Homicidios

Shamuel Manrrique NIP:802400

10/20/2020

Ejercicio propuesto 2

En el fichero de datos Homicidios.RData se tiene la variable cuantitativa Homicidios que indica el número de homicidios al año en las distintas regiones y subregiones de diferentes países del mundo.Además, se dispone de las variables cualitativas (factor) Region y Subregion.

1. Realizar una estadística descriptiva de la variable Homicidios agrupando por regiones y subregiones.
2. Agrupando por regiones/subregiones de dos en dos, decidir si el número de homicidios dependen de las regiones/subregiones consideradas.

Import libraries

Cargar los datos de Homicidios.RData

```
load("C:/Users/smmanrrique/3D Objects/unizar/add/aad_practices/practice1/datasets/Homicidios.RData")
```

Resumen general de la información del dataset

```
summary(Dataset)
```

```
##          País       Homicidios        Región                Subregión
##  Afganistán:  1   Min.   :    1   África :53   Caribe           : 22
##  Albania   :  1   1st Qu.:   93   América:47   África Oriental  : 17
##  Alemania  :  1   Median :  914   Asia   :50   Oeste de Asia    : 17
##  Andorra   :  1   Mean   : 5722   Europa :43   África Occidental: 16
##  Angola    :  1   3rd Qu.: 7366   Oceanía:14   Sudamérica       : 13
##  Anguila   :  1   Max.   :76409                Sur de Europa    : 13
##  (Other)   :201                                (Other)          :109
```

1. Estadística descriptiva

Realizar una estadística descriptiva de la variable Homicidios agrupando por regiones y subregiones.

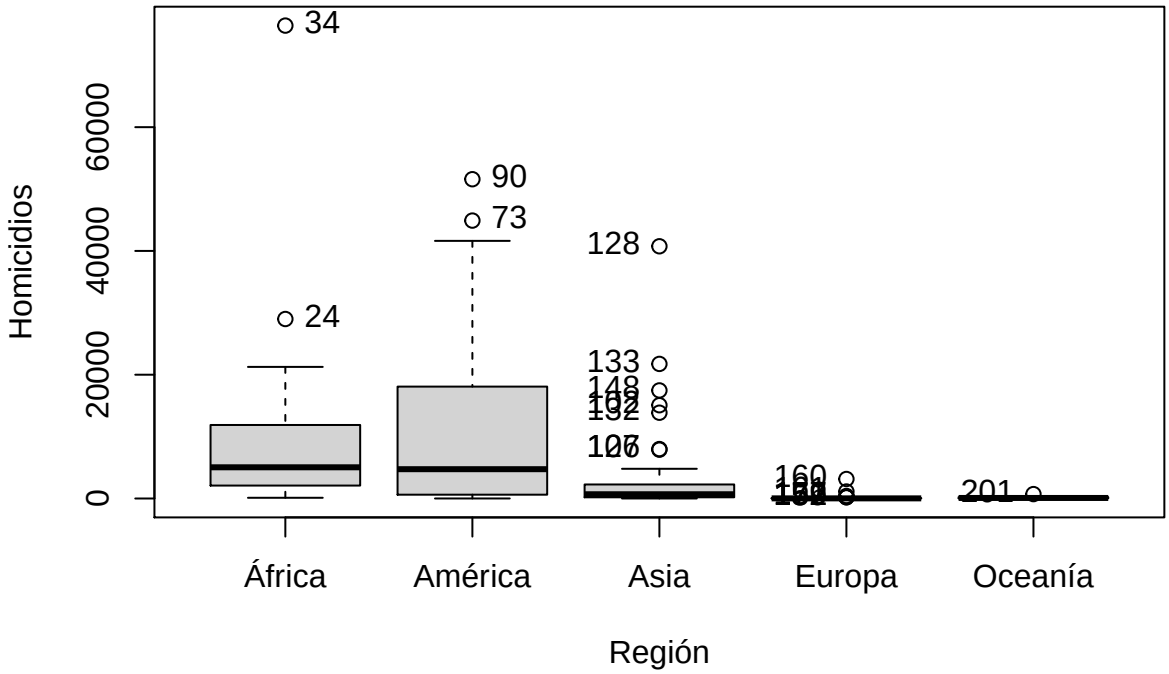
Region

```
# Homicidios por regiones
numSummary(Dataset[,c("Homicidios"), drop=FALSE], groups = Dataset$Región, statistics=c("mean", "sd", "
```

