Bivariate data analysis

Josep L. Carrasco
Bioestadística. Departament de Fonaments Clínics
Universitat de Barcelona

Setting

Two ways of identifying the setting:

1) One variable is measured in two samples.

For example, is there a difference in blood pressure between men and women?

Variable: blood pressure.

Samples: One sample of men and other of women.

2) Two variables are measured in one sample.

For example, is the failure of a treatment related to smoking?

Variables: Failure/success of treatment, smoking.

Sample: One sample from the aimed population.

In any case the analysis will involve assessing the association between two variables.

The statistical methodology will depend on the nature of the two variables (qualitative/quantitative).

Type of variables	Analysis
Both qualitative	Homogeneity of proportions
One quantitative and one qualitative	Homogeneity of means
Both quantitative	Correlation

Moreover, the analysis will vary depending on whether data design is **independent** or **paired**.

- **Independent data**.

Suppose two populations A and B.

Data from population A do not bring any information about data from population B.

Every subject is measured once.

- **Paired data**.

Data from population A are conditioned by data from population B (or B conditioned by A).

Every subject is measured twice (at least).

Usually the power is large with paired data.

Limitation: it is not always possible to develop a paired data design.

Example.

- A study wants to compare the blood concentration means of two drug formulations.
- Independent data

1st sample: n_A subjects receiving formulation A.

2st sample: n_B subjects receiving formulation B.

Every subject takes A **or** B formulations

- Paired data

1st sample: n subjects receiving formulation A.

2st sample: After the drug has been excreted, the same subjects receive formulation B.

Every subject takes A **and** B formulations

Homogeneity of proportions

Independent data

Example. Set of 100 subjects that suffered pleural effusion. Some of these patients required draining the effusion. It is desired to check if the need of draining is related to subjects' gender (what is equivalent to test if the probability of draining is different for each gender).

1) Statistical hypothesis

$$H_0 : \pi_M = \pi_F$$
$$H_1 : \pi_M \neq \pi_F$$

where π_M and π_F stand for the probability of draining for male and female.

2) Statistical test.

- Statistic

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{\left(O_{ij} - E_{ij}\right)^2}{E_{ij}}$$

where i and j are the row and column identifiers.

O_{ij}: observed frequencies at cell i,j. E_{ij}: expected columns, if H_0 is true, at cell i,j.

- Distribution under H_0: Chi-square with $(r-1)(c-1)$ degrees of freedom.
- Applicability conditions: Expected frequencies equal or greater than 5.

Let's load the data

```
fib<-read.table("fibrinol.txt",header=T)
```

Data is coded as follows:

- Gender. "Sexo" in data set. 1="Female"; 2="Male"
- Draining. "Drenaje" in data set. 1="Yes"; 2="No"
- Describing data

```
tab<-table(fib$sexo,fib$drenaje,dnn=list("Gender","Drain"))
cat("Counts")
```

```
Counts
```

```
addmargins(tab)
```

```
      Drain
Gender   1   2 Sum
   1    38  17  55
   2    40   5  45
   Sum  78  22 100
```

```
cat("\n")
```

```
cat("Proportions")
```

```
Proportions
```

```
prop.table(tab,margin=1)
```

```
      Drain
Gender         1         2
     1 0.6909091 0.3090909
     2 0.8888889 0.1111111
```

- Applicability conditions

```
chisq.test(fib$drenaje,fib$sexo)$expected
```

```
           fib$sexo
fib$drenaje    1    2
          1 42.9 35.1
          2 12.1  9.9
```

All expected counts are greater than 5.

- Test

```
chisq.test(fib$drenaje,fib$sexo)
```

```

    Pearson's Chi-squared test with Yates' continuity correction

data:  fib$drenaje and fib$sexo
X-squared = 4.5584, df = 1, p-value = 0.03276
```

If the type-I error level is set to $\alpha = 0.05$, the null hypothesis is rejected. The proportion of draining is different according to gender.

Males tend to have a higher probability, but how much high? To answer this question we have to assess the **magnitude of the difference** or the **magnitude of the association**.

Magnitude of the difference

To assess the magnitude of the difference let's estimate a 95% confidence interval for the difference of proportions.

Note: Remember that here the applicability conditions are that the observed counts must be greater than 5.

```
prop.test(c(38,40),c(55,45))
```

```

    2-sample test for equality of proportions with continuity correction

data:  c(38, 40) out of c(55, 45)
X-squared = 4.5584, df = 1, p-value = 0.03276
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.37097836 -0.02498124
sample estimates:
   prop 1    prop 2
0.6909091 0.8888889
```

The proportion of draining is higher in males. With a confidence of 95%, the difference of proportions is within $[0.025, 0.371]$.

Magnitude of the association

Alternatively, when analyzing two qualitative variables, researchers may wish to give a measure of association rather than a difference.

Most common index of association is the **odds ratio** (see Notes 1.1).

Remember that an odds is the ratio between a probability and its complementary. In this case the probability of draining.

We may compute two odds, one for men and other for women.

The odds ratio will be the ratio of such odds.

Interpretation:

Let's suppose that we put the odds for men on the numerator and that for women in the denominator.

- OR > 1. Draining is associated with the fact of being male (level in the numerator)
- OR = 1. Independence. There is no association between draining and gender.
- OR < 1. Draining is associated with the fact of being female (level in the denominator)

The value of the OR can be understood as how much times is more associated.

Note that our table was ordering as first column "Draining".

```
tab
```

```
      Drain
Gender  1  2
     1 38 17
     2 40  5
```

That will cause that the probability involved in the OR will be "draining".

However, the first row is "Female", so the odds in the OR's numerator will be such of women. Since we have seen that the proportion of draining in males was higher it will be more convenient to put the males' probability on the numerator for the sake of interpretation.

So that, let's change the order of rows.

```
tab[2:1,]
```

```
      Drain
Gender  1  2
     2 40  5
     1 38 17
```

To compute the OR and its 95% confidence interval let's use the *DescTools* package.

```
library(DescTools)
OddsRatio(tab[2:1,],conf.level = 0.95)
```

```
odds ratio     lwr.ci     upr.ci 
  3.578947   1.201489  10.660824 
```

The odds of draining is 3.6 times higher in Males than in Females. Threfore, the necessity of draining is 3.6 times more associated to Males.

Fisher's exact test

It is applied when the applicability conditions for the chi-square test are not met.

It is computationally more intensive.

Suppose now that data was 20 females of which 15 required draining. Additionally 9 out of 10 males needed to be drained.

```
Counts


      Drain
Gender   1   2 Sum
   2     9   1  10
   1    15   5  20
   Sum  24   6  30

Proportions


      Drain
Gender    1    2
     2 0.90 0.10
     1 0.75 0.25
```

The expected counts are now:

```
chisq.test(sexo,drenaje)$expected[2:1,]
```

```
    drenaje
sexo  1 2
   2  8 2
   1 16 4
```

Some of the expected counts are lower than 5. The applicability conditions for the chi-square test are not met.

