudio que pretendemos realizar.

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

Tal y como comentaba en el primer apartado , la idea de tratamiento de este dataset es la de demostrar o no que los incidentes de terrorismo en el mundo desarrollado son menores a los que se producen en paises en desarrollo o subdesarrollados.Por ello vamos a generar un conjunto de nuevas variables que nos permitiran realizar el estudio. La primera variable sera la frecuencia de inicidentes por pais, para poder acceder facilmente a los valores la convertimos en un dataframe.

```
incipercountry=as.data.frame(table(gtd2$country_txt))
```

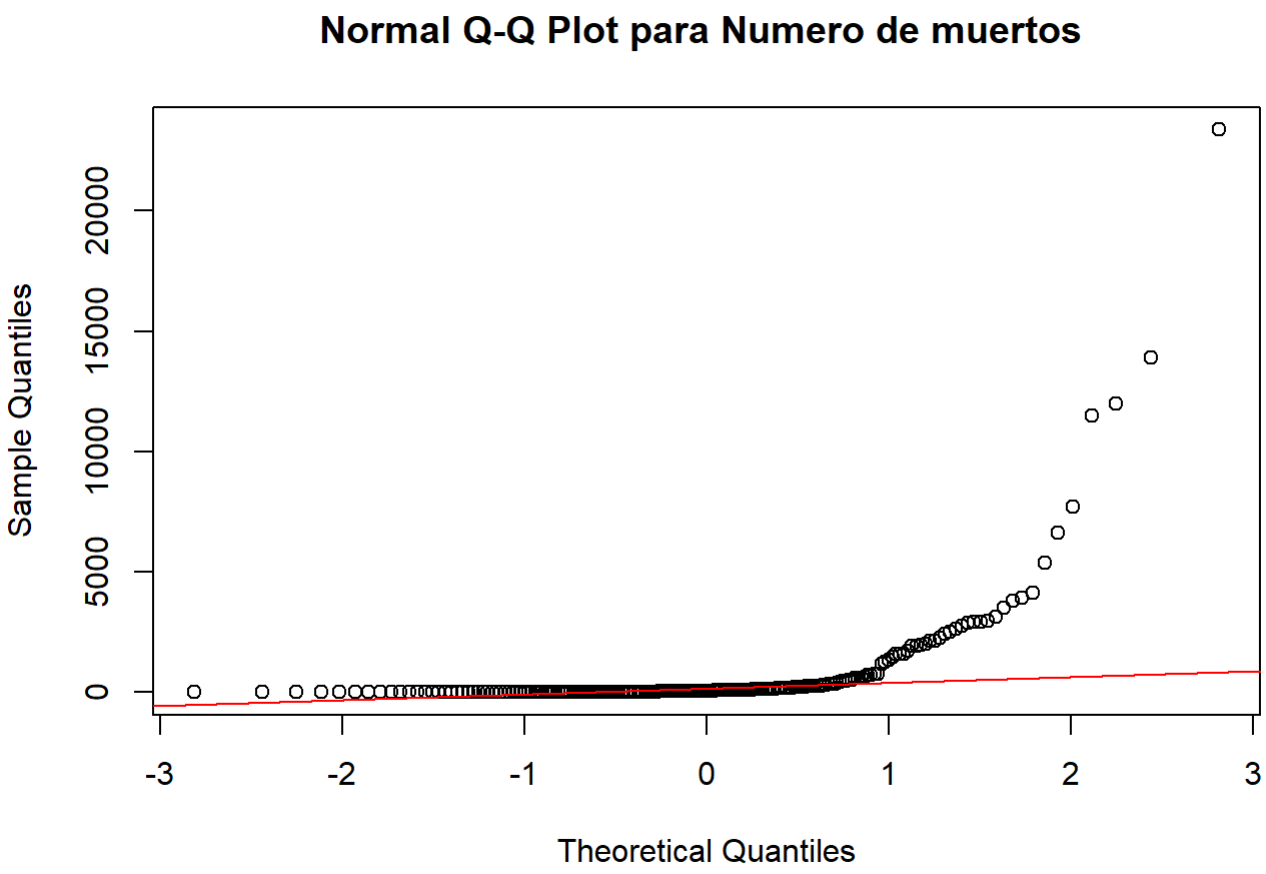
En segundo lugar una nueva variable basada en region en la que separaremos los 3 tipos de paises, haremos una categorizacion bastante gruesa , ya que de 12 regiones pasaremos a tres tipos La primera categoria sera de pasis desarrollados que corresponden a la region de norteamerica, westerneurope,eastasia y australasia. la segunda categoria de paise en desarrollo corresponde a central america, sudamerica, south asia, south east asia y eastern europe. Por ultimo la tercera categoria de paises subdesarrollados lo conpondran las regiones de centralasia, middle est,sub saharaian africa. Tambien se establecera un ranking de los paises con menos y mas atentados.

