Lab09 - Correlation and Agreement

Sometimes we do not care so much about the central tendency (means and medians) and we only want to know if there is an association between both samples.

The **correlation** measures the strength of association between two numeric variables as well as the direction. The correlation value lies between -1 and 1, the value represents what happends to one variable when the other is increased or decreased.

• 0: if the value is 0 then that means that there is **no association**, and they are **independent**

• (0 , 1): between 0 and 1 the correlation is said to be positive and it means that **if one of the variables increases the other does the same.** If one decreases the other follows suit.

• (-1 , 0): between -1 and 0 the **correlation is said to be inverse** and it means that if one of the variables increases the other decreases. If one decreases the other increases.

STRICTLY 1 = perfect positive correlation

STRICTLY -1 = perfect inverse correlation

Names for each range of correlation, for positive and inverse correlations:

- (|0,0.4|]: very weak
- (|0.4,0.6|]: weak
- (|0.6,0.8|]: moderate
- (|0.8,0.9|]: strong
- (|0.9,1|): very strong
- |1|: perfect

💡 In this type of test, the p-value will not so important as before, but the correlation coefficient. For example:

A correlation coefficient of 0.1 will be pretty much negligible whatever the p-value is. It might be statistically significant but it sure is irrelevant.

▼ Pearson's Correlation.

We use this to determine if there is an association between two random variables with normal distribution. Two random variables following normal dists.

- Pearson's correlation coefficient for a population is

$$Pxy = cov(X, Y)/(oxoy)$$

Cov: covariance, its the expected value/ mean of the product of deviation of observations of each variable from expected values.

Ox: standard deviation of X

Oy: standard deviation of Y

$$Pxy = E((X - ux)(Y - uy))/(oxoy)$$

E: expected value

Ux and Uy the means.

Pearson's correlation in R is a combination of 2 commands.
• cor gives the correlation coefficient for a given set of variables
• cor.test returns the p.value of that comparison.

The hypothesis are:

$H0$: "$thereisnoassociation$" → $r \approx 0$
$H1$: "$thereisassociation$" → $r \neq 0$

▼ Example

Using the data base mtcars we could consider if there is an association between the time the cars take to run the 1/4 mile and their weight (in 1000 lb)

We compute the correlation using cor. Read through the help of the command and make sure you understand what the use parameter do

▼ ¿Cómo funciona use?

It is an optional character string giving a method for computing covariances in the presence of missing values.

-If use is "everything", NAs will propagate conceptually, i.e., a resulting value will be NA whenever one of its contributing observations is NA.
-If use is "all.obs", then the presence of missing observations will produce an error.

-If use is "complete.obs" then missing values are handled by casewise deletion (and if there are no complete cases, that gives an error).
-"na.or.complete" is the same unless there are no complete cases, that gives NA.

- Finally, if use has the value "pairwise.complete.obs" then the correlation or covariance between each pair of variables is computed using all complete pairs of observations on those variables. This can result in covariance or correlation matrices which are not positive semi-definite, as well as NA entries if there are no complete pairs for that pair of variables. For cov and var, "pairwise.complete.obs" only works with the "pearson" method. Note that (the equivalent of) var(double(0), use = *) gives NA for use = "everything" and "na.or.complete", and gives an error in the other cases.

```
cor(mtcars$qsec, mtcars$wt, use = "complete.obs")
#El orden de los samples da igual
```

Vemos que nos da -0.1747159, esto es como EL NIVEL DE CORRELACIÓN **QUE HABRÍA SI HAY CORRELACIÓN, PERO...:**

> 💡 So the correlation coefficient points to a negative and very weak correlation. IF they are associated or not it is still unknown and we must know.
> One thing is dismissing the association because the correlation coefficient is too weak and
> a different thing is to identify if this association exists or not.
> Now we will determine if the association
> **exists or not** using **cor.test**

```
cor.test(mtcars$qsec, mtcars$wt, use = "complete.obs")
#El orden de los samples da igual
```

We can see that the p-value is 0.3389 and therefore well above 0.05, meaning that we must accept the null hypothesis. There is no correlation.

▼ **Spearman's Correlation**. Any or both of the variables are not normal

The Spearman's correlation computes the correlation between the rank of x and the rank of y.

Then we could use the formula to find the **rho value.**
We will not do it by hand, but it's important to understand what is the command doing so we can interpret the results accordingly.

The commands for Spearman's correlation are the same as for Pearson's, but we include the method parameter with the corresponding "spearman" to choose the correct one.
Spearman's correlation coefficient is called Spearman's rho.

▼ Example

Using mtcars if we want to find out if there is an association between horse-power (hp)
and miles per galon (fuel consumption) we begin by describing the variables and since we
find out that hp is not normal we describe them accordingly

```
knitr::kable(data.frame(var = c("mpg", "hp"), median =
c(median(mtcars$mpg,
na.rm = T), median(mtcars$hp, na.rm = T)), IQR = c(IQR(mtc
na.rm = T), IQR(mtcars$hp, na.rm = T)), p.norm =
c(shapiro.test(mtcars$mpg)$p.value,
shapiro.test(mtcars$hp)$p.value)))
#Esto es solo como antes para tener una idea de que
#es cada cosa, se ponen la mediana e IQR ya que no son
#normales. Si lo fueran ( los dos) se pondrían la sd y mea
cor(mtcars$mpg, mtcars$hp, method = "spearman")
rho = -0.8946646
```

We can see that the correlation is strong and negative meaning that when one of the variables go up the other goes down. The higher the horse-power of a car the less miles can make with a galon of fuel. Higher power means higher fuel consumption.

Nevertheless, we need the p_value to know if the H0 is held.

```
cor.test(mtcars$mpg, mtcars$hp, method = "spearman")
```

The p is < 0.05 so there is A CORRELATION.

But we had a warning about ties, so let's use a test that handles ties better: Kendall tau

▼ Kendall tau rank Correlation. Same as Spearman but better with ties.

WARNING
There are three Kendall tau statistics:
- tau-a
- tau-b
- tau-c

tau-b is specifically adapted to handle ties

Usually we get the correlation coefficient first and the p-value later, but since Kendall tau-b is better implemented (as per R documentation) in the Kendall package we will use that package and get both the correlation coefficient and p-value at the same time

```
Kendall::Kendall(mtcars$mpg, mtcars$hp)
tau = -0.743, 2-sided pvalue =4.7775e-09
```

Podemos ver que tau (”rho”) es distinta a la de antes. Aun así hay una correlación ($p<0.05$) y es bastante fuerte.