BIOSTATISTICS

Statistical hypothesis testing. Parametric and non-parametric methods for quantitative and qualitative variables.

Angela Jimeno Martin (ajimeno@usj.es)

June 2022

Contents

Introduction to hypothesis testing

Hypothesis testing is another very important part of statistical inference. The previous unit focused on estimating parameters: finding an expected value of a parameter that was unknown in a population. But hypothesis testing is about deciding whether the value we have found or some other value is the true value. To do this, two or more hypotheses are formulated and, by means of various statistical tests, with the information that comes from the population sample, it is decided which of the hypotheses is accepted.

In this unit the hypotheses to be put forward will be:

- **Null hypothesis**, H_0 , which tries to reflect that the observed phenomenon is the result of chance and that "there is nothing interesting to study."
- **Alternative hypothesis**, H_1 , which is the hypothesis that is put forward and that you want to prove. There are three types of alternative hypotheses:
  - population parameter > hypothesised value (estimated value). One tail on the right.
  - population parameter < hypothesised value (estimated value). One tail on the left.

– population parameter $\neq$ hypothesised value (estimated value). Two-tailed.

Hypothesis evaluation procedure

The most common approach is to focus on the null hypothesis, which is always simpler than the alternative, and decide whether to accept or reject it. This is done by examining the evidence provided by the observed data against the null hypothesis, if the evidence is large, H_0 is rejected. The evidence sought has two components:

- Apply a statistical test (t-Student, χ_2 , F, etc.) on the sample.
- Evaluate the possible values obtained from the test to determine which values lead to accepting the null hypothesis and which to rejecting it.

Given some aleatory event, it is possible to determine the probability of that event to occur using probability distribution functions. In this way, we can calculate if the event is very likely to occur or if it is very rare, and the ocurrance of that event is not due to aleatory facts.

If we think of the distribution of the sample statistic (t-Student, χ_2, F), the calculated statistic (mean or variance, respectively) will be located somewhere on the graph. If its location is at some extreme of the graph, where the probability states it is very unlikely to occur, the null hypothesis will be rejected. But it is necessary to establish the criterion that indicates up to here the value of the statistic validates the null hypothesis and from here, it rejects it. The criterion is the same α used to establish confidence intervals, the level of error, which usually takes the value of 0.05.

Actually, there are two types of errors when assessing with a statistical test whether we accept or reject the null hypothesis:

- `Type I error: the null hypothesis is rejected when it is true.`
- `Type II error: the null hypothesis is accepted when in fact it should have been rejected.`

The probability of committing the type I error is α and the probability of committing the type II error is β. The fact that α is used to determine whether the value of the statistic is too extreme for the null hypothesis to be accepted is because the intention is not to reject H_0 unless the evidence against it is very large, i.e. making a type II error is tolerable, as the repercussions of making such an error are not as serious as making a type I error. The value of β refers to the power of the test.(1 - β) corresponds to the probability that the null hypothesis is rejected as false and, therefore, the alternative is accepted as true, i.e. the probability that the statistical test detects the true positive. And although it is desired that β be as small as possible and α as small as possible, these two errors are complementary, so that what is prioritised is that α be small and 1 - β large.

Evaluation of hypotheses for the mean of the population: calculate the p-value

Statistics are informative values calculated from the observed data in a sample. Since these statistics are used to evaluate hypotheses that are put forward, we will refer to them as test statistics. Thus, to evaluate hypotheses about the population mean, the sample mean, $\overline{X}$, will be used as the test statistic.

A statistical test will be considered as such if its distribution is known (or approximate) for the null hypothesis. We are going to load the body temperature data (ºF) in a dataset from the *BodyTemperature.txt* file.

```r
bodytemperature <- read.table(file="BodyTemperature.txt",header = TRUE)
temperature <- bodytemperature$Temperature
```

The hypotheses to establish could be set as:

$$\begin{cases} H_0: & \mu = 98.6 \\ H_1: & \mu < 98.6 \end{cases}$$

To begin with, let us establish that we know the population variance, which is $\sigma = 1$, and that 25 temperatures are chosen at random from the dataset.

```
set.seed(123)
temperature.smpl <- sample(temperature, size=25)
```

We may assume that the sample distribution can be state using the normal distribution function for the mean,$\overline{X}$:

$$\overline{X} \sim N(\mu, \sigma^2/n) \rightarrow \overline{X} \sim N(\mu, 1/25)$$