Esta variable se creara una vez tratada totalmente la variable frecuencia de incidentes

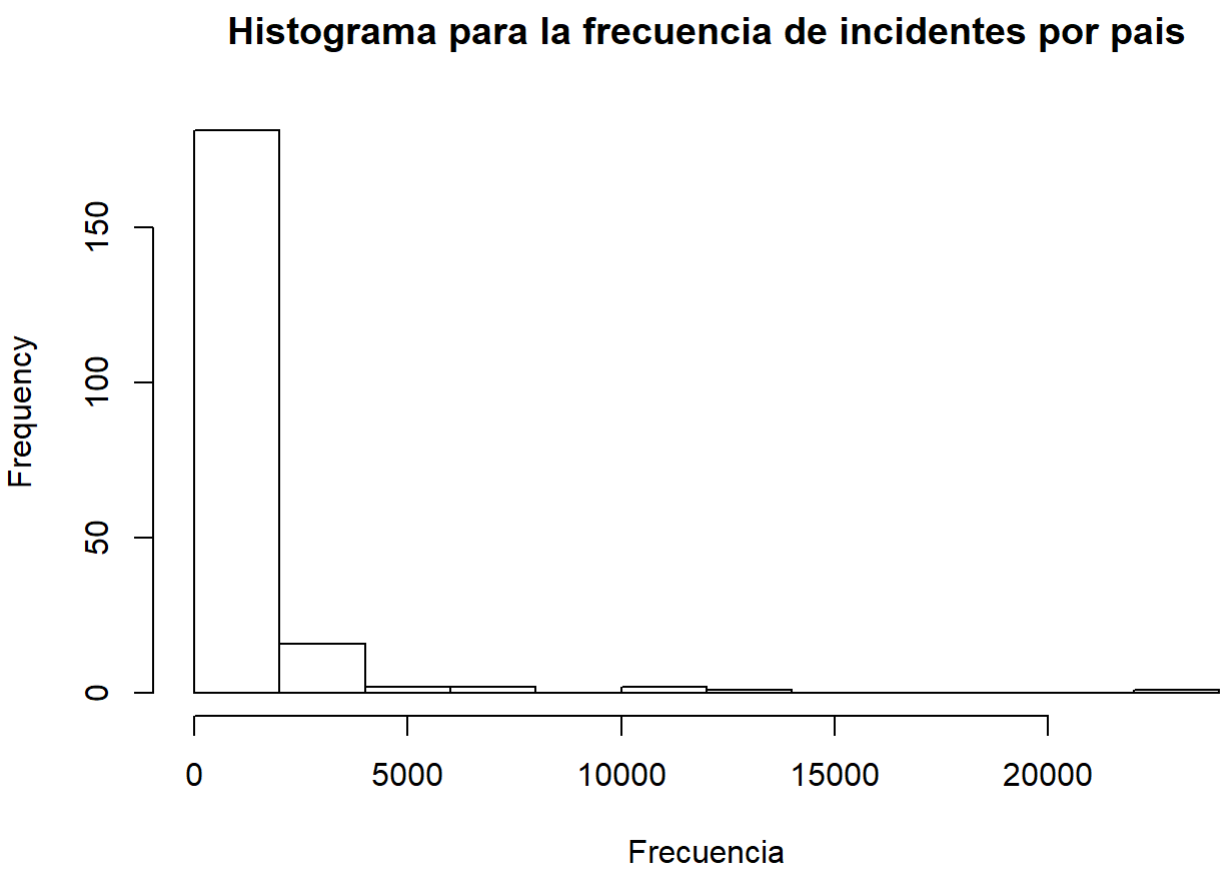
4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Pasamos a comprobar si las variables pueden ser candidatas a la normalización para ello generamos las graficas de quantile-quantile plot y el histograma de las variables.

```
qqnorm(incipercountry$Freq,main = paste("Normal Q-Q Plot para Numero de muertos "))
qqline(incipercountry$Freq,col="red")
```




```
hist(incipercountry$Freq,main=paste("Histograma para la frecuencia de incidentes por pais"),xlab="Frecu
encia",freq = TRUE)
```



Los resultados del quantile-quantile plot nos indican que la variable frequencia de incidentes no es candidata a la normalización.

