Examen 4

3/2/2021

Dimas Ramirez Luis Daniel

Palma Ponce Adriana Lizeth

Peñaloza Ponce Nayeli

Primero vamos a preparar los datos para utilizarlos en los ejercicios. A nuestro equipo le toco

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
##
## 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
```

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.6.3
```

```
base <- read_excel("F:/7mo semestre/Estadistica aplicada/SO2.xls")
#head(base)
b <- select(base,FECHA, HORA,SFE,TAH)
b$SFE[b$SFE == -99] <- mean(b$SFE[b$SFE != -99])
b$TAH[b$TAH ==-99] <- mean(b$TAH[b$TAH != -99])

head(b)
```

```
## # A tibble: 6 x 4
##   FECHA                HORA   SFE   TAH
##   <dttm>              <dbl> <dbl> <dbl>
## 1 2019-01-01 00:00:00     1  2.45  1.60
## 2 2019-01-01 00:00:00     2  2.45  1.60
## 3 2019-01-01 00:00:00     3  2.45  1.60
## 4 2019-01-01 00:00:00     4  2     3
## 5 2019-01-01 00:00:00     5  2     2
## 6 2019-01-01 00:00:00     6  1     1
```

```
str(b$FECHA)
```

```
##  POSIXct[1:8760], format: "2019-01-01" "2019-01-01" "2019-01-01" "2019-01-01" "2019-01-01" ...
```

```
Table <- data.frame(months.POSIXt(b$FECHA),b$HORA,b$SFE,b$TAH )

Table1 <- subset(Table, months.POSIXt.b.FECHA.== "septiembre")
head(Table1)
```

```
##      months.POSIXt.b.FECHA. b.HORA b.SFE    b.TAH
## 5833             septiembre      1     0 1.596303
## 5834             septiembre      2     0 1.596303
## 5835             septiembre      3     0 1.596303
## 5836             septiembre      4     3 1.596303
## 5837             septiembre      5     2 1.596303
## 5838             septiembre      6     1 1.596303
```

Ejercicio 1 .- Prueba de signos

Escoger un día de la semana y un punto de inspección para realizar una prueba de signos.

Ahora crearemos una data frame específico para este ejercicio, solamente consideramos una estación correspondiente a Cuajimalpa (SFE) y las filas 360:384 porque son las 24 horas del 15/09/2019. Primero sacaremos la mediana del mes de septiembre para utilizarlo como H_0.

```
eje0 <- data.frame(Table1$months.POSIXt.b.FECHA., Table1$b.SFE)
summary(eje0)
```

```
##  Table1.months.POSIXt.b.FECHA.  Table1.b.SFE
##  septiembre:720                Min.   : 0.000
##  abril     :  0                1st Qu.: 1.000
##  agosto    :  0                Median : 1.000
##  diciembre :  0                Mean   : 2.141
##  enero     :  0                3rd Qu.: 2.454
##  febrero   :  0                Max.   :23.000
##  (Other)   :  0
```

Del código anterior obtenemos que la mediana del mes deseptiembre fue 1

Por lo tanto se desea probar $H_0 : \hat{\mu} = 1$; $H_A : \hat{\mu} \neq 1$ con $\alpha = 0.05$

Finalmente, realizando el procedimiento para aplicar la prueba de signos

```
med1 <- 1
eje1 <- data.frame(Table1$months.POSIXt.b.FECHA., Table1$b.SFE)
a <- eje1$Table1.b.SFE[360:384]

b <- length(a[a>med1])
b
```

```
## [1] 16
```

```
c <- length(a[a==med1])
c
```

```
## [1] 7
```

```
d<- length(a)-c
d
```

```
## [1] 18
```

```
binom.test(b,d)
```

```
##
##  Exact binomial test
##
## data:  b and d
## number of successes = 16, number of trials = 18, p-value = 0.001312
```

```
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.6528796 0.9862488
## sample estimates:
## probability of success
##              0.8888889
```

Valor $P = 0.88$; por lo tanto no existe suficiente evidencia para rechazar la hipótesis nula.

Al inicio del planteamiento de este ejericio se planteo la posibilidad de que, debido a que es un dia festivo y patriotico para México, podría existir mayor movilidad vehicular y de emisores de este contaminante, por lo tanto habría una mayor cantidad de S02 reflejada en la mediana del 15 de Septiembre comparada con la mediana del año en general, sin embargo con la Prueba de los signos, notamos que no existe suficiente evidencia para apoyar nuestro planteamiento teórico, si no que la mediana del día 15-Septiembre es muy seguramente igual a la mediana de todo el año.