Let's use the Fisher's exact test to get the p-value.

```
fisher.test(tab)
```

```

    Fisher's Exact Test for Count Data

data:  tab
p-value = 0.6328
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.2595879 157.2640953
sample estimates:
odds ratio
  2.905401
```

In this case the p-value is not lower than 5% so the null hypothesis of independence is not rejected. We could not demonstrate a relation between *draining* and *gender*.

Fuatheremore, notice that the odds ratio is also in the output wit a point estimate of 2.9 but with a extremely wide confidence interval.

Limitations: computationally intensive and low power.

Sample size

One key point when designing a research is to compute the number of subjects to include in the sample. To compute the sample size we will use the function *pwr.2p.test(h,power)* from *pwr* package.

- *h*. The effect size. It can be computed usign the function *ES.h(p1, p2)* where:
  - *p1*. Proportion for group 1.
  - *p2*. Proportion for group 2, or equivalently the proportion for group 1 plus the **difference to detect**. That answers the question: from what value do we consider the difference between proportions to be clinically important.
- *Power*: The power of test.

Example. How many subjects should be sampled to demonstrate the association between draining and gender?

- *p1*. The proportion is unknown so we have to guess it. In the analysis we saw that the sample proportion of draining in women was 0.69. So it could have sense to use a value around this estimate as, for instance, 0.7.

- *p2*. Which difference do we wish to detect? That means, from what value we consider the difference between proportions clinically important. We could consider two values and compare the results. For example, 0.05 and 0.1.

- *Power*. Usually the power is set to values higher than 0.7. Let's set it at 70% and 80% so we will be able to compare the results.

```
library(pwr)
# Difference 0.05. Power 0.7
h=ES.h(0.75,0.7)
pwr.2p.test(h=h,power=0.7)
```

```

     Difference of proportion power calculation for binomial distribution (arcsine trans

              h = 0.1120819
              n = 982.6186
      sig.level = 0.05
          power = 0.7
    alternative = two.sided

NOTE: same sample sizes
```

```
# Difference 0.1. Power 0.7
h=ES.h(0.8,0.7)
pwr.2p.test(h=h,power=0.7)
```

```

     Difference of proportion power calculation for binomial distribution (arcsine trans

              h = 0.2319843
              n = 229.3713
      sig.level = 0.05
          power = 0.7
    alternative = two.sided

NOTE: same sample sizes
```

```
# Difference 0.05. Power 0.8
h=ES.h(0.75,0.7)
pwr.2p.test(h=h,power=0.8)
```

```
     Difference of proportion power calculation for binomial distribution (arcsine trans

              h = 0.1120819
              n = 1249.584
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: same sample sizes
```

```
# Difference 0.1. Power 0.8
h=ES.h(0.8,0.7)
pwr.2p.test(h=h,power=0.8)
```

```
     Difference of proportion power calculation for binomial distribution (arcsine trans

              h = 0.2319843
              n = 291.6887
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: same sample sizes
```

More than two proportions

In some cases it is possible that more than two proportions have to be compared.

Data is stored in a $r \times c$ contingency table.

Example. Are the blood group frequencies (A,B,AB,O) the same in 3 populations?

Blood group: 4 categories.

Population: 3 categories.

- Observed data

```
X<-matrix(c(568,152,31,1058,272,102,0,951,902,278,66,2611),ncol=4,byrow=T)
colnames(X)<-c("A","B","AB","O")
rownames(X)<-c("1","2","3")
X
```

```
    A   B AB    O
1 568 152 31 1058
2 272 102  0  951
3 902 278 66 2611
```

1) Statistical hypotheses

$$
\begin{aligned}
&H_0 : \\
&(\pi_{1A} = \pi_{2A} = \pi_{3A} = \pi_A) \\
&\cap (\pi_{1B} = \pi_{2B} = \pi_{3B} = \pi_B) \\
&\cap (\pi_{1AB} = \pi_{2AB} = \pi_{3AB} = \pi_{AB}) \\
&\cap (\pi_{1O} = \pi_{2O} = \pi_{3O} = \pi_O)
\end{aligned}
$$

$$H_A : \pi_{ij} \neq \pi_{i'j}$$

2) Statistical test: Chi-Square test.

- Applicability conditions

```
chisq.test(X)$expected
```

```
         A        B       AB         O
1 450.7621 137.6610 25.09984 1195.4770
2 330.1602 100.8296 18.38435  875.6258
3 961.0777 293.5094 53.51581 2548.8972
```

All the expected counts are greater than 5.

- Test

```
chisq.test(X)
```

```

    Pearson's Chi-squared test

data:  X
X-squared = 93.19, df = 6, p-value < 2.2e-16
```

The P-value is lower than 5% (common value for type-I error rate), so the null hypothesis of equality of proportions is rejected. So that, **some of the proportions are different, but which ones?**

When comparing two proportions this question is not necessary because the proportions are complementary, so if one increases the other decreases.

This rule does not work with more than two proportions because there are more possibilities. For example, in the case of three proportions, one proportion could increase and the two remaining proportion decrease, or just one could decrease and the other remain steady.

To answer this question it is necessary to carry out an additional analysis: the **residual analysis**.

Note: a residual analysis only makes sense if the **null hypothesis is rejected**.

We are going to use the **Pearson's residuals** (r_{ij}).

$$r_{ij} = \frac{O_{ij} - E_{ij}}{\sqrt{E_{ij}}}$$

A value of $|r_{ij}| > 2$ will be considered as a high residual: the deviation from H_0 in this cell is large.

```
chisq.test(X)$residuals
```

```
          A          B        AB         O
1  5.521972  1.2221194  1.177682 -3.976121
2 -3.200837  0.1165537 -4.287698  2.547202
3 -1.905654 -0.9052805  1.706552  1.230086
```

- The sign of the residuals indicates if there are more or less observed than expected counts (positive or negative sign respectively).
- Population 1 has a larger proportion of type "A" than expected and lower of type "O".
- Population 2 has larger proportion of type "O" and lower of types "A" and "AB".
- Population 3 has no conclusive results.

Sample size

To compute the sample size we will use the function *pwr.chisq.test(w,power,df)* from *pwr* package. The arguments are:

- *w.* The effect size.

$$w = \sqrt{\frac{\chi^2}{n \cdot df}}$$

It is commonly accepted that a value of 0.1 stands for a small effect, a value of 0.3 means medium effect, and a value of 0.5 involves a large effect.

- *power.* The desired power of the test.