Let us now assume that the null hypothesis is true and that the population mean is $\mu = 98.6$. Let us plot the distribution of the values that $\overline{X}$ can take:

```
par(mfrow=c(1,2))
x <- seq(97.85, 99.3, length.out = 10000)

probX <- dnorm(x,mean=98.6,sd=0.2)

plot(x, probX, type="l", xlab="mean"
     ,main="Temperature mean"
     ,ylab="Density", xlim=c(97.85,99.3))
abline(h=0, col="gray")
segments(x0 = 98.4, y0 = 0, y1 = 1.2, x1 = 98.4)

z <- (x-98.6)/(1/sqrt(25))
probZ <- dnorm(z,mean=0,sd=1)

plot(z, probZ, type="l", xlab="z"
     ,main="Z value"
     ,ylab="Density", xlim=c(-4, 4))
abline(h=0, col="gray")
segments(x0 = -1, y0 = 0, y1 = 0.24, x1 = -1)
```

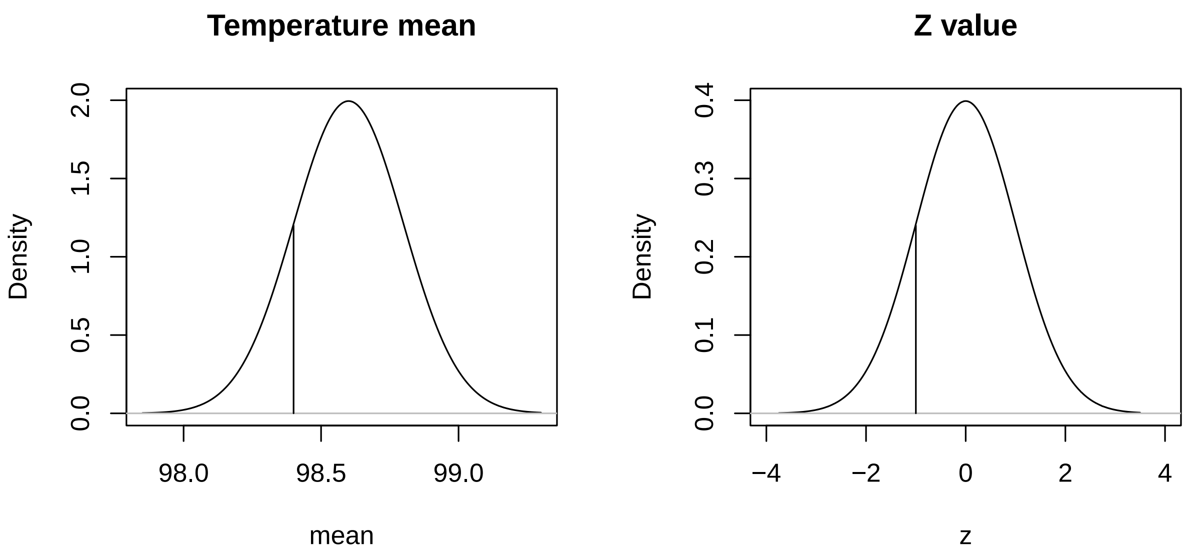

Figure 1: Null hypothesis

These graphs show the distribution of temperatures for the null hypothesis, and the corresponding value of the calculated z-statistic. In this case, the population deviation is known and is indicated in both graphs by a vertical line. This is known as the lower probability tail, and shows those temperature values that have a probability equal to or less than 98.4 of occurring. More extreme values will be found thereafter.

The probability values corresponding to these extreme values are called the **significance level** for the test statistic and are represented as p**-values**.

$$p = P(\overline{X} \leq \overline{x} | H_0)$$

The value of p is a probability conditional on H_0 , since it is assumed to be true. The alternative hypothesis, H_1 , states that the value of the mean will be below μ.

```
temperature.mean <- mean(temperature.smpl)
temperature.mean
```

```
## [1] 98.368
```

Therefore the p-value is calculated as:

$$p = P(\overline{X} \leq 98.37)$$

As this is a normal distribution, this value can be calculated with `pnorm`

```
pnorm(temperature.mean,mean = 98.6, sd = 0.2)
```

```
## [1] 0.1230244
```

Interpretation of the p-value

The p-value is the conditional probability of the extreme values of a statistical test assuming that the null hypothesis is true. As the p-value becomes larger it means that the value of the mean of the alternative hypothesis is not so far away from the null hypothesis, but closer to it, fitting the generated distribution.

For the p-value to help decide whether to accept or reject the null hypothesis, it is necessary to establish a threshold probability value at which the null hypothesis no longer agrees with the observed data. This threshold is called the significance level or test size. Typical significance levels are: 0.01, 0.05 or 0.1. This threshold corresponds to α, and if values smaller than this are obtained, the data are said to provide statistically significant evidence against H_0 .

When significant differences are found between the value of the analysed sample statistic and that of the population, it is assumed that this difference is very unlikely to be due to chance alone, and that the sample is actually different from the proposed null hypothesis. In the case of small sample sizes, it may be the case that we reject H_0 but it is true, leading to an inconclusive result, as it is not really known which of the two scenarios we may be in. In case more than 2 hypotheses are tested at the same time, more advanced statistical inference methods are usually needed.