Ejercicio 2.- Prueba de signos pareados

Escoger dos días consecutivos y hacer una prueba de hipotésis con prueba de signos pareada.

Para este inciso, continuaremos con la línea del planteamiento anterior. Sin embargo aquí buscaremos saber si la mediana del día 15-Sept y 16-Sept (con los registros la estación de Cuajimalpa) son iguales. De inicio esperaríamos nuevamente que las medianas no sean iguales, debido a que el 15 es un día bastante activo refiriendonos a movilidad y el 16 al ser un día festivo, de manera general, hay menos movilidad debido a que la gente descansa. Entonces el equipo espera que la prueba nos indique que existe suficiente evidencia para rechazar la H_0

Por lo tanto nuestras hipotésis, con un nivel de significancia de 0.05 quedarían definidas como:

$$H_0 : \hat{\mu}_1 - \hat{\mu}_2 = 0$$

$$H_A : \hat{\mu}_1 - \hat{\mu}_2 \neq 0$$

```
e <- eje1$Table1.b.SFE[360:384]
f <- eje1$Table1.b.SFE[384:408]
g <- e-f
g
```

```
##  [1]  2.453588  2.453588  2.453588  2.453588  2.000000  1.000000  0.000000
##  [8]  2.000000  8.000000  8.000000  9.000000  7.000000  5.000000  4.000000
## [15]  4.000000  4.000000  2.000000  1.000000  0.000000  0.000000  1.000000
## [22]  1.000000  0.000000  0.000000 -1.000000
```

```
med2 <- 0

h <- length(g[g>med2])
i<- length(g[g=med2])
j <- length(g)-i

binom.test(h, j )
```

```
##
##  Exact binomial test
##
## data:  h and j
## number of successes = 19, number of trials = 25, p-value = 0.01463
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.5487120 0.9064356
## sample estimates:
## probability of success
##                   0.76
```

Valor $P = 0.88$; por lo tanto no existe suficiente evidencia para rechazar la hipótesis nula.

Este resultado nos indica que la mediana del registro de SO2 es el mismo tanto el 15 como el 16 de septiembre.

Ejercicio 3.- Prueba de Rangos con Signos de Wilcoxon

Usando los mismos datos del anterior, usar la prueba de Wilcoxon

Para esta prueba estableceremos la hipotésis de que los dos días pertenecen población esperando que la prueba nos indique que existe suficiente evidencia para rechazar la hipotésis.

```
wilcox.test(e,f, paired = TRUE)
```

```
## Warning in wilcox.test.default(e, f, paired = TRUE): cannot compute exact p-
## value with ties
```

```
## Warning in wilcox.test.default(e, f, paired = TRUE): cannot compute exact p-
## value with zeroes
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  e and f
## V = 207, p-value = 0.0001434
## alternative hypothesis: true location shift is not equal to 0
```

A un nivel de significancia de 0.05 y con un $p - value = 0.0001434$ podemos concluir que existe sufuciente evidencia para rechazar la hipotésis nula, por lo tanto, las emisiones de S02 registradas en la estación de monitoreo de Cuajimalpa el día 15 y 16 de Septiembre corresponden a diferentes poblaciones.

Ejercicio 4.- Kruskal-Wallis

Realizar pruebas de Kruskal-Wallis para observar si existen diferencias entre la concentración de SO2 entre los meses del año 2019

Para este ejercicio utilizaremos la información correspondiente a Tlahuac, y elegimos la hora pico para realizar la prueba, las 15:hrs.

Nuetra hipotésis nula para este caso es

H_0 :El SO2 es idéntico a través de los meses durante el 2019

```r
eje4 <- data.frame(Table$months.POSIXt.b.FECHA., Table$b.HORA ,Table$b.TAH)
Table4 <- subset(Table, b.HORA == 15 )
#Table4
names(Table4) <- c("Mes", "HORA","SFE","TAH")
head(Table4)
```

```
##       Mes HORA SFE TAH
## 15  enero   15   2   1
## 39  enero   15   5   1
## 63  enero   15   6   1
## 87  enero   15   5   1
## 111 enero   15   5   8
## 135 enero   15   4   4
```

```r
kruskal.test(SFE~Mes, data=Table4)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  SFE by Mes
## Kruskal-Wallis chi-squared = 38.519, df = 11, p-value = 6.391e-05
```

Con un $p - value = 6.391e - 05$, existe suficiente para rechazar la hipotésis nula, por lo tanto el registro de SO2 fue el mismo durante el 2019 en la estación de Tlahuac.