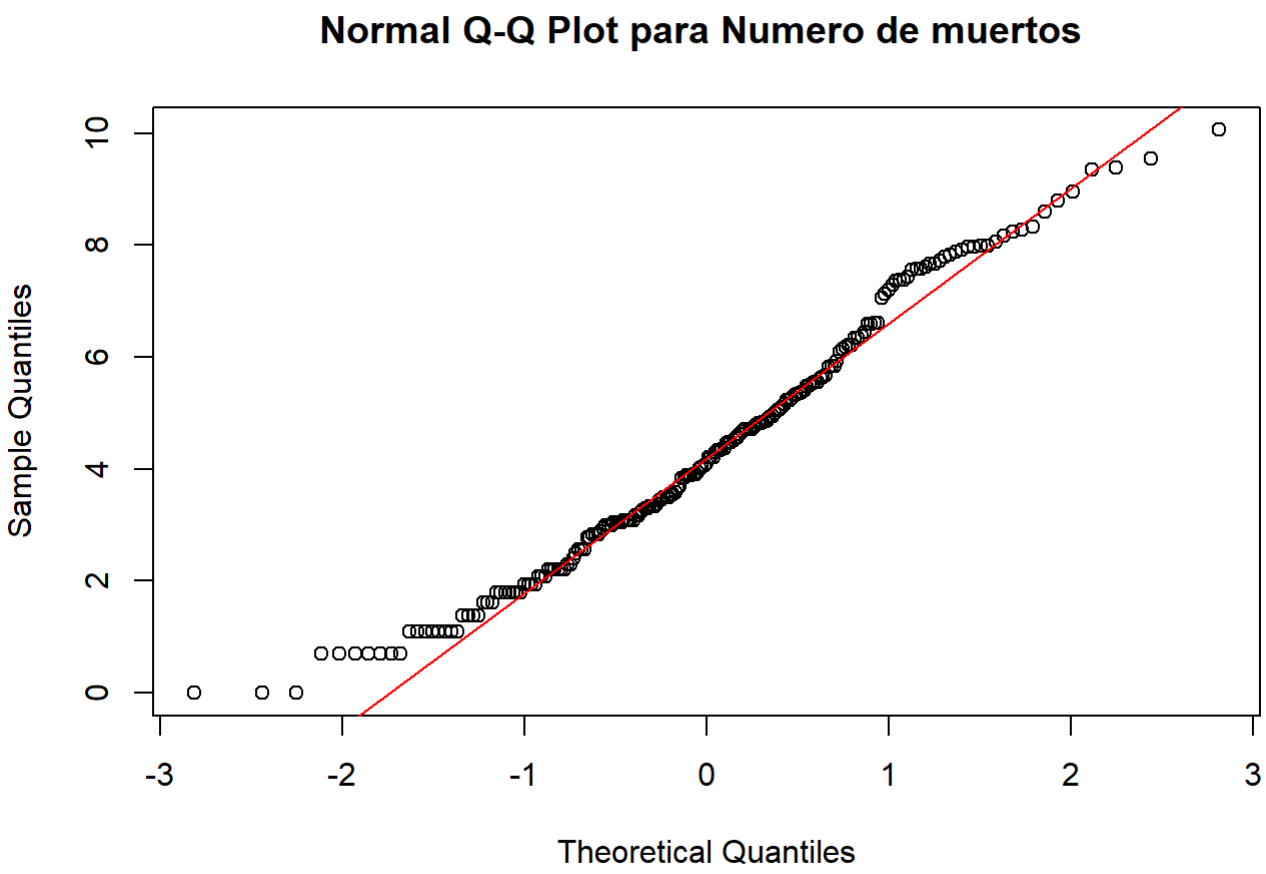
No podemos revisar si las variables estan normalizadas aplicando el test de Shapiro Wilk ya que este esta limitado a 5000 componentes.