Normality test

As it was mentioned previously, depending on the statistical distribution of sample data, the type of hypothesis test will vary. Many hypothesis testing approaches are based on the fact that the population from which the samples come follow a normal distribution. To confirm these assumptions, the **Shapiro-Wilk** test should be used. The statistic for this test is W , which would be calculated:

$$W = \frac{(\sum_{i=1}^{n} a_i X_i)^2}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

Where a_i is an expression calculated from the median and the covariance matrix. The value of W is between 0 and 1. In the normality, the following hypothesis test is performed:

$$\begin{cases} H_0: & \text{Normally distributed population} \\ H_1: & \text{Not normally distributed population} \end{cases}$$

The higher the value of W the lower the p-value, if the p-value found is lower than the established significance level α, then the null hypothesis that the population follows a normal distribution is rejected, and the tests explained above cannot be used to test their hypotheses.If a given variable follows a normal distribution, then it is recommended to use a **parametric tests** to complete the hypothesis testing analysis. On the contrary, **non parametric tests** are preferred. In R, the functions to perform the normality test are:

- `shapiro.test`, which receives as argument the vector with the sample.
- Realisation of a Q-Q plot. It is called Q-Q, because it represents the *quantile*. By plotting the quantiles of two distributions you can compare them and determine if they are the same.

`Example:` From the BodyTemperature dataset used before, apply the Shapiro-Wilk normality test and a Q-Q plot for the variable Temperature.

```
shapiro.test(bodytemperature$Temperature)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  bodytemperature$Temperature
## W = 0.98171, p-value = 0.1803
```

The result of the normality test is that in the case of Temperature, the sample, and therefore the population follow a normal distribution since the p-value is greater than 0.05 (significance level). It is possible to check this graphically assessing through a Q-Q plot if the distribution of the variable fits the normal. This is directly done with the R function `qqnorm`.

```
qqnorm(bodytemperature$Temperature)
qqline(bodytemperature$Temperature,lwd=2,col="green")
```

It can be clearly seen how the Temperature variable fits well to the trend line of the normal distribution (green).

Parametric tests

Normal distribution, α known, large samples

The statistical test to be used will be z, the distribution of z is shown to be a standardized normal distribution. We will have two values of z: * Z, which will be the one calculated for the null hypothesis (theoretical) * z for the alternative hypothesis.

The formula for z will be:

$$p = P(\frac{\overline{X} - \mu}{\sigma})$$

Where $\overline{X}$ is the mean of the sample, μ is the mean of the population and σ the standard deviation.

The hypotheses in this situation can be established as follows:

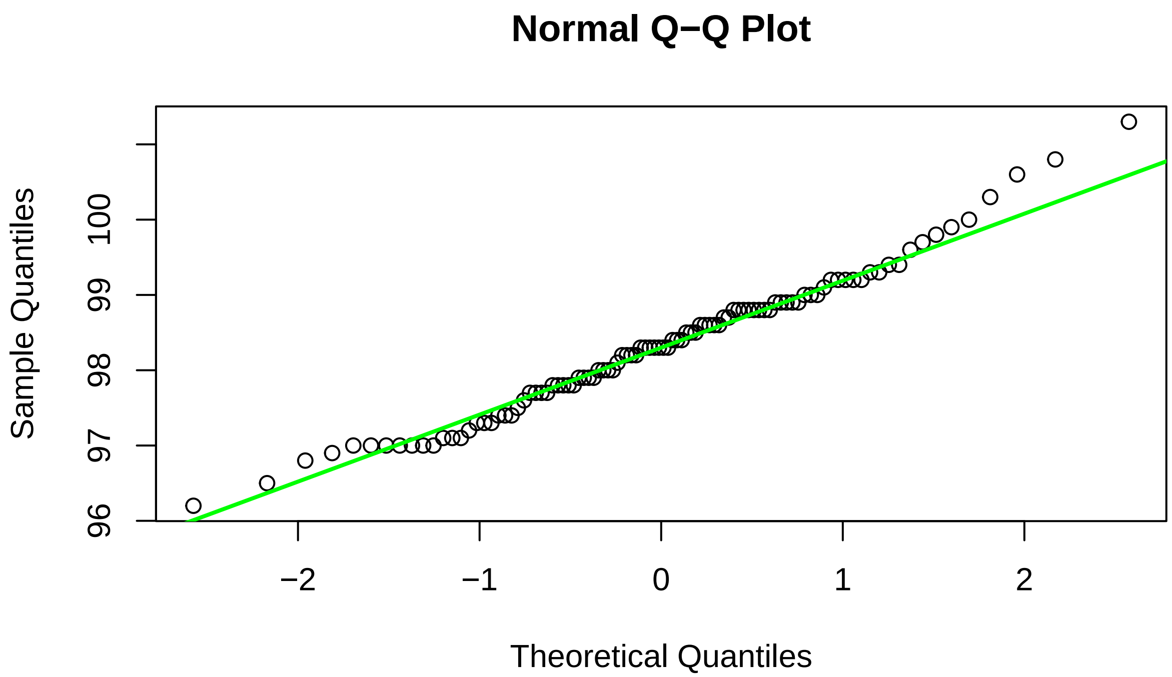

Figure 2: QQPlot

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu < \mu_0 \\ H_1: & \mu > \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases}$$