Ejercicio 5.- Prueba de Rachas

Empleando los datos del ejercicio uno realizar la prueba de rachas para determinar si la concentración de SO2 en un día es una varaible aleatoria

H_0 : Los registros de SO2 en la unidad de Cuajimalpa en el mes de septiembre son variables aleatorias.

En este ejercicio esperariamos que las varibles no sean aleatorias, es decir, que exista suficiente evidencia para rechazar la hipotésis nula lo cual nos indicaría indicios de un posible fenomeno que explique el comportamiento de los datos.

```r
library(tseries)
```

```
## Warning: package 'tseries' was built under R version 3.6.3
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

```r
a
```

```
##  [1]  2.453588  2.453588  2.453588  2.453588  2.000000  1.000000  1.000000
##  [8]  3.000000  9.000000 10.000000 11.000000  8.000000  6.000000  6.000000
## [15]  7.000000  6.000000  4.000000  2.000000  1.000000  1.000000  1.000000
## [22]  1.000000  0.000000  1.000000  0.000000
```

```r
runs.test(as.factor(a>median(a)))
```

```
##
##  Runs Test
##
## data:  as.factor(a > median(a))
## Standard Normal = -4.264, p-value = 2.008e-05
## alternative hypothesis: two.sided
```

Con un $p - value = 2.008e - 05$ y a un nivel de significancia del 0.05 existe suficiente evidencia para rechzar la hipotésis nula, por lo tanto los registros no son variable aleatorias y tienen un comportamiento que podría ser estudiado para explicarlo.

Ejercicio 6.- Prueba de Kolmogorov-Smirnov

Empleando los datos del ejercicio 4, hacer una prueba de Kolmogorov-Smirnov, proponga tres distribuciones de probabilidad que se pueda ajustar a los datos. Realizar un histograma con los datos.

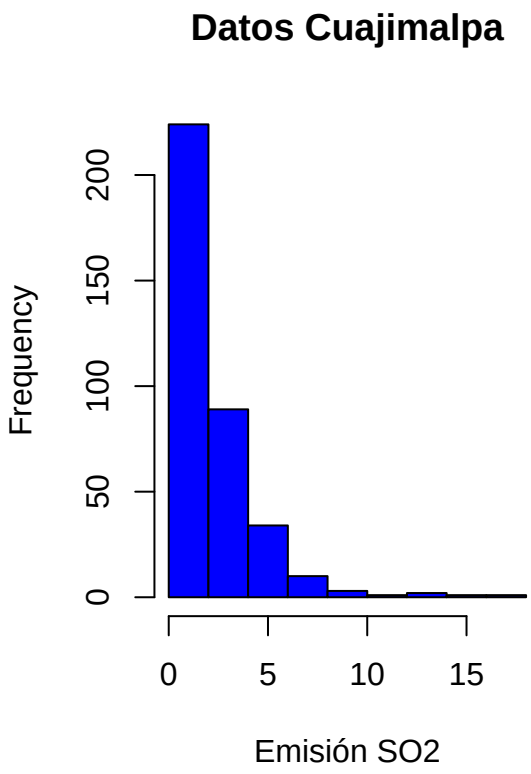
Primero denominamos una variable con todos los datos de la estación Cuajimalpa a las 15 horas del mes de septiembre, y propusimo s 3 distribuciones y comparamos cada una.

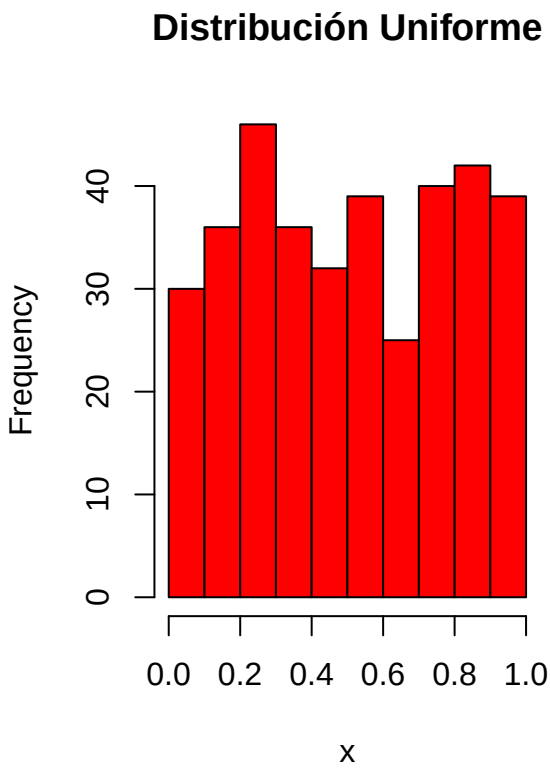
H_O : Las poblaciones responden a la misma distribución de probabilidad.