Datos de Homicidios agrupando por regiones.

```
##               mean         sd      IQR skewness   kurtosis  0%     25%    50%
## África   8929.8868 11531.6791  9790.00 4.068252 22.4819674 117 2089.00 5039.0
## América 11433.2128 13937.9778 17463.00 1.282990  0.7715068   3  616.00 4734.0
## Asia     3296.4200  7191.0775  2061.75 3.648488 15.4359882   2  193.25  683.5
## Europa    157.1163   503.2272    68.00 5.414308 31.4845973   1    6.00   23.0
## Oceanía   153.5714   194.9464   173.75 2.077224  4.6613133   1   30.25   85.5
##           75%  100% Homicidios:n
## África  11879 76409           53
## América 18079 51589           47
## Asia     2255 40752           50
## Europa     74  3149           43
## Oceanía   204   709           14
```

```
# Grafía de Boxplot de Homicidios por regiones
Boxplot(Homicidios~Región, data=Dataset, id=list(method="y"))
```




```
##  [1] "24"  "34"  "73"  "90"  "102" "106" "127" "128" "132" "133" "148" "152"
## [13] "156" "159" "160" "161" "172" "177" "201"
```

```
sqldf("SELECT  Región, COUNT(País) AS Count FROM Dataset GROUP BY Región order by count desc")
```

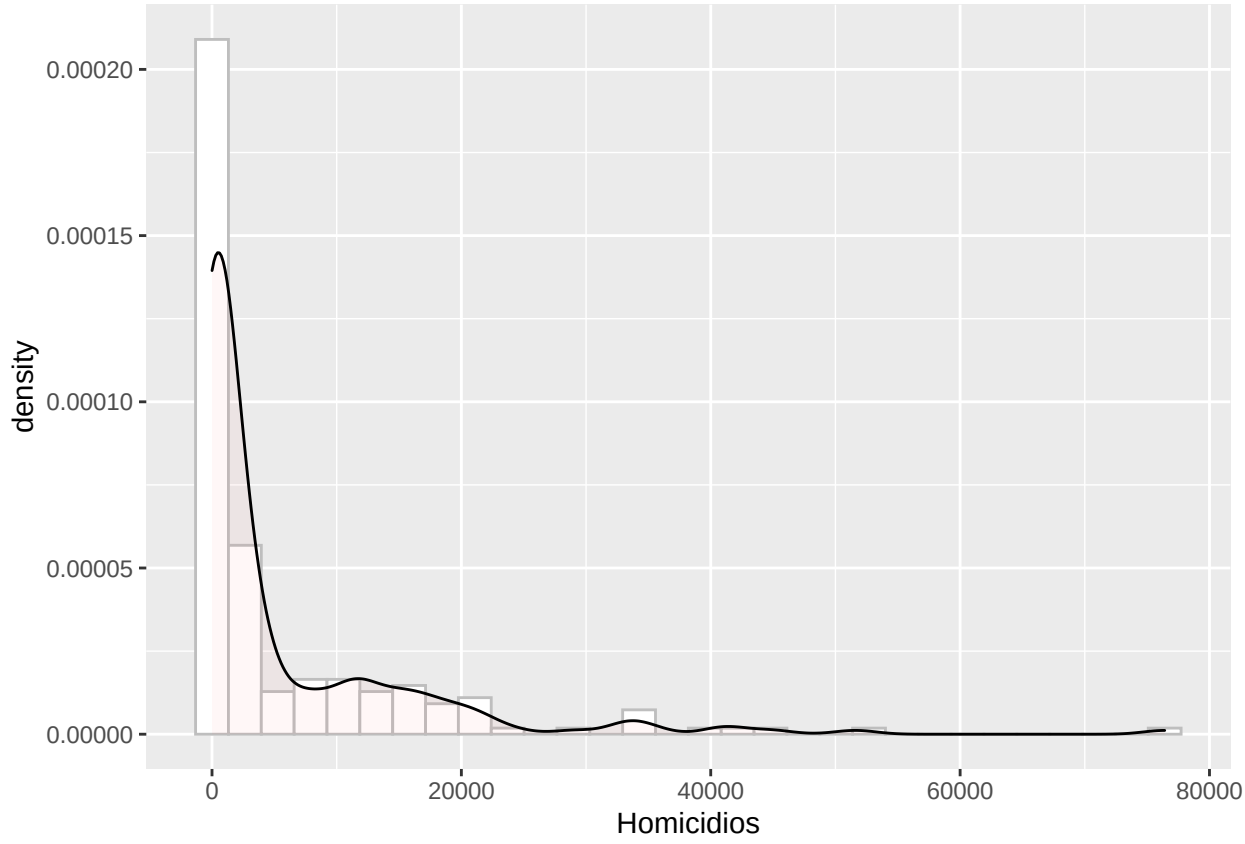
```
##    Región Count
## 1  África    53
## 2    Asia    50
## 3 América    47
## 4  Europa    43
```

```
## 5 Oceanía    14
```

Observando los datos por región muestra una relación entre el número de países que conforman una región y la desviación estándar. El valor del skewness es alto en las regiones indica que la distribución tiene una asimetría positiva. Casi todas las regiones excepto América tienen un valor de kurtosis mayor a tres(leptokurtosis) es decir que existen valores atípicos en estas distribuciones. Sin embargo, aunque el valor de kurtosis en América es bajo es la región con mayor desviación estándar(11531.6791 sd) y esparcimiento(17463.00 IQR).