In the cases of $H_1 : \mu < \mu_0$ and $H_1 : \mu > \mu_0$ the contrasts are called **one-tailed**, since only one direction of the graph is observed. Regarding continuous variables $\leq$ or $<$ are equivalent. For the third case of H_1, the contrast tests are called **two-tailed**, since the extreme values of the sample mean can be found in both directions of the graph. In addition, the significance level must be divided by two. To know if the null hypothesis is rejected, we may use two distinct criteria: comparing the values of z and Z_α or using p-value. Regarding this last criteria, if $p < \alpha$ the null hypothesis is rejected. According to the others:

- For $H_1 : \mu < \mu_0$ null hypothesis is rejected if (negative z value) $z \leq Z_\alpha$ or $z < Z_\alpha$
- For $H_1 : \mu > \mu_0$ null hypothesis is rejected if $z \geq Z_\alpha$ or $z > Z_{1-\alpha}$.
- For $H_1 : \mu \neq \mu_0$ null hypothesis is rejected if $z \leq Z_{\frac{\alpha}{2}}$ (z valor negativo) or $z \geq Z_{1-\frac{\alpha}{2}}$.

`Example:` In the genomes of prokaryotes, genes are organised into operons. Genes within an operon tend to have similar expression levels. An experiment is carried out to find the average true expression of an operon. The average expression of an operon consisting of 34 genes is found to be 0.20. In the literature, the mean expression is found to be 0.28 for that operon when the population variance is 0.14. The significance level is set at $\alpha = 0.05$. Is it a significant difference that has been found in the experiment?

Assumptions:

- Population presents a normal distribution
- The variance is known
- Sample is large

The hyphothesis test proposed:

$$\begin{cases} H_0: & \mu = 0.28 \\ H_1: & \mu \neq 0.28 \end{cases}$$

If the value of the mean coincides with 0.28, it will be accepted that the experiment agrees with the bibliographic data. Otherwise, the expression found for the operon is different and therefore, it may be another operon or different conditions than those proposed in the literature.

Applying the statistical test:

$$z = \frac{0.2 - 0.28}{\frac{\sqrt{0.14}}{\sqrt{34}}} = -1.24$$

Let's check the rejection criteria:

- According to z value: $H_1 : \mu \neq \mu_0$ then $z \leq Z_{\frac{\alpha}{2}}$ (negative z value) or $z \geq Z_{1-\frac{\alpha}{2}}$
- According to p-value: $p = 2P(Z \geq |z|) = 2 \cdot (1 - P(Z \leq |z|))$

```r
Z <- qnorm(0.05/2)
round(Z,2)
```

```
## [1] -1.96
```

To reject the hypothesis $z le Z_{\alpha}$. Substituting the values: -1.24 > -1.959964, the null hypothesis is not rejected. Using the second rejection criterion:

```r
z <- abs((0.2 - 0.28)/(sqrt(0.14)/sqrt(34)))
1-pnorm(z)
```

```
## [1] 0.1062519
```

```r
pvalue <- 2*(1-pnorm(z))
```

A value of 0.1062519 is obtained, which is $> \alpha$ (0.05). The null hypothesis is therefore accepted, in the same way as for the first criterion.

Normal distribution, α unknown or small samples

The statistical test to be used will be the t-test. This statistic follows a t-Student distribution, which is similar to a standardised normal distribution, whose degrees of freedom, v , will correspond to $n - 1$ (n corresponding to the sample size). The hypotheses to contrast are establised the same way as in the previous section with z, however the statistic here is t and its expression is as follows:

$$t = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

In this situation, depending if the contrast is one-tail or two-tail we will find different rejection criteria:

- For $H_1 : \mu < \mu_0$ null hypothesis is rejected if $t \leq T_{\alpha,v}$ or $t < T_{\alpha,v}$.
- For $H_1 : \mu > \mu_0$ null hypothesis is rejected if $t \geq T_{\alpha,v}$ or $t > T_{1-\alpha,v}$.
- For $H_1 : \mu \neq \mu_0$ null hypothesis is rejected if $t \leq T_{\frac{\alpha}{2}}$ or $t \geq T_{1-\frac{\alpha}{2}}$

`Example:` In DNA, the ratio of G and C must be the same, which is why it is often referred to as the G+C ratio. A sample of 16 DNAs is taken from that species, and it is found that the standard deviation of the sample is 15 and the number of G+C 310. The significance level is set at $\alpha = 0.05$. Does the result agree with the hypothesis that it is greater than 300?