- *df*. The degrees of freedom of the chi-square test.

Using the setting of the blood groups let us compute the sample size necessary to detect a small effect with a power of 0.8.

```
library(pwr)
pwr.chisq.test(w=0.1,power=0.8,df=6)
```

```

     Chi squared power calculation 

              w = 0.1
              N = 1362.429
             df = 6
      sig.level = 0.05
          power = 0.8

NOTE: N is the number of observations
```

The result is 1363 subjects.

Paired data

Example. A sample of 51 subjects with pleural effusion that were assessed for the presence of infection in the pleural liquid using two different approaches (A and B).

- Load the data.

```
infec<-read.table("infeccion.txt",header=T)
```

The aim is to check if the proportion of a positive result (value 1 in variables), i.e. presence of infection, is different depending on the approach applied.

- Statistical hypothesis

$$\begin{aligned} H_0 &: \pi_A = \pi_B \\ H_1 &: \pi_A \neq \pi_B \end{aligned}$$

Where π_A and π_B stand for the probability of infection in approaches A and B respectively.

- Notice that every subject has a reading with every approach –> **paired data**.

Describing data

```
tab<-table(infec$prueba_a,infec$prueba_b,dnn=list("Test A","Test B"))
cat("Counts")
```

```
Counts
```

```
addmargins(tab)
```

```
      Test B
Test A  1  2 Sum
   1   19 16  35
   2    1 15  16
   Sum 20 31  51
```

```
cat("\n")
```

```
cat("Proportions")
```

```
Proportions
```

```
addmargins(prop.table(tab))
```

```
      Test B
Test A          1          2        Sum
   1   0.37254902 0.31372549 0.68627451
   2   0.01960784 0.29411765 0.31372549
   Sum 0.39215686 0.60784314 1.00000000
```

Applicability conditions

The expected counts have to be greater than 5, or equivalently the sum of the observed counts on the frequency table diagonal have to be greater than 10.

$n_{10} + n_{01} = 16 + 1 = 17 > 10$

Test

The test to assess an homogeneity of two proportions with paired data is known as **McNemar's test**. The R function to run this test is *mcnemar.test(x)* where

- x: two-dimensional contingency table or the factor variables separated by a comma.

```
mcnemar.test(infec$prueba_a,infec$prueba_b)
```

```
    McNemar's Chi-squared test with continuity correction

data:  infec$prueba_a and infec$prueba_b
McNemar's chi-squared = 11.529, df = 1, p-value = 0.000685
```

P-value is lower than 5%, so that the proportion of positives is different depending on the approach.

Magnitude of the difference

To assess the magnitude of the difference let's compute a 95% confidence interval for the difference of proportions with paired data. To do so we may use the *diffpropci.mp(b, c, n, conf.level)* function from *PropCIs* package. The arguments are:

- *b*: off-diag count
- *c*: off-diag count
- *n*: sample size
- *conf.level*: confidence level

```
library(PropCIs)
diffpropci.mp(tab[2,1], tab[1,2], sum(tab), conf.level=0.95)
```

```



data:

95 percent confidence interval:
 0.1458682 0.4201696
sample estimates:
[1] 0.2830189
```

The proportion of positives is larger when using test A. With a confidence of 95% the difference is within $[0.146, 0.420]$.

Magnitude of the association

The Odds Ratio for a 2x2 contingency table with paired data is somewhat different to that of independent data, but easier to compute.

$$OR = \frac{b}{c}$$

where b and c are the off-diagonal counts. The confidence interval can be computed by applying the **oddsratioci.mp** function.

```
b<-tab[1,2]
c<-tab[2,1]
cat("Odds Ratio")
```

```
Odds Ratio
```

```
b/c
```

```
[1] 16
```

```
oddsratioci.mp(tab[2,1], tab[1,2], conf.level=0.95)
```

```
data:

95 percent confidence interval:
  2.70617 94.59863
```

Sample size

To compute the sample size for an homogeneity of proportions with paired data let us create an ad-hoc function:

```
samsize.mcnemar <- function(pi.01, pi.10, alpha=0.05, beta=0.2, sided=2)
{
  pi.d <- (pi.01 + pi.10)
  N <- (qnorm(1 - alpha/sided) * sqrt(pi.d) + qnorm(1 - beta) *
          sqrt(pi.d - (pi.01 - pi.10)^2))^2/(pi.01 - pi.10)^2
  return(ceiling(N))
}
```

The argument of this function are:

- *pi.01* and *pi.10* are the off-diagonal proportions.
- *alpha* is the level of significance. The default value is set to 0.05.
- *beta* is the probability of type-II error, i.e. the opposite of the power. Default value is set to 0.2.
- *sided*. Is the the test one-sided (value of 1) or two-sided (value of 2). Default value is set to two-sided.

In the example we saw the off-diagonal proportions were around 0.3 and 0.02. In a case like that, what would it be the sample size necessary to reject the hypothesis with a power of 0.8? (Remember execute the function code to load it).

```r
samsize.mcnemar(pi.01 = 0.3, pi.10 = 0.02)
```

```
[1] 30
```

The result is 30 subjects.

Homogeneity of means

Independent data

Example. To analyze the factors influencing newborns' weight one sample of 105 newborns from primiparous mothers, and other sample of 551 newborns from multiparous mothers. The variable to analyze is the weight at birth in Kg that is assumed to follow a Normal distribution at each population.

- Hypotheses

$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 = \mu \\ H_A: &\quad \mu_1 \neq \mu_2 \end{aligned}$$

A one-sided test would be also possible.

- Statistical test: two options depending on the equality of variances.

Therefore, we need to compare the variances before testing the difference of means

$$\begin{aligned} H_0: &\quad \sigma_1^2 = \sigma_2^2 = \sigma^2 \\ H_A: &\quad \sigma_1^2 \neq \sigma_2^2 \end{aligned}$$

- Applicability conditions:
  - X follows a Normal distribution in both populations **OR**
  - Sample sizes are large ($n_1 \geq 30$ **and** $n_2 \geq 30$).

Homogeneity of variances

- If the analyzed variable X follows a Normal distribution **in both populations**: F-test
- If X does not follow a Normal distribution **in some population**: Fligner-Killeen test.