Validar que los datos siguen una distribución Normal

```
# Histogram with density plot
ggplot(data=Dataset , aes(x=Homicidios)) +
 geom_histogram(aes(y=..density..), colour="gray", fill="white", bins = 30) +
 geom_density(alpha=.05, fill="#FF6666")
```




```
# Prueba de normalidad por Región
normalityTest(Homicidios~Región, test="shapiro.test", data=Dataset) # Shapiro
```

```
##
##  --------
##  Región = África
##
##  Shapiro-Wilk normality test
##
## data:  Homicidios
## W = 0.61579, p-value = 1.568e-10
##
```

```
##  --------
##  Región = América
##
##  Shapiro-Wilk normality test
##
## data:  Homicidios
## W = 0.80131, p-value = 1.698e-06
##
##  --------
##  Región = Asia
##
##  Shapiro-Wilk normality test
##
## data:  Homicidios
## W = 0.49685, p-value = 7.525e-12
##
##  --------
##  Región = Europa
##
##  Shapiro-Wilk normality test
##
## data:  Homicidios
## W = 0.31683, p-value = 6.975e-13
##
##  --------
##  Región = Oceanía
##
##  Shapiro-Wilk normality test
##
## data:  Homicidios
## W = 0.75465, p-value = 0.001447
##
##  --------
##
##  p-values adjusted by the Holm method:
##         unadjusted adjusted
## África  1.5679e-10 4.7036e-10
## América 1.6977e-06 3.3954e-06
## Asia    7.5246e-12 3.0099e-11
## Europa  6.9752e-13 3.4876e-12
## Oceanía 0.0014467  0.0014467
```

```
normalityTest(Homicidios~Región, test="ad.test", data=Dataset)   # Anderson-Darling
```

```
##
##  --------
##  Región = África
##
##  Anderson-Darling normality test
##
## data:  Homicidios
## A = 4.1222, p-value = 2.137e-10
##
##  --------
##  Región = América
```

```
##
##  Anderson-Darling normality test
##
## data:  Homicidios
## A = 3.4322, p-value = 1.014e-08
##
##  --------
##  Región = Asia
##
##  Anderson-Darling normality test
##
## data:  Homicidios
## A = 9.2964, p-value < 2.2e-16
##
##  --------
##  Región = Europa
##
##  Anderson-Darling normality test
##
## data:  Homicidios
## A = 10.72, p-value < 2.2e-16
##
##  --------
##  Región = Oceanía
##
##  Anderson-Darling normality test
##
## data:  Homicidios
## A = 1.2426, p-value = 0.001971
##
##  --------
##
##  p-values adjusted by the Holm method:
##         unadjusted adjusted
## África  2.1374e-10 6.4123e-10
## América 1.0139e-08 2.0279e-08
## Asia    < 2.22e-16 < 2.22e-16
## Europa  < 2.22e-16 < 2.22e-16
## Oceanía 0.0019706  0.0019706
```

```
normalityTest(Homicidios~Región, test="lillie.test", data=Dataset)
```

```
##
##  --------
##  Región = África
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Homicidios
## D = 0.22236, p-value = 6.023e-07
##
##  --------
##  Región = América
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
```

```
##
## data:  Homicidios
## D = 0.22053, p-value = 4.739e-06
##
##  --------
##  Región = Asia
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Homicidios
## D = 0.334, p-value = 1.763e-15
##
##  --------
##  Región = Europa
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Homicidios
## D = 0.37819, p-value < 2.2e-16
##
##  --------
##  Región = Oceanía
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  Homicidios
## D = 0.27052, p-value = 0.006518
##
##  --------
##
##  p-values adjusted by the Holm method:
##         unadjusted adjusted
## África  6.0231e-07 1.8069e-06
## América 4.7386e-06 9.4771e-06
## Asia    1.7635e-15 7.0538e-15
## Europa  < 2.22e-16 < 2.22e-16
## Oceanía 0.0065179  0.0065179
```

```
normalityTest(Homicidios~Región, test="cvm.test", data=Dataset)  # Cramer
```

```
##
##  --------
##  Región = África
##
##  Cramer-von Mises normality test
##
## data:  Homicidios
## W = 0.64668, p-value = 1.411e-07
##
##  --------
##  Región = América
##
##  Cramer-von Mises normality test
##
## data:  Homicidios
```

```
## W = 0.58949, p-value = 3.989e-07
##
##  --------
```