Assumptions:

- Population presents a normal distribution
- The variance is known
- Sample is large

The hypothesis test proposed:

$$\begin{cases} H_0: & \mu = 300 \\ H_1: & \mu \geq 300 \end{cases}$$

Applying the statistical test:

$$t = \frac{310 - 300}{\frac{15}{\sqrt{16}}} = 2.67$$

For $H_1 : \mu > \mu_0$ null hypothesis is rejected if $t \geq T_{\alpha,v}$ or $t > T_{1-\alpha,v}$.

```
Te <- qt(1-0.05/2,15)
round(Te,2)
```

```
## [1] 2.13
```

Substituting the values: $2.67 > 2.13$, therefore the null hypothesis is rejected. Moreover, using the second criteria $p = 1 - P(T \leq t)$, if we calculate the p-value:

```
t <- abs((31.71 - 30)/(8.01/sqrt(200)))
pvalue <- 1-pt(t,199)
pvalue
```

```
## [1] 0.001433573
```

A value of 0.001433573 is obtained, which is $< \alpha$ (0.05). Therefore, the null hypothesis is rejected, confirming what was obtained with the first criterion.

The R, `t.test` function performs the hypothesis test only when a vector of sample values is passed as a parameter.

Evaluation of hypotheses for population proportions

For a binary random variable, X, the possible values are 0 and 1. In hypothesis testing one is often interested in assessing the proportion of the population that has had the outcome of interest, i.e. $X = 1$.

The hypothesis testing of two population proportions is equivalent to the comparison of means of a population with a normal distribution, with known variance or large population size. Consequently, the statistical test to be applied is z.The null hypothesis will therefore follow a normal distribution.

$$H_0 \sim N(\mu_0, \mu_0 \cdot (1 - \mu_0))$$

$$Z = \frac{\overline{X} - \mu_0}{\sqrt{\frac{\mu_0 \cdot (1 - \mu_0)}{n}}}$$

For the alternative hypotheses, we may observe the same situations as previously, one or two-tail contrasts:

1. $H_1 : \mu < \mu_0$
2. $H_1 : \mu > \mu_0$
3. $H_1 : \mu \neq \mu_0$

Rejection criteria for null hypothesis are also the same as with the case of a continuous variable with z distribution.

Example: Using the Melanoma dataset of the MASS package, we may say the null hypothesis is that the mean number of ulcerations observed in patients is 0.5, i.e. 50% of patients have ulcerations. As an alternative hypothesis, it is believed to be less than this proportion. Consider a significance level of 0.05 and determine with hypothesis prevails.

```
require(MASS)

# We use data from column ulcer as cualitative variable.
Melanoma$ulcer <- as.factor(Melanoma$ulcer)

summary(Melanoma)
```

```
##       time          status          sex              age             year     
##  Min.   :  10   Min.   :1.00   Min.   :0.0000   Min.   : 4.00   Min.   :1962  
##  1st Qu.:1525   1st Qu.:1.00   1st Qu.:0.0000   1st Qu.:42.00   1st Qu.:1968  
##  Median :2005   Median :2.00   Median :0.0000   Median :54.00   Median :1970  
##  Mean   :2153   Mean   :1.79   Mean   :0.3854   Mean   :52.46   Mean   :1970  
##  3rd Qu.:3042   3rd Qu.:2.00   3rd Qu.:1.0000   3rd Qu.:65.00   3rd Qu.:1972  
##  Max.   :5565   Max.   :3.00   Max.   :1.0000   Max.   :95.00   Max.   :1977  
##    thickness     ulcer  
##  Min.   : 0.10   0:115  
##  1st Qu.: 0.97   1: 90  
##  Median : 1.94          
##  Mean   : 2.92          
##  3rd Qu.: 3.56          
##  Max.   :17.42          
```

Let's calculate p, the proportion of the sample that has ulcers on their melanoma.

```
p <- length(Melanoma$ulcer[Melanoma$ulcer == 1])/nrow(Melanoma)
p <- round(p,2)
p
```

```
## [1] 0.44
```

For this situation, the contrast is as follows:

$$\begin{cases} H_0: & \mu = 50 \\ H_1: & \mu < 50 \end{cases}$$

Where z value would be:

$$z = \frac{0.44 - 0.5}{\sqrt{\frac{(0.5 \cdot (1-0.5))}{205}}} = -1.72$$

The first criteria states that for $H_1 : \mu < 0.5$ then we reject null hypothesis if $z \leq Z_{0.05}$

```
Z <- qnorm(0.05)
round(Z,2)
```

```
## [1] -1.64
```

We observe that it is satisfied, $z\ le\ Z_{0.05}$, since -1.72 < -1.64. Therefore, the null hypothesis is rejected. Furthermore, according to the p-value criterion, we calculate $p = P(Z \leq z)$ and check if $p < \alpha$.

```
z <- (0.44 - 0.5)/sqrt((0.5*(1-0.5))/(205))
pvalue <- pnorm(z)
pvalue
```

```
## [1] 0.04288568
```

And we observe that $0.0428 < 0.05$, then again null hypothesis is rejected.