```r
eje6 <- Table4$SFE
#str(eje6)
x<- runif(365)
y<- rnorm(365)
z<- rf(365,1,10)
```

Primero comparamos los datos de la estación con una distribución Uniforme

```r
par(mfrow=c(1,2))
hist(eje6, col="Blue", main="Datos Cuajimalpa", xlab="Emisión SO2")
hist(x, col="Red", main="Distribución Uniforme")
```






```
ks.test(eje6, x)
```

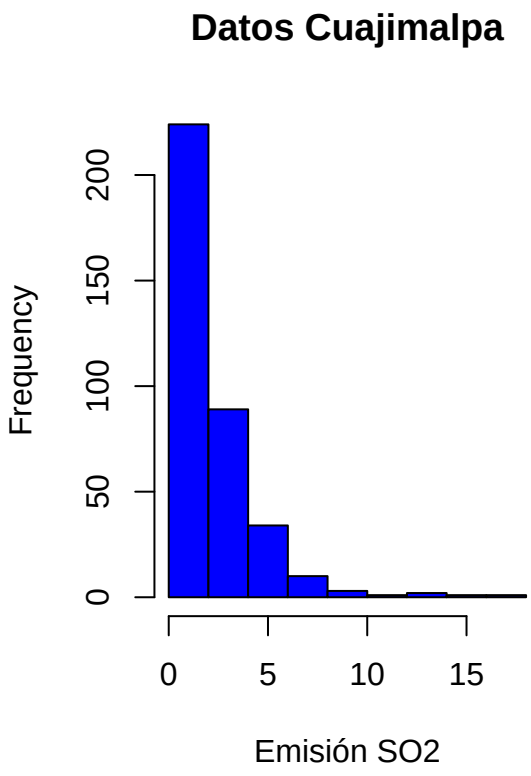
```
## Warning in ks.test(eje6, x): p-value will be approximate in the presence of ties
```

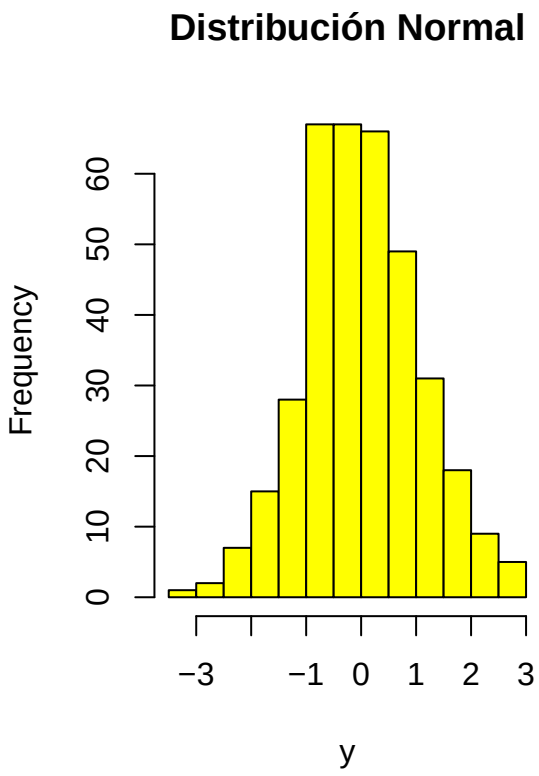
```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  eje6 and x
## D = 0.94521, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Como p-value = 0.009683, existe suficiente evidencia para rechzar la hipotésis nula, es evidente que las distribuciones no son las mismas.

Segundo, comparamos los datos de la estación con una distribución Normal

```
par(mfrow=c(1,2))
hist(eje6, col="Blue", main="Datos Cuajimalpa", xlab="Emisión SO2")
hist(y, col="Yellow", main="Distribución Normal")
```






```
ks.test(eje6, y)
```

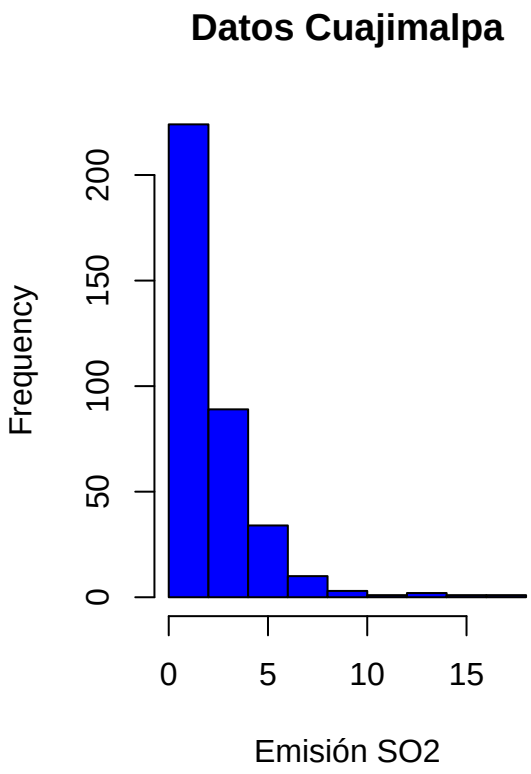
```
## Warning in ks.test(eje6, y): p-value will be approximate in the presence of ties
```