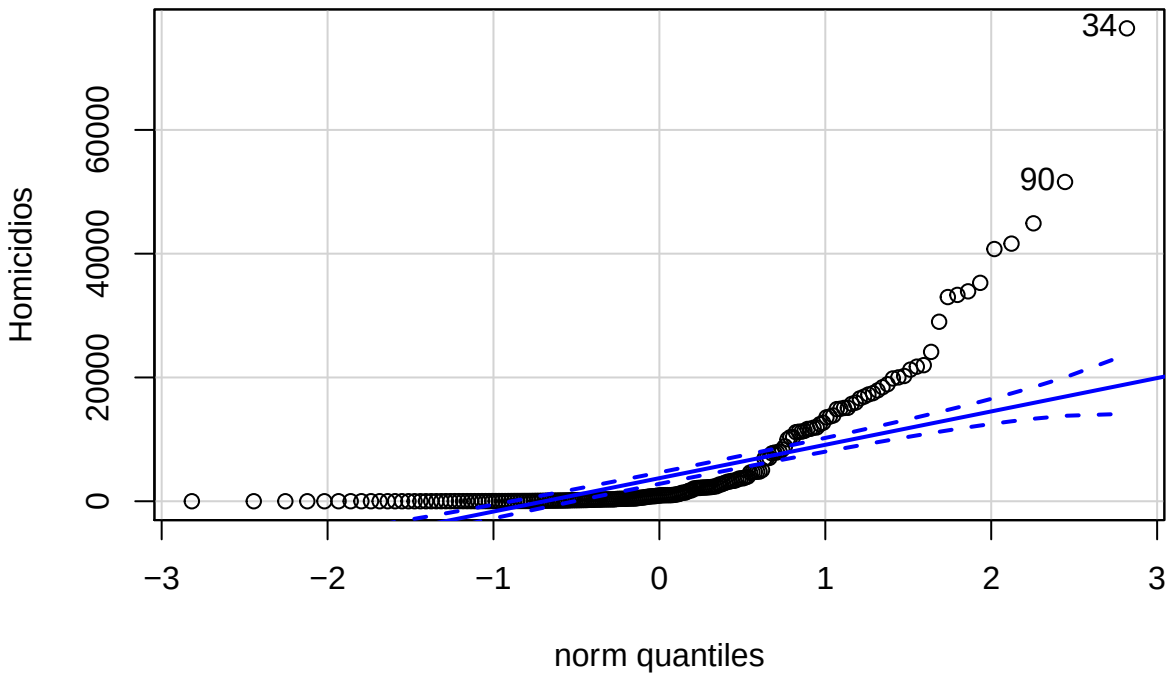
```
## Warning in cvm.test(x = c(1418L, 15072L, 143L, 203L, 831L, 7990L, 192L, : p-
## value is smaller than 7.37e-10, cannot be computed more accurately
```

```
##
##  Región = Asia
##
##  Cramer-von Mises normality test
##
## data:  Homicidios
## W = 1.8888, p-value = 7.37e-10
##
##  --------
```

```
## Warning in cvm.test(x = c(162L, 185L, 18L, 42L, 25L, 403L, 15L, 21L, 264L, : p-
## value is smaller than 7.37e-10, cannot be computed more accurately
```

```
##
##  Región = Europa
##
##  Cramer-von Mises normality test
##
## data:  Homicidios
## W = 2.1748, p-value = 7.37e-10
##
##  --------
##  Región = Oceanía
##
##  Cramer-von Mises normality test
##
## data:  Homicidios
## W = 0.21392, p-value = 0.002892
##
##  --------
##
##  p-values adjusted by the Holm method:
##         unadjusted adjusted
## África  1.4114e-07 4.2341e-07
## América 3.9892e-07 7.9784e-07
## Asia    7.3700e-10 3.6850e-09
## Europa  7.3700e-10 3.6850e-09
## Oceanía 0.0028922  0.0028922
```

```
# qqPlot para la validación de normalidad
with(Dataset, qqPlot(Homicidios, dist="norm", id=list(method="y", n=2, labels=rownames(Homicidios))))
```




```
## [1] 34 90
```

Los valores obtenidos en cada uno de los p-valores indican que la función de probabilidad no sigue una distribución normal, para realizar las pruebas de varianza y media iguales se usaran otras heurísticas que no requieran normalidad.

Validar si las varianzas son iguales o no

```
# Test Levene para dos varianzas con distribuciones distintas a la normal
leveneTest(filter(Dataset, Región=="África" )$Homicidios~filter(Dataset, Región=="África" )$Subregión, c
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##       Df F value    Pr(>F)
## group  4  6.8809 0.0001839 ***
##       48
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(filter(Dataset, Región=="África" )$Homicidios~filter(Dataset, Región=="África" )$Subregión, c
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  4  1.7519 0.1541
##       48
```

Con el valor obtenido menor a 0.05 no cumple con la hipótesis para decir que las varianzas son iguales, por lo que asumimos varianzas distintas.