Hytphotesis testing for comparing two populations

In this case the null hypothesis will respond to the following structure:

$$H_0 = \mu_1 - \mu_2 = \Delta$$

When comparing two mean values in the hypothesis, the tests to be performed are called two-sample tests, whereas those seen so far would be one-sample tests. The two samples from which the test is carried out to compare the means can be either independent (where samples come from distinct populations) or dependent (where samples come from the same population and suppose observations in different time points).

The statistical values are calculated and analyses following an analogous approach as with one-sample tests. Such that:

- Normal populations, known variances or large sample size where samples are independent use the Z-statistic:

$$Z = \frac{\overline{X_1} - \overline{X_2} - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Normal populations, unknown but equal variances or small sizes where samples are independent use t-statistic:

$$t = \frac{\overline{X_1} - \overline{X_2} - \Delta}{\sqrt{\frac{(n_1-1)\cdot S_1^2+(n_2-1)\cdot S_2^2}{n_1+n_2-2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- Normal populations, unknown and unequal variances or small sizes where samples are independent use t-statistic modified:

$$t = \frac{\overline{X_1} - \overline{X_2} - \Delta}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

- Normal populations, unknown variance or small sizes where samples are dependent use t-statistic modified:

$$t = \frac{\overline{d} - \Delta}{\frac{s_d}{\sqrt{n}}}$$

Where d is the mean of the pairwise difference, n the number of pairs and sd the standard deviation of the pairwise differences. The degrees of freedom correspond to $v = n - 1$. Actually, this test is like a one- sample t-test.

Non-parametric tests

In the study situation where a hypothesis test has to be performed and the populations do not follow a normal distribution, the tests studied in the previous unit may not yield conclusive results when applied to the samples. In this case, it will be necessary to apply the corresponding non-parametric test to perform the test.

These tests assume little or nothing about the probability density obtained from the data. Therefore, they are often used in cases where the samples are not normally distributed, or it is not known which distribution they follow, or their sample sizes are very small. The analysis of non-numerical data would also fit here.

Chi-square test

This test is used to analyze categorical data. The organization of data for this analysis uses cross tables. Cross tables are a simple and powerful method for aggregating count data into different categories (absolute frequencies per category).

These tables consist of f rows and c columns. The simplest case would be a 2x2 table. The rows represent different categories of one categorical variable and the columns represent different categories of another categorical variable. What the contingency tables show is a "summary" of the categorical data. If we have the categorical variable `Genotype`, with 3 possible categories, and another variable `Body Mass Index` with 5 categories:

- Cell 11 shall indicate the count of elements where the first genotype matches the first BMI category.
- Cell 12 shall indicate the count of items where the second genotype matches the first BMI category.
- Cell 21 shall indicate the count of items where the first genotype matches the second BMI category.
- ...

Cross table (example)

—	categoryA1	categoryA2	Total
categoryB1	a	b	a+b
categoryB2	c	d	c+d
Total	a+c	b+d	n=a+b+c+d

$\frac{a+c}{n}$ y $\frac{b+d}{n}$ are marginal probabilities of categories A1 and A2, respectively. While $\frac{a+b}{n}$ y $\frac{c+d}{n}$ are marginal probabilities for categories B1 and B2.

The analysis of contingency tables allows us to discriminate whether two categorical variables are related to each other or not by performing the χ^2 test on them. This test is based on a comparison of expected values against observed ones obtained from the sample analysis.

$$\chi^2 = \sum_j \sum_i \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

with degrees of freedom, $(f-1) \cdot (c-1)$, where f is the number of rows and c the number of columns of the contigency table, o_{ij} is the observed absolute frequency and e_{ij} as:

$$N \cdot \left(\frac{F_i}{N} \cdot \frac{C_j}{N} \right)$$

Where F_i is the marginal probability of column i and C_i is the marginal probability of column j. The null hypothesis is that the two variables are independent, while the alternative is that they are dependent.

$$\begin{cases} H_0: & \text{Independent variables} \\ H_1: & \text{Dependent variables} \end{cases}$$

If the p-value obtained with the `chisq.test` function is lower than the significance level α, the alternative hypothesis that the variables analysed are related is accepted.

