Lab08-Non Parametric Test- NON-NORMAL DISTS.

IMPORTANT!
The reasons why a variable is not normal can be important.
• If the variable was not expected to be normal, then there is no problem.
• If the variable was expected to be normal but does not follow a normal distribution it could mean that our sample is not representative of the population we want to study.

▼ The sign test

Under non parametric procedures, the mean u is replaced by the median. Remember that it is also the description parameter
of choice when the variable does not follow a normal distribution.

As always the Hypothesis are :
$H0$: $\tilde{\mu} = \tilde{\mu}0$
$H1$: $\tilde{\mu} \neq \tilde{\mu}0$

The test consists basically in replacing each exceeding
$\tilde{\mu}$ with a plus sign and each sample value bellow 0
$\tilde{\mu}$ with a minus sign.

If the null hypothesis is true and the population
is symmetric, the sum of the plus signs should be approximately equal to the sum of the minus signs. All values exactly equal to 0
$\tilde{\mu}$ are excluded from the analysis and the sample
size is reduced.

Since the values we are using (after the sign assignation is finished) are either positive or negative (2 values) the test is a binomial random variable X representing the number of plus signs in our random sample.

If the null hypothesis $H0 : \tilde{\mu} = \tilde{\mu}0$ is true theprobability that a sample value results in either a plus or a minus is 1/2.

We can do it easily with:

```
table(DATA$EMISION < median(DATA$EMISION))
FALSE TRUE
2970 2970
table(DATA$EMISION > median(DATA$EMISION))
FALSE TRUE
2970 2970

#This would be the perfect scenario, after checking
#the data is NOT NORMAL with shapirio
```

▼ One sample comparison

ONE-SAMPLE WILCOXON SIGNED RANK TEST, when variable doesnt follow normal distribution

For comparing one sample with a fixed value.

1. We check the normality
2. We do the sign test

Now this test tell us if the median is equal to mu. H0

```
wilcox.test(DATA$EMISION, mu = 200)
```

And with the p-value we make the same assumptions as always.

If the p value is less than 0.05, we reject the null hypothesis.

▼ 2 independent samples

We will use the **Willcoxon-Mann-Whitney U test**, for comparing the means of two independent continuous distributions.

STEPS

- Order observations of both samples and we asign a rank to each observation.

  $Samplex$ = 100, 110, 130
  $Sampley$ = 105, 115, 125
  $JoinedSamples$ = 100, 105, 110, 115, 125, 130
  $Ranks$ = 1, 2, 3, 4, 5, 6

- If there are repeated observations we replace them by the mean of the ranks that observations would have if they were different.

  $Samplex$ = 100, 110, 130
  $Sampley$ = 105, 110, 125
  $JoinedSamples$ = 100, 105, 110, 110, 125, 130
  $Ranks$ = 1, 2, 3, 4, 5, 6 → 1, 2,((3 + 4)/2), 5, 6
  → 1, 2, 3.5, 3.5, 5, 6

- Now we sum the ranks by sample.

  $Samplex$ = 100, 110, 130, 140
  $Sampley$ = 105, 110, 125, 150
  $JoinedSamples$ = 100, 105, 110, 110, 125, 130, 140, 150
  $Ranks$ = 1, 2, 3, 4, 5, 6, 7, 8 → 1, 2,
  (3 + 4)
  2
  , 5, 6, 7, 8
  → 1, 2, 3.5, 3.5, 5, 6, 7, 8
  wx = 1 + 3 + 6 + 7 = 16
  wy = 2 + 4 + 5 + 8 = 19

- Now we find the u values of samples(ux and uy).

  ux=wx-((nx*(nx+1))/2)= 16-((4*(4+1))/2)=6

  uy=13

  min(ux,uy)=6

EXAMPLE

For example, we want to test if emissions of automatic cars are different than those of the manual cars:

So, we need to obtain only the data from automatic and manual. CHANGE == A & M.

```
selection_A <- DATA$CHANGE == "A"
selection_M <- DATA$CHANGE == "M"
selection <- selection_A | selection_M

wilcox.test(DATA$EMISION[selection] ~ DATA$CHANGE[selection])
#Esto es como siempre, la variable numerica a la izquierda y
#la categórica a la derecha, pero solo con los valores que no
#hacen falta. OBtenemos esto:

Wilcoxon rank sum test with continuity correction
data: DATA$EMISION[selection] by DATA$CHANGE[selection]
W = 1089082, p-value = 2.287e-14
alternative hypothesis: true location shift is not equal to
0
```

Podemos decir que sus medianas no son estadísticamente iguales porque la p < 0.05

Añado esto porque dice que podría ser importante para un examen tenerlo automatizado y no perder tiempo. A saber.
Es como para tenerlo más claro la diferencia entre las medianas.

▼ ¿Qué es el IQR?

El IQR (rango intercuartílico, por sus siglas en inglés) es una medida de dispersión que se utiliza en estadística para describir la variabilidad de un conjunto de datos. Se calcula como la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1) de un conjunto de datos ordenados.

El IQR se calcula de la siguiente manera:

1. Ordena los datos de menor a mayor.
2. Encuentra el primer cuartil (Q1), que es el valor que deja el 25% de los datos por debajo de él.
3. Encuentra el tercer cuartil (Q3), que es el valor que deja el 75% de los datos por debajo de él.
4. Calcula el rango intercuartílico (IQR) como la diferencia entre Q3 y Q1: IQR = Q3 - Q1.

```
knitr::kable(data.frame(group = c("Automatic", "Manual"),
 median = c(median(DATA$EMISION[selection_A]),
median(DATA$EMISION[selection_M])),
IQR = c(IQR(DATA$EMISION[selection_A]),
IQR(DATA$EMISION[selection_M]))))
```