Validar si las dos medias son iguales o no

Para aplicar el test de wilcoxon si la variable factor como en este caso tiene más de dos alternativas, se debe evaluar de cada una por pares, de lo contrario no será posible hacer la prueba con wilcox.test.

```
wilcox.test( filter(Dataset, Región=="África")$Homicidios, filter(Dataset, Región=="América")$Homicidios
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  filter(Dataset, Región == "África")$Homicidios and filter(Dataset, Región == "América")$Homici
## W = 1279.5, p-value = 0.817
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test( filter(Dataset, Región=="África")$Homicidios, filter(Dataset, Región=="Asia")$Homicidios)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  filter(Dataset, Región == "África")$Homicidios and filter(Dataset, Región == "Asia")$Homicidi
## W = 2091, p-value = 4.39e-07
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test( filter(Dataset, Región=="África")$Homicidios, filter(Dataset, Región=="Europa")$Homicidios)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  filter(Dataset, Región == "África")$Homicidios and filter(Dataset, Región == "Europa")$Homici
## W = 2240, p-value = 5.268e-16
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test( filter(Dataset, Región=="África")$Homicidios, filter(Dataset, Región=="Oceanía")$Homicidios
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  filter(Dataset, Región == "África")$Homicidios and filter(Dataset, Región == "Oceanía")$Homici
## W = 730, p-value = 3.226e-08
## alternative hypothesis: true location shift is not equal to 0
```

Por ser muchas las comparativas que se tendrían que hacer se tomó como región base África y se comparó si tenía igual media con el resto de regiones. El resultado arrojó que en todas las regiones se rechaza la hipótesis nula por lo que la media es distinta, excepto com America tienen la misma media con un p-value = 0.817.

Subregion

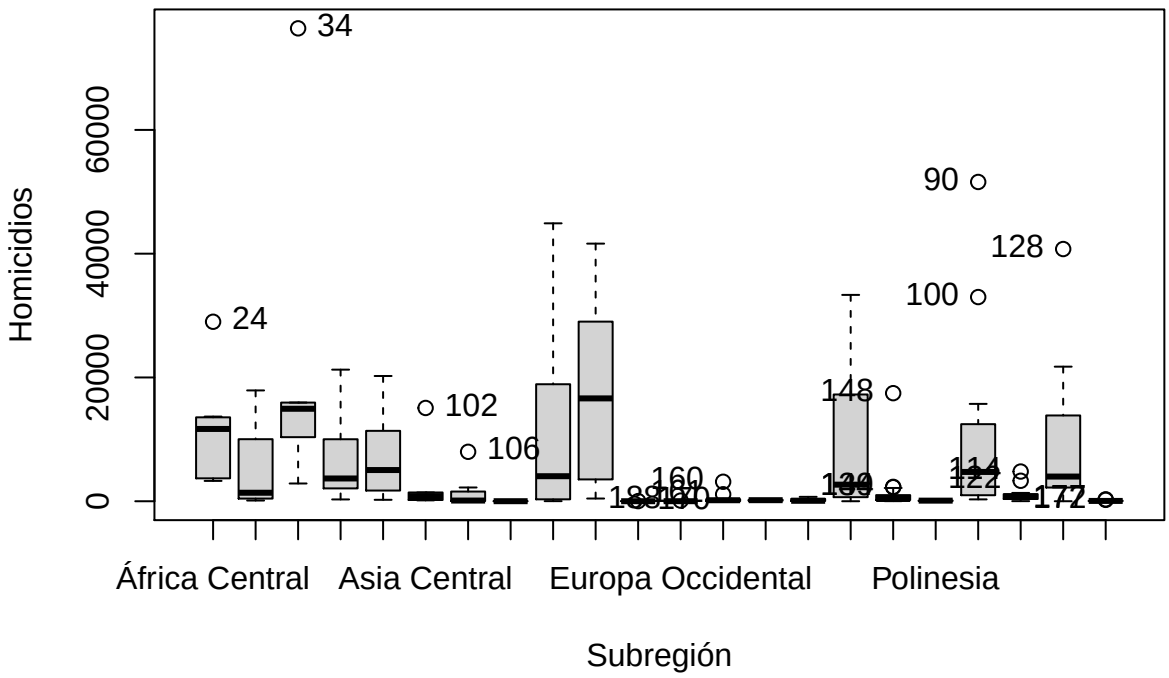
Datos de Homicidios agrupando por subregiones.

```
# Homicidios por regiones
numSummary(Dataset[,c("Homicidios"), drop=FALSE], groups = Dataset$Subregión, statistics=c("mean", "sd"
```

```
##                         mean           sd      IQR   0%      25%     50%
## África Central    11361.88889  7984.469056  9858.00 3289  3700.00 11680.0
## África del Norte   5205.16667  7266.183631  7614.00  117   458.75  1374.5
## África del Sur    24102.40000 29691.853837  5588.00 2870 10352.00 14941.0
```

```