Para hacer que nuestra variable sea normal, calcularemos el logaritmo de los datos , ya que tenemos datos con valor cero le sumaremos 1 antes de aplicar el logaritmo, evitando los alores menos infinito

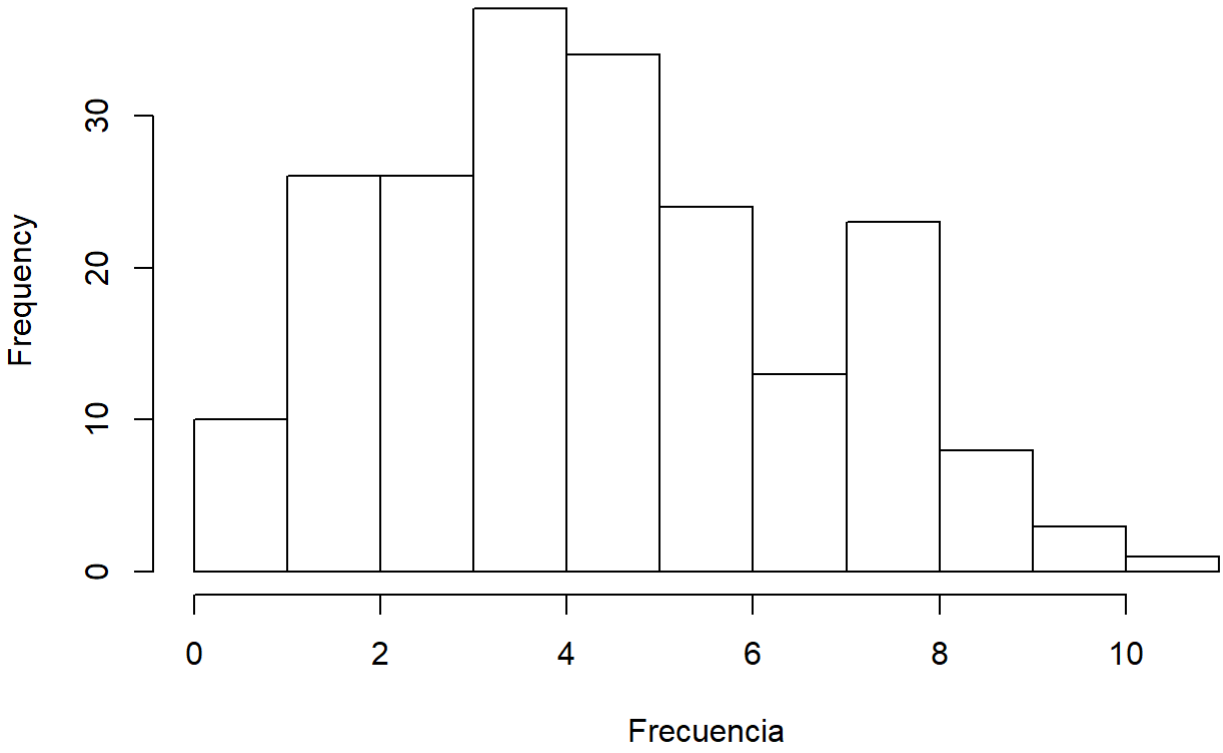
```
logfreq=log(incipercountry$Freq+1)
```

Aplicaremos ahora el test de quantile-quantile plot para los nuevos valores

```
qqnorm(logfreq,main = paste("Normal Q-Q Plot para Numero de muertos "))
qqline(logfreq,col="red")
```




```
hist(logfreq,main=paste("Histograma para el logaritmo dela frecuencia de incidentes por pais"),xlab="Fr
ecuencia",freq = TRUE)
```



Viendo estos datos si podemos considerar que el log de los datos se conporta de manera normal.

Sobre esta variable si podemos aplicar el test de Shapiro Wilk ya que tiene un numero de elementos

```
shapiro.test(logfreq)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  logfreq
## W = 0.9763, p-value = 0.001545
```

Como el p-value es menor que 0.05 consideramos que no es normal.