Fisher's exact test

In case the expected values, calculated as described above, result in less than 5 in more than 20% of the cells, Fisher's exact test must be performed. Fisher's exact test calculates every possible combination of the N values in the table (it performs a permutation of all values in all existing categories, as long as the value of the sum of the rows and columns remains the same). With the resulting values it creates a distribution in which calculates how extreme the results are for the resulting contingency table.

Fisher's test is usually used in scenarios where two variables with two categories each are compared.

The assumption is the same as for χ^2 for independence.

$$\begin{cases} H_0: & \text{Independet variables} \\ H_1: & \text{Dependent variables} \end{cases}$$

R is able to compute Fisher's test through the function `fisher.test` which must receive a data matrix as parameter.

`Example:` You want to relate the result of a genetic test carried out on gene A, which has two alleles, to the presence of a disease. The results obtained are shown in the following table:

—	Disease (+)	Disease (-)
Allele1	45	122
Allele2	67	38

Using the χ^2 test and Fisher's exact test, determine whether there is a relationship between alleles and disease. Consider 5% significance.

The hypothesis test would be the same for both tests:

$$\begin{cases} H_0: & \text{gene and disease are unrelated} \\ H_1: & \text{gene and disease are related} \end{cases}$$

The matrix is created with the data as follows:

```r
crosstab <- matrix(c(45,67,122,38),nrow = 2,ncol = 2,byrow = FALSE
                            ,dimnames = list("Gene"=c("Allele1","Allele2"),"Disease"=c("Yes","No")))
crosstab
```

```
##          Disease
## Gene       Yes  No
##    Allele1  45 122
##    Allele2  67  38
```

We then compute the marginal probabilities:

—	Disease (+)	Disease (-)	Total
Allele1	45	122	167

—	Disease (+)	Disease (-)	Total
Allele2	67	38	105
Total	112	160	272

Test χ^2

Expected values are calculated for each element in the table.

$$N \cdot \left(\frac{F_i}{N} \cdot \frac{C_j}{N} \right)$$

$e_{11} = 272 \cdot \left(\frac{167}{272} \cdot \frac{112}{272} \right) = 67.76$

$e_{12} = 272 \cdot \left(\frac{167}{272} \cdot \frac{160}{272} \right) = 98.23$

$e_{21} = 272 \cdot \left(\frac{105}{272} \cdot \frac{112}{272} \right) = 43.23$

$e_{22} = 272 \cdot \left(\frac{105}{272} \cdot \frac{160}{272} \right) = 61.76$

The degrees of freedom are : $(2-1) \cdot (2-1) = 1$

Finally, if we susbstitute the expression of χ^2:

$$\chi^2 = \sum_j \sum_i \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \frac{(45 - 67.76)^2}{67.76} + \frac{(122 - 98.23)^2}{98.23} + \frac{(67 - 43.23)^2}{43.23} + \frac{(38 - 61.75)^2}{61.75} =$$

$$7.64 + 5.75 + 13.08 + 9.14 = 35.61$$

The expected values for χ^2_1:

```r
e11 <- (112/272)*(167/272)*272
e12 <- (160/272)*(167/272)*272
e21 <- (112/272)*(105/272)*272
e22 <- (160/272)*(105/272)*272

expected <- matrix(c(e11,e12,e21,e22),nrow = 2,byrow = FALSE)
observed <- matrix(c(45,122,67,38),nrow = 2,ncol = 2,byrow = FALSE)
expected
```

```
##          [,1]     [,2]
## [1,] 68.76471 43.23529
## [2,] 98.23529 61.76471
```

```r
observed
```

```
##      [,1] [,2]
## [1,]   45   67
## [2,]  122   38
```

```r
diference <- expected - observed
diference.square <- diference * diference
diference.square.fraction <- diference.square / expected
diference.square.fraction
```

```
##          [,1]      [,2]
## [1,] 8.212952 13.062505
## [2,] 5.749067  9.143754
```

```r
value.chi <- sum(diference.square.fraction)

# Calculate p-value:
pchisq(value.chi,df = 1,lower.tail = FALSE)
```

```
## [1] 1.80993e-09
```

As p-value $< \alpha$, we then reject null hypothesis and we may conclude variables are related or dependent.

Also `chisq.test` function performs directly the operation.

```r
chisq.test(crosstab)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  crosstab
## X-squared = 34.662, df = 1, p-value = 3.921e-09
```

For the Fisher's test, the `fisher.test` function computes all possible contingency tables and calculates the p-value.

```r
fisher.test(crosstab)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  crosstab
## p-value = 2.041e-09
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.1195958 0.3652159
## sample estimates:
## odds ratio
##  0.2105543
```

The p-value is in the same order of magnitude as that calculated for the original table as already advanced. This value demonstrates that the null hypothesis that the variables are independent of each other must be rejected, showing that they are related.