## África Occidental  6623.31250  6603.743249  7364.50  294  2081.75  3678.0
## África Oriental    6665.35294  5593.366685  9647.00  233  1726.00  5039.0
## Asia Central       3533.40000  6471.158111  1215.00  143   203.00   831.0
## Asia Oriental      1425.75000  2762.069967  1215.50    2    26.25   112.0
## Australasia          10.00000     9.899495     7.00    3     6.50    10.0
## Caribe            10161.95455 12659.968741 18030.25   17   375.50  4045.0
## Centroamérica     17682.42857 16215.537249 25495.00  439  3526.50 16618.0
## Europa del Norte      6.60000     6.995236     6.50    1     2.25     3.0
## Europa Occidental    21.55556    31.642974    17.00    2     6.00     7.0
## Europa Oriental     492.18182   939.381799   310.50   15    23.00   162.0
## Melanesia           166.75000    92.128081   135.25   65   101.75   172.0
## Micronesia          247.20000   298.258613   349.00   27    40.00    71.0
## Norteamérica      10786.40000 14448.701924 16590.00    3   660.00  2678.0
## Oeste de Asia      1662.47059  4147.174974   793.00   18   197.00   366.0
## Polinesia            75.66667    65.957057    62.50    1    50.50   100.0
## Sudamérica        10468.38462 15369.566897 11483.00  300   975.00  4734.0
## Sudeste de Asia    1227.36364  1486.503567   791.00   16   363.50   759.0
## Sur de Asia       10442.77778 13394.458843 11645.00    5  2215.00  3988.0
## Sur de Europa        83.23077    93.666210    60.00    1    26.00    52.0
##                        75%  100% Homicidios:n
## África Central    13558.00 29000            9
## África del Norte   8072.75 17906            6
## África del Sur    15940.00 76409            5
## África Occidental  9446.25 21262           16
## África Oriental   11373.00 20239           17
## Asia Central       1418.00 15072            5
## Asia Oriental      1241.75  7990            8
## Australasia          13.50    17            2
## Caribe            18405.75 44907           22
## Centroamérica     29021.50 41624            7
## Europa del Norte      8.75    23           10
## Europa Occidental    23.00   102            9
## Europa Oriental     333.50  3149           11
## Melanesia           237.00   258            4
## Micronesia          389.00   709            5
## Norteamérica      17250.00 33341            5
## Oeste de Asia       990.00 17463           17
## Polinesia           113.00   126            3
## Sudamérica        12458.00 51589           13
## Sudeste de Asia    1154.50  4800           11
## Sur de Asia       13860.00 40752            9
## Sur de Europa        86.00   307           13
```

```
# Grafía de Boxplot de Homicidios por regiones
Boxplot(Homicidios~Subregión, data=Dataset, id=list(method="y"))
```




```
##  [1] "24"  "34"  "102" "106" "170" "188" "160" "161" "134" "148" "149" "90"
## [13] "100" "114" "122" "128" "172" "177"
```