Let's load the example data.

```r
pes<-read.table("pesnen.txt",header=T,sep="\t")
head(pes)
```

```
  paridad pes.nen
1       7     1.3
2       2     1.4
3       2     1.5
4       7     1.8
5       1     1.8
6       1     1.8
```

The variable *paridad* indicates the count of labours. Let's transform it to a factor showing if the mother is primiparous or not.

```r
library(dplyr)
pes$rpar<-factor(pes$paridad==1)
```

Let's describe the two samples:

```r
by(pes$pes.nen,pes$rpar,mean)
```

```
pes$rpar: FALSE
[1] 3.040835
------------------------------------------------------------
pes$rpar: TRUE
[1] 2.90381
```

```r
by(pes$pes.nen,pes$rpar,sd)
```

```
pes$rpar: FALSE
[1] 0.4201974
------------------------------------------------------------
pes$rpar: TRUE
[1] 0.4274081
```

```r
by(pes$pes.nen,pes$rpar,length)
```

```
pes$rpar: FALSE
[1] 551
------------------------------------------------------------
pes$rpar: TRUE
[1] 105
```

- Check the homogeneity of variances

In the heading of the example it has been said that the weight at birth can be assumed to follow a Normal distribution. Therefore we will use the F-test to check the homogeneity of variances. The function to apply is *var.test(x,y)* where x and y are the numeric vectors with the data values.

```
x1<-pes$pes.nen[pes$rpar=="TRUE"]
x2=pes$pes.nen[pes$rpar=="FALSE"]
var.test(x1,x2)
```

```

    F test to compare two variances

data:  x1 and x2
F = 1.0346, num df = 104, denom df = 550, p-value = 0.7948
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.780061 1.414780
sample estimates:
ratio of variances
          1.034615
```

The p-value is greater than 5%, so we could assume that the variances are similar.

If it was not possible to assume normality of data, the function *fligner.test(x,g)* runs the Fligner-Kileen test. The function arguments are:

- x: numeric vector of data values;
- g: a vector or factor object giving the group for the corresponding elements of x.

In the case example:

```
fligner.test(pes$pes.nen,pes$rpar)
```

```

    Fligner-Killeen test of homogeneity of variances

data:  pes$pes.nen and pes$rpar
Fligner-Killeen:med chi-squared = 0.17917, df = 1, p-value = 0.6721
```

Next, let's test the homogeneity of means. To do that we will use the function *t.test(x,y,var.equal)* where:

- x and y are the numeric vectors with the data values.
- *var.equal*: a logical variable indicating whether to treat the two variances as being equal.

```
t.test(x1,x2,var.equal=T)
```

```

    Two Sample t-test

data:  x1 and x2
t = -3.054, df = 654, p-value = 0.00235
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.22512603 -0.04892461
sample estimates:
mean of x mean of y 
 2.903810  3.040835 
```

The p-value is lower than 5% so the null hypothesis of equality of means is rejected.

The interval that appears in the output is the 95% confidence interval for the difference of means, so we may use this interval as an indicator of the **magnitude of the difference**.

Sample size

To compute the sample size we will use the function *pwr.t.test(delta,sd,power)* from *pwr* package. The function has the following arguments:

- *d*. Effect size. Difference between the means divided by the pooled standard deviation:

$$d = \frac{\mu_1 - \mu_0}{s}$$

- *Power*: The aimed power of the test.

For example, how many newborns should be analyzed to find as significant a difference of 0.2 Kg if the standard deviation was 0.42 Kg with a power of 80%.

```
d=0.2/0.42
pwr.t.test(d=d,power=0.8)
```

```

     Two-sample t test power calculation 

              n = 70.20035
              d = 0.4761905
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group
```

The results is 71 newborns by group.

Alternative analysis

How can we proceed if applicability conditions do not meet?

That means, the variable does not follow a Normal model **and** the sample size is low (lower than 30).

Two options:

1) Permutation test.
2) Non-parametric analysis: Mann-Whitney test.

Example. In a research on cephalea 8 subjects were treated with a drug, say A, whereas other 10 subjects where treated with placebo. One hour after receiving the treatment, the 18 subjects evaluated their pain relief in a visual analogue scale (VAS) of 10cm. A value of 0 means "no improvement", a value of 10 means "complete improvement". Let's suppose that it is not possible to assume Normality of the variable.

The results were

```
tractament=c(rep("A",8),rep("P",10))
vas=c(8.3,9.1,6.2,5.4,8.3,6.5,8.4,7.5,3.1,5.6,4.5,6.2,5.1,5.3,5.5,4.1,4.3,4.2)
dades=data.frame(vas,tractament)
dades
```

```
   vas tractament
1  8.3          A
2  9.1          A
3  6.2          A
4  5.4          A
5  8.3          A
6  6.5          A
7  8.4          A
8  7.5          A
9  3.1          P
10 5.6          P
11 4.5          P
12 6.2          P
13 5.1          P
14 5.3          P
15 5.5          P
16 4.1          P
17 4.3          P
18 4.2          P
```

Permutations test

- Aim: to compute the p-value by simulating the distribution under H_0.
- No parametric model is assumed.

Steps:

1) Choose the appropriate test statistic.

For example, the difference of sample means.

```
theta=mean(vas[1:8])-mean(vas[9:18])
theta
```

```
[1] 2.6725
```

2) Generate the distribution of the test statistic under H_0.

The hypotheses to test still are:

$$\begin{array}{rl} H_0: & \mu_1 = \mu_2 = \mu \\ H_A: & \mu_2 \neq \mu_2 \end{array}$$

A permutation is just the random switch of data between the two samples maintaining the design structure.

Original data

```
t(dades)
```

```
           [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]  [,10] [,11]
vas        "8.3" "9.1" "6.2" "5.4" "8.3" "6.5" "8.4" "7.5" "3.1" "5.6" "4.5"
tractament "A"   "A"   "A"   "A"   "A"   "A"   "A"   "A"   "P"   "P"   "P"  
           [,12] [,13] [,14] [,15] [,16] [,17] [,18]
vas        "6.2" "5.1" "5.3" "5.5" "4.1" "4.3" "4.2"
tractament "P"   "P"   "P"   "P"   "P"   "P"   "P"  
```

One permutation. The labels of the treatment are permuted.

```
t(cbind(dades$vas,sample(dades$tractament)))
```

```
     [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9]  [,10] [,11] [,12]
[1,] "8.3" "9.1" "6.2" "5.4" "8.3" "6.5" "8.4" "7.5" "3.1" "5.6" "4.5" "6.2"
[2,] "P"   "P"   "P"   "P"   "P"   "A"   "P"   "A"   "P"   "P"   "A"   "A"  
     [,13] [,14] [,15] [,16] [,17] [,18]
[1,] "5.1" "5.3" "5.5" "4.1" "4.3" "4.2"
[2,] "A"   "A"   "P"   "A"   "P"   "A"  
```

The rationale behind is it does not matter if data comes from "A" or "P" because they come from populations with the same mean. By permuting the labels we mixed up the data, so all the data come from the same distribution model with a common mean, therefore we are generating data under the null hypothesis.