Ahora crearemos la variable categ, para ello primero asignaremos a cada pais su region , y a partir de las regiones se recategorizaran en desarrollados, en desarrollo o subdesarrollados.

```
#install.packages(BBmisc)
library(BBmisc)
```

```
## Warning: package 'BBmisc' was built under R version 3.5.2
```

```
##
## Attaching package: 'BBmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     coalesce, collapse
```

```
## The following object is masked from 'package:base':
##
##     isFALSE
```

```
regio=as.character(incipercountry$Var1)
n=nrow(incipercountry)
for (j in 1:n){

  cou=as.character(incipercountry$Var1[j])
  cou2=which.first(gtd$country_txt==cou)
  regio[j]=as.numeric(gtd$region[cou2])
  }
```

en regio tenemos las regiones alos que pertenecen los difrentes paises crearemos la variable categ y anyadiremos al datset con un cbind. Recordemos que la clasificacion se hara con el siguiente criterio: La primera categoria sera de pasis desarrollados que corresponden a la region de norteamerica, westerneurope,eastasia y australasia. la segunda categoria de paise en desarrollo corresponde a central america, sudamerica, south asia, south east asia y eastern europe. Por ultimo la tercera categoria de paises subdesarrollados lo conpondran las regiones de centralasia, middle est,sub saharaian africa.

```
categ=regio


m=nrow(incipercountry)
for (s in (1:m)){
  if (regio[s]==1){categ[s]=1
  next}
  if (regio[s]==4){categ[s]=1
  next}
  if (regio[s]==8){categ[s]=1
  next}
  if (regio[s]==12){categ[s]=1
  next}
  if (regio[s]==2){categ[s]=2
  next}
  if (regio[s]==3){categ[s]=2
  next}
  if (regio[s]==5){categ[s]=2
  next}
  if (regio[s]==6){categ[s]=2
  next}
  if (regio[s]==9){categ[s]=2
  next}
  if (regio[s]==7){categ[s]=3
  next}
  if (regio[s]==10){categ[s]=3
  next}
  if (regio[s]==11){categ[s]=3
  next}


}

gtd3=cbind(incipercountry,categ,regio,logfreq)
```

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc.

Pasamos ahora a comprobar nuestras especulaciones, en primer lugar revisaremos la distribucion de los incidentes por paises con la categoria que se establecieron el punto 1 de este ejercicio.

Utilizaremos un contraste de hipotesis para ello enunciaremos la hipotesis nula como que los paises desarrollados tienen el mismo nivel de terrorismo que el resto de los paises . Y definiremos la hipotesis alternativa como que los paises desarrollados tienen menos riesgo que el resto de paises.

Seleccionamos los incidentes de los paises desarrollados y por otro lado el del resto de paises.

```
desarro=gtd3$logfreq[gtd3$categ==1]
nodesarro=gtd3$logfreq[gtd3$categ==3]
```

Suponemos que tiene varianzas iguales y que siguen distribuciones normales

$ t_{0}=) $ donde $ $ y $ $ son las medias muestrales $ n_{1} $ y $ n_{2} $ son el tamaño de cada muestra y $ s_{p}= $

Se compara el valor de este estadistico con el valor de una distribucion t de Student

$ t_{;n_{1}+n_{2}-2 } $ Si $ |t_{0} |>t_{;n_{1}+n_{2}-2 } $ se rechaza $ H_{0} $

```
desvdesa=sd(desarro)
desvnodesa=sd(nodesarro)
ndesa=length(desarro)
nnodesa=length(nodesarro)
medidesa=mean(desarro)
medianodesa=mean(nodesarro)
```

Calculamos S_p

```
subp=((((ndesa-1)*desvdesa))+((nnodesa-1)*(desvnodesa)))/(ndesa+nnodesa-2)
# y t0

t0=(medidesa-medianodesa)/subp*(sqrt(1/ndesa +1/nnodesa))
t0
```

```
## [1] -0.02919568
```

$t_{0} $= -0.01821381 En las tablas consultamos $t_{;n_{1}+n_{2}-2 } $ obteniendo 1,960 Por lo que no podemos rechazar la hipotesis nula ya que el valor absoluto de t_0 es menor que $t_{\frac{\alpha}{2};n_1+n_2-2} = 1,960$