Se obtuvo la información de las medias y desviación estándar por subregión, los valores de skewness y kurtosis no se pudieron calcular dado que muchas de las subregiones no tenía muestras mayores a cuatro lo cual era requerido para para estos test. Vemos que los valores de desviación son bastante altos.

Subregiones de la región Áfricana

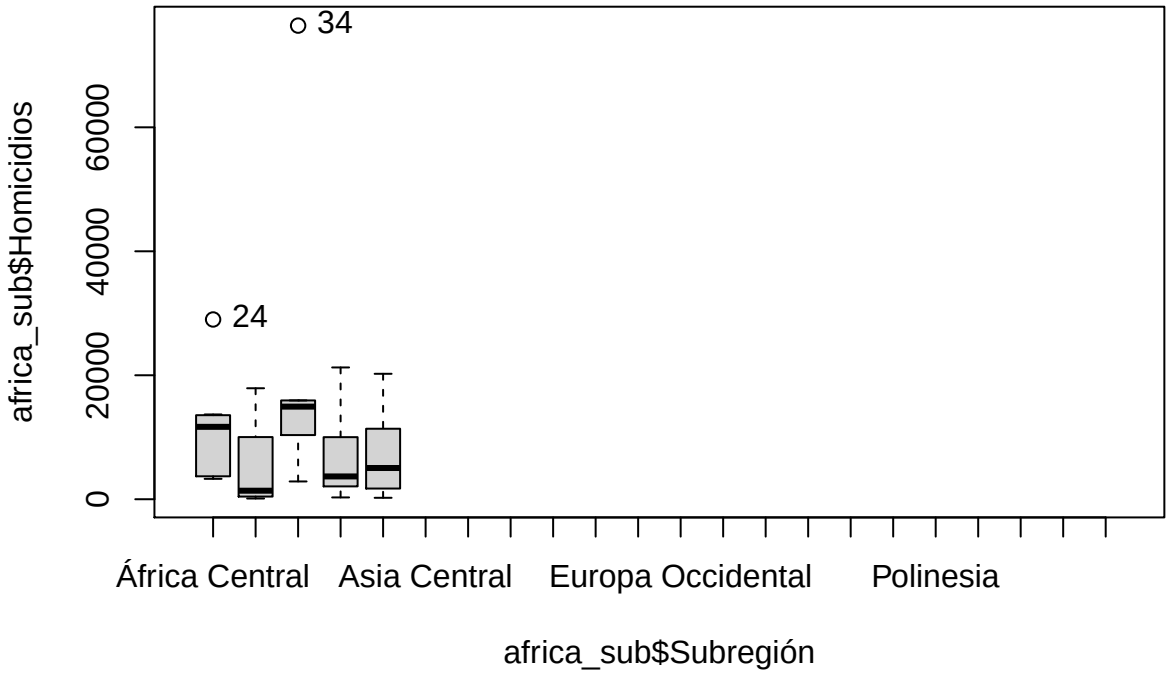
```
africa_sub = filter(Dataset, Región=="África")

# Homicidios por regiones
numSummary(africa_sub[,c("Homicidios"), drop=TRUE], groups = africa_sub$Subregión, statistics=c("mean",
```

```
## Warning in numSummary(africa_sub[, c("Homicidios"), drop = TRUE], groups
## = africa_sub$Subregión, : the following groups are empty: Asia Central,
## Asia Oriental, Australasia, Caribe, Centroamérica, Europa del Norte, Europa
## Occidental, Europa Oriental, Melanesia, Micronesia, Norteamérica, Oeste de Asia,
## Polinesia, Sudamérica, Sudeste de Asia, Sur de Asia, Sur de Europa
```

```
##                         mean        sd    IQR  skewness  kurtosis   0%      25%
## África Central    11361.889  7984.469 9858.0 1.2745837 2.5792753 3289  3700.00
## África del Norte   5205.167  7266.184 7614.0 1.4013812 0.8579916  117   458.75
## África del Sur    24102.400 29691.854 5588.0 2.0633211 4.4417164 2870 10352.00
## África Occidental  6623.312  6603.743 7364.5 1.1728218 0.2599456  294  2081.75
## África Oriental    6665.353  5593.367 9647.0 0.8212739 0.2958608  233  1726.00
##                        50%      75%  100% data:n
```

```
## África Central    11680.0 13558.00 29000      9
## África del Norte   1374.5  8072.75 17906      6
## África del Sur    14941.0 15940.00 76409      5
## África Occidental  3678.0  9446.25 21262     16
## África Oriental    5039.0 11373.00 20239     17
```

```
# Grafía de Boxplot de Homicidios por regiones
Boxplot(africa_sub$Homicidios~africa_sub$Subregión)
```




```
## [1] "24" "34"
```

Por subregión africana todas menos África Oriental son asimétricas positivas, además que los valores de la media de África Central y África del Sur son doble y triple respecto a las demás. También son estas dos subregiones que tienen este valor de kurtosis más alto, es decir que posiblemente existan valores atípicos en estas distribuciones.