Next let's create a function that generates a permutation and computes the difference of means.

```
dif.mean <- function(x,y) {
  xstar<-sample(x)
  mean(y[xstar=="A"])-mean(y[xstar=="P"])
}
```

Run this function a huge number of times, for example 1000 times.

```
perm=replicate(1000, dif.mean(tractament, vas))
```

3) Compute the p-value using the data simulated in step 2.

The P-value is computed as the proportion of differences that are more extreme than the observed difference.

```
((sum(perm>theta)+sum(perm<(theta*-1))))+1/(1000+1)
```

```
[1] 0.000999001
```

A value of 1 is added in both numerator and denominator. The reason of that is we have 1000 estimates from the permutations plus one estimate from the original sample.

Mann-Whitney test

The Mann-Whitney test (also known as Wilcoxon–Mann–Whitney test or wilcoxon rank sum test) is a non-parametric test. It is probably the most applied alternative to t-test.

It is based on ranking the data and comparing the sum of ranks of each group. It is commonly accepted that the test is about equality of **medians**. However, one limitation is the magnitude of the difference of means is missing.

The R function to run the Mann-Whitney test is *wilcox.test(x,y)* where x and y are the numeric vectors of data values.

```
x1<-dades[dades$tractament=="A",]$vas
x2<-dades[dades$tractament=="P",]$vas
wilcox.test(x1,x2)
```

```

    Wilcoxon rank sum test with continuity correction

data:  x1 and x2
W = 76.5, p-value = 0.001365
alternative hypothesis: true location shift is not equal to 0
```

Alternatively the function may used a **formula** with the variables:

```
wilcox.test(vas~tractament, dades)
```

```

    Wilcoxon rank sum test with continuity correction

data:  vas by tractament
W = 76.5, p-value = 0.001365
alternative hypothesis: true location shift is not equal to 0
```

The P-value is lower than 5%, so the null hypothesis is rejected.

Magnitude of difference

Permutations and Mann-Whitney's tests can be also complemented with the magnitude of the difference. It is possible to give a confidence interval on the difference of means using bootstrap as that from the former point.

However, when applying the Mann-Whitney's test the null hypothesis is the equality of original data medians.Therefore, in this case it makes sense to give the confidence interval on the difference of medians.

First we create the function that generates one resample and computes the difference of means (and medians).

```
theta.star.mean=function(){
x1=sample(vas[1:8],replace=T)
x2=sample(vas[9:18],replace=T)
mean(x1)-mean(x2)
}
```

```
theta.star.median=function(){
x1=sample(vas[1:8],replace=T)
x2=sample(vas[9:18],replace=T)
median(x1)-median(x2)
}
```

Next these functions are ran so many times as bootstrap samples are wished. For example 1000 times.

```
dif.boot.mean=replicate(1000,theta.star.mean())
dif.boot.median=replicate(1000,theta.star.median())
```

So that the 95% confidence interval is obtained as:

```
quantile(dif.boot.mean,probs=c(0.025,0.975))
```

```
    2.5%    97.5%
1.572438 3.647937
```

```
quantile(dif.boot.median,probs=c(0.025,0.975))
```

```
   2.5%   97.5%
1.09875 4.05000
```

Paired data

In a research study patients with vasculitis were analyzed. This pathology is characterized by the active and inactive phases of disease. Serum level of e-selectine molecule was measured in 33 patients twice, one measure by phase. The aim is to test if the serum levels of e-selectine are different at each phase of the disease.

Let's load the data. The file has two variables: *esela* for active phase and *eseli* for inactive phase.

```
esel=read.table("eselec.txt",header=T)
head(esel)
```

```
  esela eseli
1  80.6  35.0
2 100.2  82.0
3  85.6  67.0
4  90.5  55.6
5  75.9  78.6
6  91.6  19.4
```

- Hypotheses

$$\begin{aligned} H_0: &\quad \mu_1 = \mu_2 = \mu \\ H_A: &\quad \mu_1 \neq \mu_2 \end{aligned}$$

However, the design of the data is paired, every subject has two measurements of the outcome.

Let's suppose that Y_1 and Y_2 are the data vectors of each measurement, in the example the two phases.

Let's construct the variable difference $D = Y_1 - Y_2$.

Testing the homogeneity of means is equivalent to test if the mean of the difference is equal to 0.

$$\begin{aligned} H_0: &\quad \mu_D = 0 \\ H_A: &\quad \mu_D = 0 \end{aligned}$$

Thus, the homogeneity of means with paired data is actually a one-sample test on the mean of the difference.

The applicability conditions are the same as those from the one-sample test for the mean:

- The outcome (the difference in this case) has to follow a Normal distribution model.

OR

- The sample size has to be larger than 30.

```
d=esel$esela-esel$eseli
length(d)
```

```
[1] 33
```

The sample size is larger than 30, so the applicability conditions are met.

The test is carried out using the *t.test* function. We have two options:

- Use the two data column vectors and specify *paired=T* in the function.

```
t.test(esel$esela,esel$eseli,paired=T)
```

```
    Paired t-test

data:  esel$esela and esel$eseli
t = 8.3531, df = 32, p-value = 1.522e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 32.05918 52.73718
sample estimates:
mean of the differences 
               42.39818 
```

- Use the vector of the differences.

```
t.test(d)
```

```
    One Sample t-test

data:  d
t = 8.3531, df = 32, p-value = 1.522e-09
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 32.05918 52.73718
sample estimates:
mean of x 
 42.39818 
```

The p-value is lower than 5% so we conclude that the means of e-selectine are different at each phase of the disease.

Magnitude of the difference

In the former output we could see that the mean estimate of the differences was 42.4 with a 95% confidence interval of 32.06 – 52.74.

Sample size

To compute th sample size necessary to reject the null hypothesis wit a specific power we will use the function *pwr.t.test(d,power,type)* from *pwr* package. The arguments are:

- *d*. Effect size. The difference between the means divided by the standard deviation of the differences.

$$\frac{\mu_1 - mu_0}{s_d}$$

- *power*. The power of the test.

- *type*. Type of the test. The value of *paired* indicates the paired data design.

Let us suppose that it is wished to detect a difference of 10 units with a standard deviation of 30 units.

```
d=10/30
pwr.t.test(d=d,power=0.8,type="paired")
```

```
     Paired t test power calculation 

              n = 72.5839
              d = 0.3333333
      sig.level = 0.05
          power = 0.8
    alternative = two.sided

NOTE: n is number of *pairs*
```

It is necessary a sample size of 73 subjects.

Alternative analysis

A research group studied 16 men with gestational desire couples and a minimum of two altered seminograms. These patients were treated and their levels of the hormone FSH was measured before and 3 months after receiving the treatment. The aim was to find out if there are differences in FSH values before and after treatment.