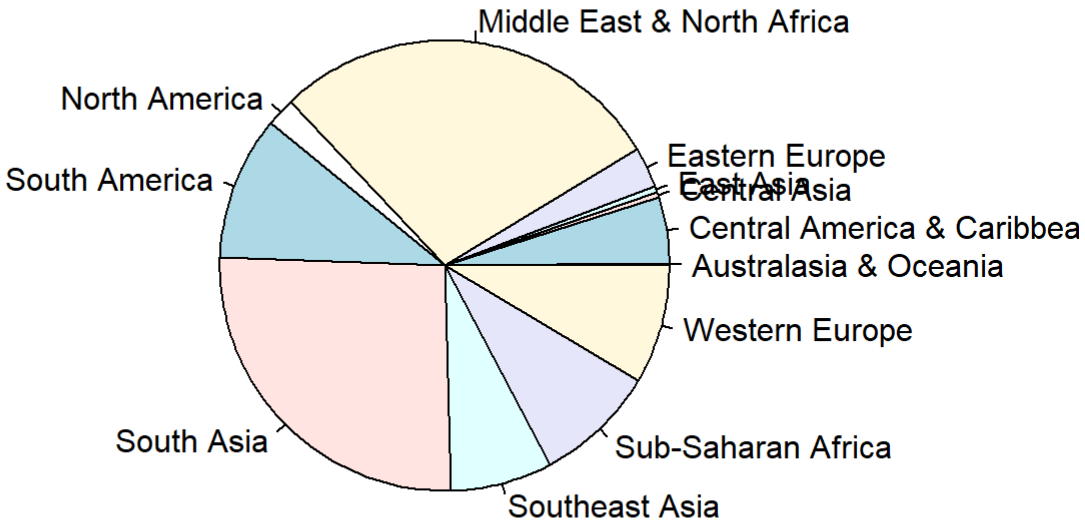
5. Representación de los resultados a partir de tablas y gráficas.

Representaremos ahora algunos de los datos obtenidos.

```
table(gtd2$region_txt)
```

```
##
##       Australasia & Oceania Central America & Caribbean
##                         274                         7874
##                Central Asia                    East Asia
##                         555                          757
##               Eastern Europe  Middle East & North Africa
##                         4892                        46918
##                North America                South America
##                         3335                        17103
##                   South Asia               Southeast Asia
##                        43082                        12023
##           Sub-Saharan Africa               Western Europe
##                        14438                        14129
```

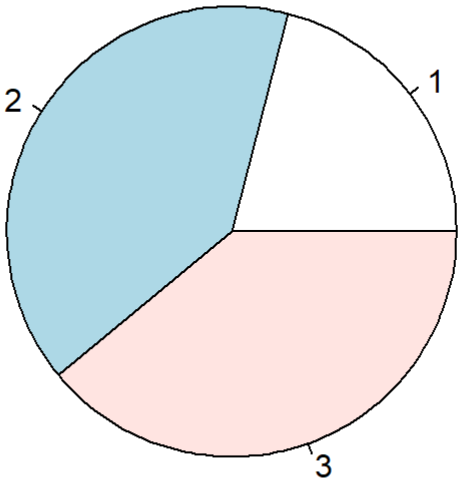
```
pie(table(gtd2$region_txt))
```




```
table(gtd3$categ)
```

```
##
##  1  2  3
## 43 82 80
```

```
pie(table(gtd3$categ))
```