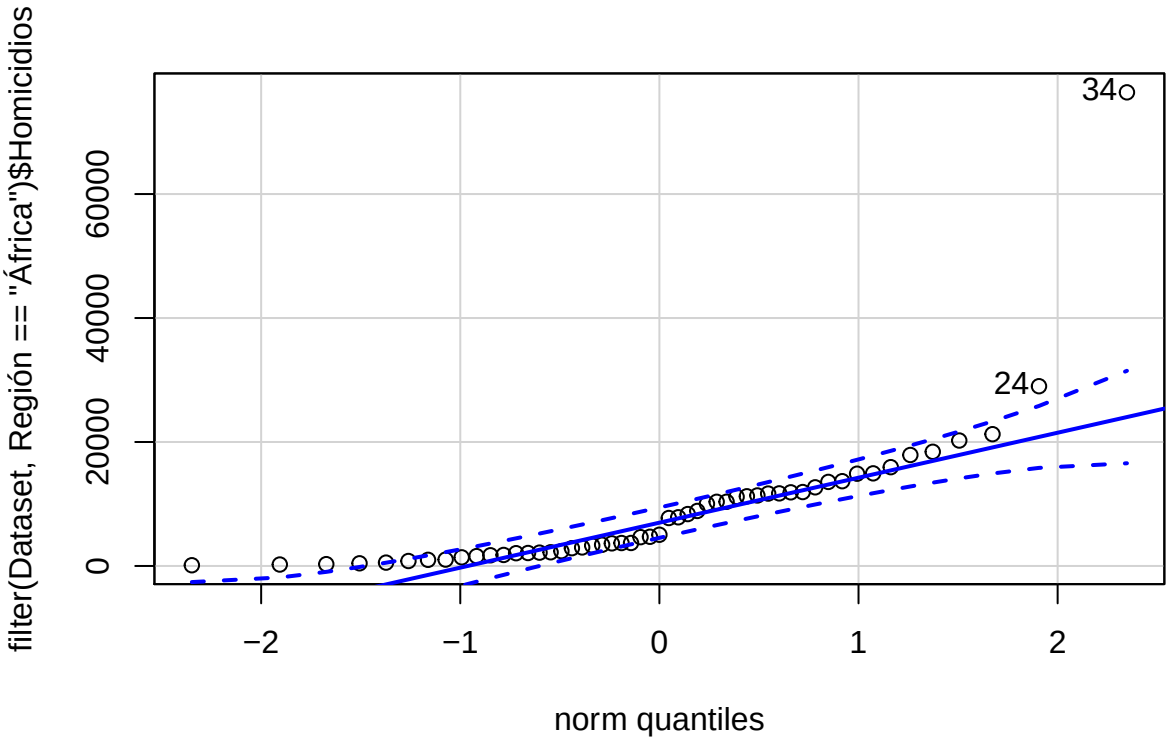
Validamos que los datos siguen una distribución Normal

```
# Prueba de normalidad por Región
#normality Test(filter(Dataset, Región=="África" )$Homicidios~filter(Dataset, Región=="África" )$Subreg
#normalityTest(filter(Dataset, Región=="África" )$Homicidios~filter(Dataset, Región=="África" )$Subregi
normalityTest(filter(Dataset, Región=="África" )$Homicidios~filter(Dataset, Región=="África" )$Subregió
```

```
##
##  --------
##  filter(Dataset, Región == "África")$Subregión = África Central
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
```

```
##
## data:  filter(Dataset, Región == "África")$Homicidios
## D = 0.27477, p-value = 0.0487
##
##  --------
##  filter(Dataset, Región == "África")$Subregión = África del Norte
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  filter(Dataset, Región == "África")$Homicidios
## D = 0.32672, p-value = 0.04431
##
##  --------
##  filter(Dataset, Región == "África")$Subregión = África del Sur
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  filter(Dataset, Región == "África")$Homicidios
## D = 0.4083, p-value = 0.006513
##
##  --------
##  filter(Dataset, Región == "África")$Subregión = África Occidental
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  filter(Dataset, Región == "África")$Homicidios
## D = 0.24269, p-value = 0.01253
##
##  --------
##  filter(Dataset, Región == "África")$Subregión = África Oriental
##
##  Lilliefors (Kolmogorov-Smirnov) normality test
##
## data:  filter(Dataset, Región == "África")$Homicidios
## D = 0.16197, p-value = 0.277
##
##  --------
##
##  p-values adjusted by the Holm method:
##                   unadjusted adjusted
## África Central    0.0487008  0.132931
## África del Norte  0.0443104  0.132931
## África del Sur    0.0065126  0.032563
## África Occidental 0.0125320  0.050128
## África Oriental   0.2769696  0.276970
```

```
#normalityTest(filter(Dataset, Región=="África" )$Homicidios~filter(Dataset, Región=="África" )$Subregi

# qqPlot para la validación de normalidad
with(Dataset, qqPlot(filter(Dataset, Región=="África" )$Homicidios, dist="norm", id=list(method="y", n=
```




```
## [1] 34 24
```

Se obtuvo que en la mayoría de los casos las subregiones no siguen una distribución normal, dado que el p-value es menor a 0.05.

Validamos si las dos varianzas son iguales o no

```
# Test Levene para dos varianzas con distribuciones distintas a la normal
leveneTest(Homicidios~Subregión, data=Dataset)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value    Pr(>F)
## group  21  3.1182 1.912e-05 ***
##       185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(Homicidios~Subregión, data=Dataset, center=mean)
```

```
## Levene's Test for Homogeneity of Variance (center = mean)
##        Df F value    Pr(>F)
## group  21  7.7883 < 2.2e-16 ***
##       185
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Al parecer con el valor obtenido menor a 0.05 no es significativo para decir que las varianzas son iguales Por lo que asumimos varianzas distintas.

Validamos si las dos medias son iguales o no

```
africa = filter(Dataset,  Región=="África")
# table(africa$Subregión)

wilcox.test( filter(africa, Subregión=="África Central")$Homicidios, filter(Dataset, Subregión=="África
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  filter(africa, Subregión == "África Central")$Homicidios and filter(Dataset, Subregión == "Áfi
## W = 43, p-value = 0.06633
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test( filter(africa, Subregión=="África Central")$Homicidios, filter(Dataset, Subregión=="África
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  filter(africa, Subregión == "África Central")$Homicidios and filter(Dataset, Subregión == "Áfi
## W = 17, p-value = 0.5185
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test( filter(africa, Subregión=="África Central")$Homicidios, filter(Dataset, Subregión=="África
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  filter(africa, Subregión == "África Central")$Homicidios and filter(Dataset, Subregión == "Áfi
## W = 103, p-value = 0.08424
## alternative hypothesis: true location shift is not equal to 0
```

```
wilcox.test( filter(africa, Subregión=="África Central")$Homicidios, filter(Dataset, Subregión=="África
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  filter(africa, Subregión == "África Central")$Homicidios and filter(Dataset, Subregión == "Áfi
## W = 111, p-value = 0.06616
## alternative hypothesis: true location shift is not equal to 0
```

Seleccionando a la región Africana y estudiando sus subregiones con el test de Wilcoxon no podemos asumir que las medias son distintas dado que los valores de los p-value son mayores que 0.05.