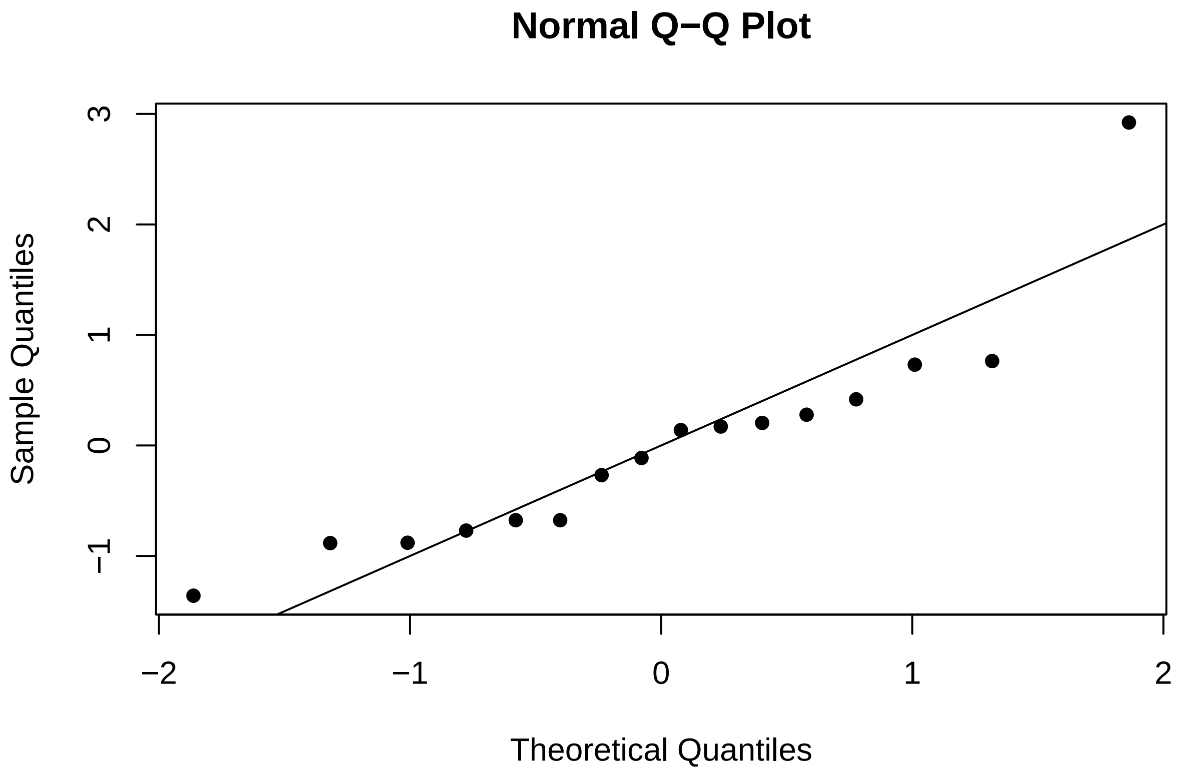
Let's load the data. The file contains two variables: *fshb* and *fshf* for the measurements before and after the treatment respectively.

```
fsh=read.table("fshdata.txt",header=T)
head(fsh)
```

```
  fshb fshf
1 3.65 6.37
2 2.21 2.21
3 1.41 3.93
4 3.92 1.81
5 1.64 5.02
6 3.65 6.60
```

Let's check the applicability conditions. The sample size is lower than 30, so we need to check the normality condition.

```
d=fsh$fshf-fsh$fshb
qqnorm(scale(d),pch=16)
abline(0,1)
```



```
shapiro.test(d)
```

```

    Shapiro-Wilk normality test

data:  d
W = 0.86464, p-value = 0.02253
```

In the Q-Q plot it can be seen that there are a couple of points that move away from the straight line of agreement. In addition the Shapiro-Wilks test gives a small p-value. Thus, we can not assume that the differences follow a Normal model. Therefore the applicability conditions are not met.

Permutations test

One option is to apply a permutation test.

Firstly, we will create a function that calculates the average of the matched differences from a permutation. Notice that the permutation must be according to the design of data, therefore the data must be permuted within the subjects. Thus, data to permute are the two data of each subject.

Actually, this is the same that randomly assigning the sign of the differences. So, we will randomly generate the sign of the difference of each individual using a Bernoulli distribution.

The code that generates these signs is:

```
sign(rbinom(16,1,0.5)-0.5)
```

```
 [1]  1 -1 -1  1 -1 -1  1  1  1  1 -1 -1  1 -1 -1 -1
```

And the function that generates a permutation of the differences and computes the mean is:

```
difpair=function(){
  theta=(fsh$fshb-fsh$fshf)*sign(rbinom(16,1,0.5)-0.5)
  mean(theta)
}
```

Now let's generate 1000 permutations

```
perm=replicate(1000,difpair())
```

Next the p-value is computed as

```
(sum(abs(perm)>abs(mean(d)))+1)/(1000+1)
```

```
[1] 0.01698302
```

which is lower than 5% so the null hypothesis of equality of means is rejected.

Magnitude of the difference

We will use bootstrap to construct a 95% confidence interval for the differences of means.

First we need to create a function that generates a bootstrap sample of the differences and computes the mean.

```
theta.star=function(){
x=sample(d,replace=T)
mean(x)
}
```

Next, this function is run so many times as bootstrap samples are desired. For example, 1000 times.

```
dif.boot=replicate(1000,theta.star())
```

To estimate the 95% confidence interval compute the quantiles 2.5 and 97.5.

```
quantile(dif.boot,probs=c(0.025,0.975))
```

```
     2.5%     97.5% 
0.7627969 3.5675937 
```

Wilcoxon's test

Another alternative is to apply the Wilcoxon's non-parametric test as we did in the one-sample block.

```
wilcox.test(d)
```

```

    Wilcoxon signed rank test with continuity correction

data:  d
V = 93, p-value = 0.01204
alternative hypothesis: true location is not equal to 0
```