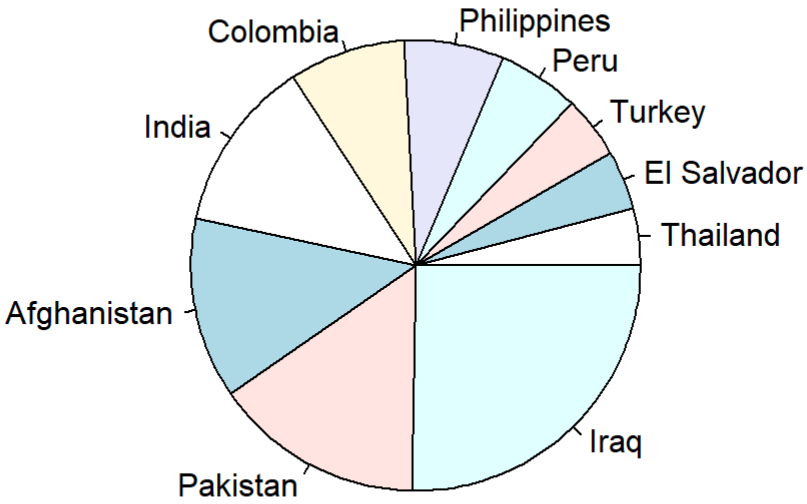


En el primer grafico y tabla se obserban los incidentes por region sin agregar, ya con este grafico podemos obserbar que el numero de incidentes en las regiones de sud asia y medio este ocupan la mayoria de los casos, con el segundo grafico vemos que el conjunto de los paises desarrollados es claramente inferior al resto de paises calculando el porcentaje obtenemos que solo el 11% de los incidentes ocurren en paises desarrollados. Pasamos a mostrar un grafico por paises para ver que paises tienen mas incidentes

```
t=table(gtd2$country_txt)
t=sort(t)
h=head(t,10)
t=tail(t,10)
t
```

```
##
##    Thailand El Salvador      Turkey        Peru Philippines    Colombia
##        3808        3901        4133        5392        6605        7722
##       India Afghanistan    Pakistan        Iraq
##       11506       11994       13899       23370
```

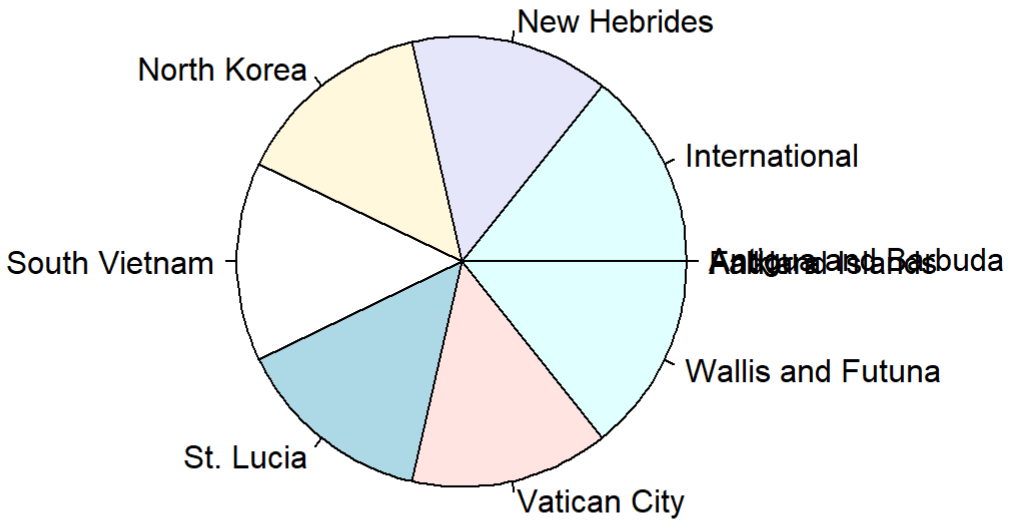
```
pie(t)
```




```
h
```

```
##
##             Andorra Antigua and Barbuda    Falkland Islands
##                   0                   0                   0
##       International        New Hebrides         North Korea
##                   1                   1                   1
##       South Vietnam           St. Lucia        Vatican City
##                   1                   1                   1
##   Wallis and Futuna
##                   1
```

```
pie(h)
```



De estos datos obtenemos que elpais con mas atentados terrroristas es Iraq seguido de Paquistan y Afganistan y los que menos son Andorra y antigua y Barbuda.

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

Los datos obtenidos con el contrate de hipotesis no permiten descartar la hipotesis nula , esto se puede deber a dos factores que realmente la hipotesis nula sea cierta y el riesgo en los paises desarrollados sea igual a los paises no desarrollados , o que hemos asumido la normalidad de los datos cuando estos realmente no lo son, es por ello que considero que el contraste de hipotesis no es valido en este estudio.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.R Markdown

El codigo y los datos tratados se encuentran como se solicita en el enunciado de la practica el codigo se encuentra en Github Para salvar los datos se utiliza el commando write.csv(gtd3, file = "../data/gtd3.csv")

```
write.csv(gtd3, file = "gtd3.csv")
```