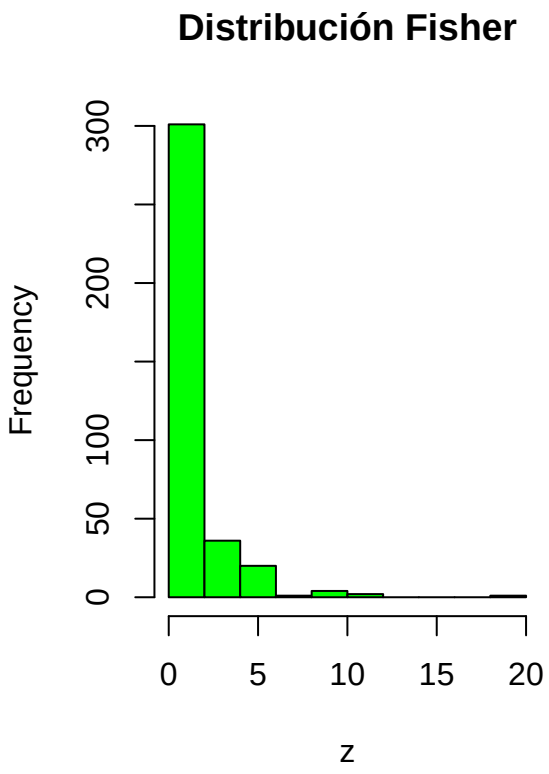
```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  eje6 and y
## D = 0.7726, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Como p-value $<$ 2.2e-16, existe suficiente evidencia para rechzar la hipotésis nula, por lo que las distribuciones no son las mismas.

Por último, comparamos los datos de la estación con una distribución Fisher

```
par(mfrow=c(1,2))
hist(eje6, col="Blue", main="Datos Cuajimalpa", xlab="Emisión SO2")
hist(z, col="Green", main="Distribución Fisher")
```






```
ks.test(eje6, z)
```

```
## Warning in ks.test(eje6, z): p-value will be approximate in the presence of ties
```

```
##
##  Two-sample Kolmogorov-Smirnov test
##
## data:  eje6 and z
## D = 0.61644, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

Como p-value $<$ 2.2e-16, existe suficiente evidencia para rechzar la hipotésis nula, por lo que las distribuciones no son las mismas sin embargo, de manera gráfica podría parecer que sí se aproximan.

Ejercicio 7.- Prueba de Correlación de Rangos

Con los datos del ejercicio dos realizar una prueba de correlación de rangos

Para este ejercicio esperariamos que exista suficiente evidencia para rechazar la hipotésis nula, lo cual nos indicaría que sí existe una correlación entre un día y otro, y nuevamente existe un fenomeno que explica la relación entre el 15 y 16 de septiembre en la emisión de S02 en la estación de Cuajimalp

$$H_0 : \rho = 0$$

```r
#e vector donde están localizados los valores del 15 de Septiembre
#f vector donde están localizados los valores del 16 de Septiembre

cor.test(e,f, method = "spearman")
```

```
## Warning in cor.test.default(e, f, method = "spearman"): Cannot compute exact p-
## value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  e and f
## S = 1110.4, p-value = 0.002759
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.5729269
```

Como el valor es p-value = 0.002759 y a un valor de significacia del 0.05, existe suficiente evidencia para rechazar la hipotésis nula y por lo tanto sí existe correlación entre un día y otro.

Conclusión General

Debido al escenario que planteamos, podemos concluir que a pesar de que el 15 de Septiembre es un día festivo y se creyó que las emisiones de S02 podrían ser mayores, no es así comparado con la mediana de las emisiones en el año 2019. Por otro lado, comparando el 15 y 16 de septiembre prácticamente no hay diferencia estadísticamente significativa para afirmar que no son iguales las emisiones de un día respecto al otro. No logramos encontrar una distribución que se asemeje a la de nuestros datos para esta estación en específico, sin embargo de manera gráfica logramos asemejarla a una distribución F~Fisher.