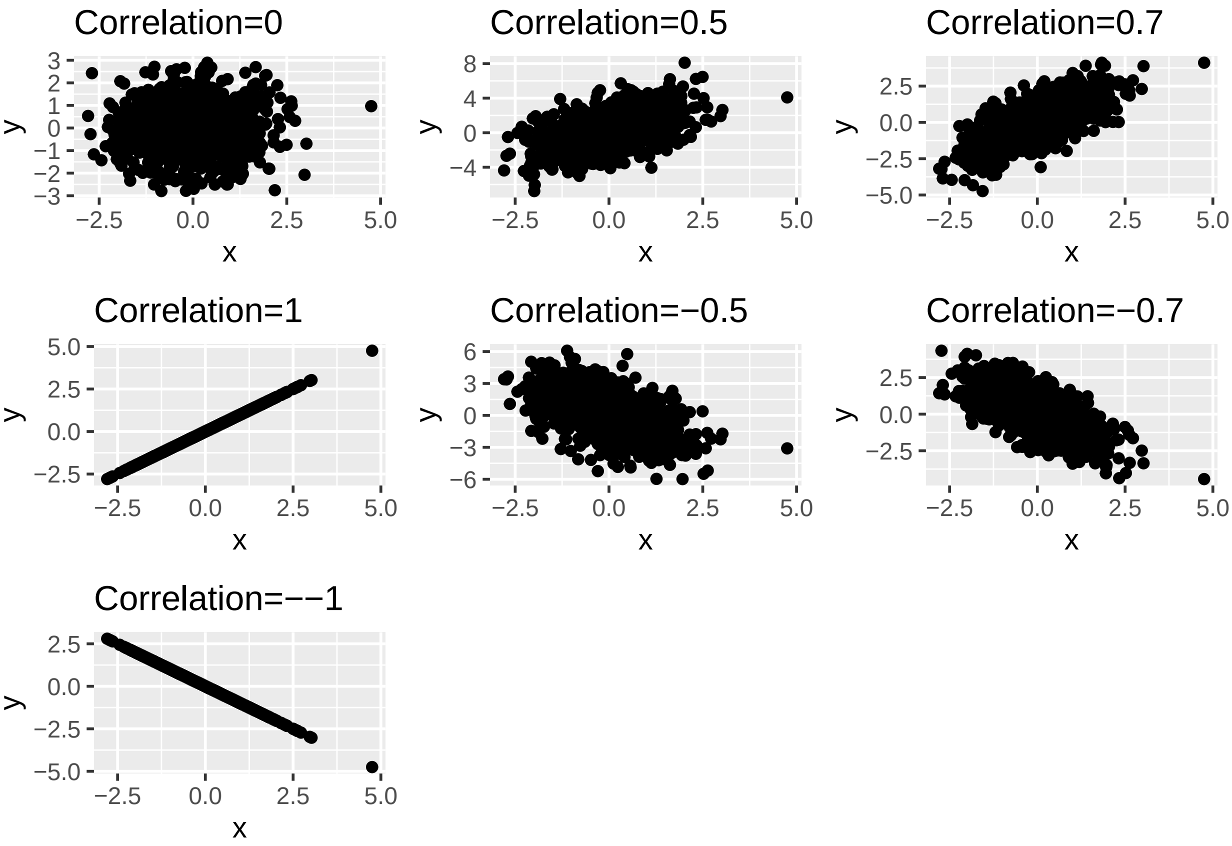
The p-value is lower than 5% so the null hypothesis is rejected.

Correlation

The degree of association or lineal correlation between two continuous variables, X and Y, is measured by the **correlation coefficient**.

The correlation coefficient takes values between -1 and 1, where:

- $-1 \rightarrow$ perfect inverse correlation.
- $0 \rightarrow$ independence
- $1 \rightarrow$ perfect direct correlation.



The most applied correlation coefficient is the **Pearson's correlation coefficient**. There are two functions in R to estimate the **Pearson's correlation coefficient**:

- *cor(x,y)*. It estimates the coefficient.

- *cor.test(x,y)*. It estimates the coefficient but also gives the 95% confidence interval and the p-value related to the following hypothesis test:

$$\begin{aligned} H_0 :& \quad \rho = 0 \\ H_A :& \quad \rho \neq 0 \end{aligned}$$

where ρ stands for the correlation coefficient. Therefore this is a test of independence between X and Y.

However, two conditions must hold to apply this coefficient:

1) The relation between X and Y must be linear.
2) X and Y must follow a Normal distribution (actually they must jointly follow a bivariate Normal distribution).

Example. A research study aims to determine the degree of relation between blood cholesterol level (mg/100ml) and the saturated fat intake (gr/week). A sample of 20 subjects was collected and their blood cholesterol and saturated fat intake was measured.

Let's load the data:

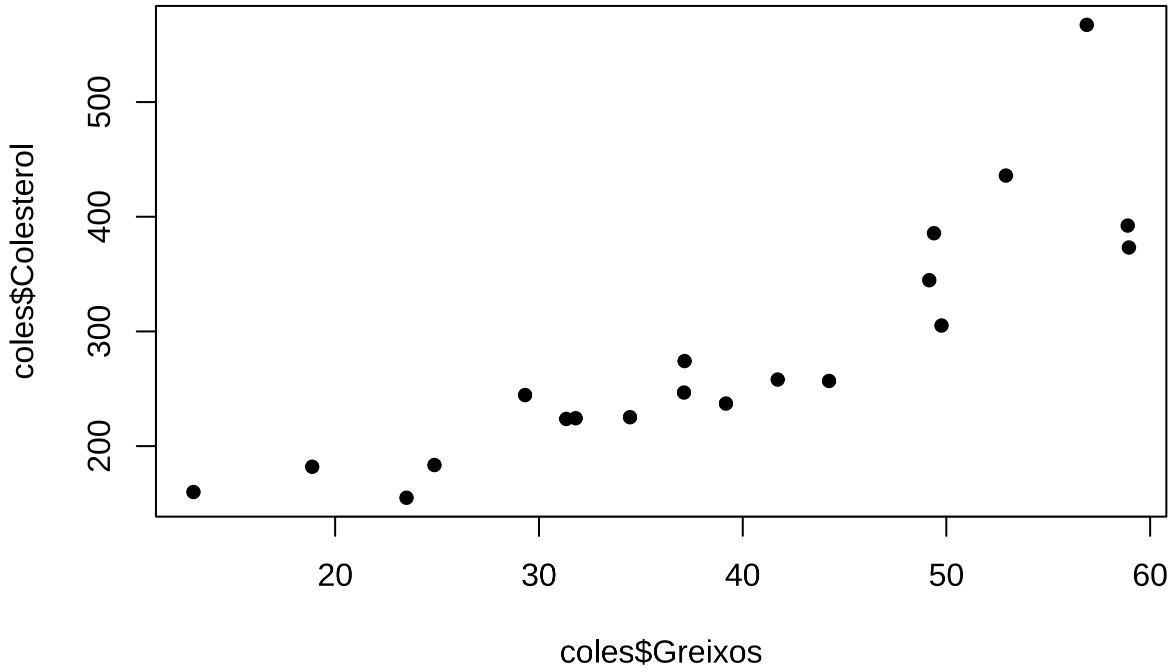
```
coles=read.table("colesterol.txt",header=T)
```

The file contains three variables:

- *Pacient*. The subject's identification number.
- *Greixos*. Weekly saturated fat intake.
- *Colesterol*. Blood cholesterol level.

Let's plot the two variables:

```
plot(coles$Greixos,coles$Colesterol,pch=16)
```



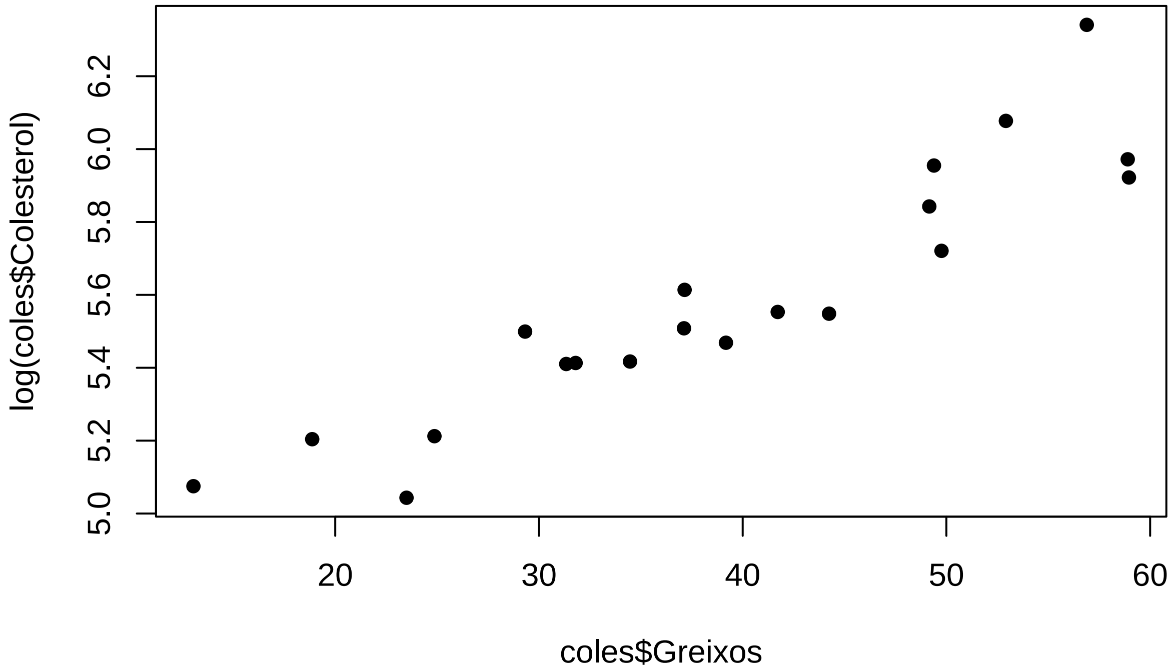
It looks like the relation between X and Y is exponential rather than linear.

Non-linear relation

1) Try to linearize the relation by transforming the data. Then apply the Pearson's correlation coefficient.

In the case example let's apply the logarithm to the blood cholesterol.

```
plot(coles$Greixos,log(coles$Colesterol),pch=16)
```



The relation between the saturated fat intake and the logarithm of blood cholesterol seems quite linear. Next let's compute the Pearson's correlation coefficient.

```
cor.test(coles$Greixos,log(coles$Colesterol))
```

```

    Pearson's product-moment correlation

data:  coles$Greixos and log(coles$Colesterol)
t = 10.347, df = 18, p-value = 5.268e-09
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8173993 0.9704317
sample estimates:
      cor
0.9252442
```

The coefficient is 0.925 indicating a high level of correlation.

2) If there is no transformation that linearizes the relation then apply the **Spearman's correlation coefficient**. This coefficient is based on ranks instead of the raw data.

Let's suppose that there is no transformation able to linearize the relation between the saturated fat intake and the blood cholesterol levels. Thus, we compute the Spearman's correlation coefficient as:

```
cor.test(coles$Greixos,coles$Colesterol,method = "spearman")
```

```
    Spearman's rank correlation rho

data:  coles$Greixos and coles$Colesterol
S = 82, p-value = 5.499e-06
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho 
0.9383459 
```

Notice that the output gives the coefficient estimate, the p-value associated to the independence test but no confidence interval is shown. One option is to use the bootstrap.

The following function generates a bootstrap sample and computes the Spearman's correlation coefficient.

```
boot.cor=function(){
mos=sample(1:20,replace=T)
cor(coles[mos,]$Greixos,coles[mos,]$Colesterol,method="spearman")
}
```

Let's replicate the function 1000 times.

```
boot.r=replicate(1000,boot.cor())
```

Finally the 95% confidence interval is obtained as:

```
quantile(boot.r,probs=c(0.025,0.975))
```

```
     2.5%     97.5% 
0.8131703 0.9803187 
```

Non normality of X or Y

If X or Y do not follow a Normal model but its relationship is linear, the Pearson's correlation coefficient can be still applied. However, the methods to get the confidence interval or the p-value to test the independence need to be modified.

Following the example, we have seen that applying the logarithm to the blood cholesterol levels the relationship have been linearized. However, let's suppose that it is not possible to assume normality of both variables.

- Confidence interval: use bootstrap

The following function generates a bootstrap sample and computes the Pearson's correlation coefficient.

```
boot.cor=function(){
mos=sample(1:20,replace=T)
cor(coles[mos,]$Greixos,coles[mos,]$Colesterol)
}
```

Let's replicate the function 1000 times.

```
boot.r=replicate(1000,boot.cor())
```

Finally the 95% confidence interval is obtained as:

```
quantile(boot.r,probs=c(0.025,0.975))
```

```
     2.5%     97.5% 
0.8224487 0.9531941 
```

- Independence test: use permutations approach.

Next function creates a data permutation and computes the coefficient.

```
perm.cor=function(){
new.dat=data.frame(sample(coles$Greixos),sample(log(coles$Colesterol)))
cor(new.dat[,1],new.dat[,2])
}
```

Let's replicate the function 1000 times.

```
perm.r=replicate(1000,perm.cor())
```

Finally, the p-value is obtained as:

```
((sum(abs(perm.r)>abs(cor(coles$Greixos,coles$Colesterol)))+1)/(1000+1)
```

```
[1] 0.000999001
```

The value of 0 means the no coefficient estimate from the 1000 permutations is larger than the estimate from the original sample. So that it would fair to say that the p-value is lower than 0.001.

Sample size

We will use the function *pwr.r.test(r,power)* to compute the sample size necessary to reject the null hypothesis of independence with a specific power. The arguments of the function are:

- *r*. The correltion coefficient under the alternative hypothesis.
- *power*. The power of the test.

Example. What is the sample size necessary to reject the null hypothesis of independence if the correlation coefficient is of 0.3?

```
pwr.r.test(r=0.3,power=0.8)
```

```

     approximate correlation power calculation (arctangh transformation) 

              n = 84.07364
              r = 0.3
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

It is necessary a sample with 5 subjects.