Non-parametric test for two samples

As we have seen above for parametric tests, it is sometimes necessary to compare random variables from two different populations, X and Y, which may have different distribution functions.

Therefore, in two-sample parameter tests of two different populations, the null hypothesis to be tested will be that the two populations to be studied follow the same distribution, while the alternative is that the populations follow different distributions due to a shift at one point in the distribution θ.

$$\begin{cases} H_0: & F_Y(x) = F_X(x) \\ H_1: & F_Y(x) = F_X(x+\theta) \text{ where } \theta \neq 0 \end{cases}$$

So there will be m + n random variables (m from the X distribution and n from the Y distribution), which can be arranged in $\binom{m+n}{m}$ different ways.

Wilcoxon test for independent samples

This statistical test would be the equivalent of the `t.test` for the comparison of means in normal populations. In this test, the m and n values of both samples are combined to give N, in order from highest to lowest or lowest to highest. Each random variable in the population X will have been assigned a rank (an ordered position), R_i . The sum of Wilcoxon ranks will be:

$$W = \sum_{i=1}^{m} R_i - \frac{n_i(n_i+1)}{2}$$

The null hypothesis will be that the location parameter, $\theta = 0$, i.e. there has been no displacement and the distribution functions of both populations are equal.

$$H_0 : \theta = 0$$

Alternative hypotheses, on the other hand, can be of various types, similar to parametric tests:

- $H_1 : \theta > 0$. Null hypothesis will be rejected if $W \geq w_\alpha$.
- $H_1 : \theta < 0$. Null hypothesis will be rejected if $W \leq n(m+n+1) - w_\alpha$
- $H_1 : \theta \neq 0$. Null hypothesis will be rejected if $W \geq w_{\frac{\alpha}{2}}$ or $W \leq n(m+n+1) - w_{\frac{\alpha}{2}}$.

Where, w_α is the constant analogous to the significance value, α, in parametric tests.

`Example`: In a study they want to determine whether a drug can change the level of expression in a certain number of genes in a cancer patient. The data are collected at two different times, and θ represents the shifted value of the expressions after application of the drug. Using a significance level of 0.05, determine whether there has been an effect, increasing gene expression.

Gene	Expression - t1	Expression - t2
g1	2670.171	1588.39
g2	2322.195	1377.756
g3	887.6829	638.3171
g4	915.1707	518.0488
g5	19858.68	4784.585
g6	14586.05	3644.561
g7	44259.61	25297.95
g8	34081.98	18560.51
g9	2381.634	1557.293

The contrast here:

$$\begin{cases} H_0 : & \theta = 0 \\ H_1 : & \theta > 0 \end{cases}$$

Then, we join and sort the values in order to obtain, finally, the rank that occupies each value at the new vector:

```r
t1 <- c(2670.171,2322.195,887.6829,915.1707,19858.68,14586.05,44259.61,34081.98,2381.634)
t2 <- c(1588.39,1377.756,638.3171,518.0488,4784.585,3644.561,25297.95,18560.51,1557.293)

N <- c(t1,t2)
```

```
# Order
sort(N)
```

```
##  [1]   518.0488   638.3171   887.6829   915.1707  1377.7560  1557.2930
##  [7]  1588.3900  2322.1950  2381.6340  2670.1710  3644.5610  4784.5850
## [13] 14586.0500 18560.5100 19858.6800 25297.9500 34081.9800 44259.6100
```

```
# Rank
t.rank <- rank(N)
t.rank
```

```
##  [1] 10  8  3  4 15 13 18 17  9  7  5  2  1 12 11 16 14  6
```

After that, it is possible to determine the value of W in case of t1 and t2:

```
# Example: 9 first elements correspond to t1

W <- sum(t.rank[1:9]) - length(t1)*(length(t1)+1)/2
W
```

```
## [1] 52
```

Theoretical value of $w_{0.05}$ for this two samples is computed using `qwilcox`.

```
w.alpha <- qwilcox(0.05,m = length(t1),n = length(t2))
w.alpha
```

```
## [1] 22
```

To reject the null hypothesis we follow the criterion $W \geq w_\alpha$. It is therefore satisfied, as 52 is greater than 22, so the null hypothesis that there is no shift in gene expression is rejected. The application of the drug affects the expression of the group of genes.

Using `wilcox.test`, the same conclusion is reached.

```
wilcox.test(t1,t2,paired = FALSE,alternative = "greater")
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  t1 and t2
## W = 52, p-value = 0.1701
## alternative hypothesis: true location shift is greater than 0
```

Wilcoxon test for dependent samples

If the random variables are paired, they are dependent, and instead of creating a new random variable with the observations of the two samples, the difference between the observations, which in reality come from the same population but collected at different times, is made. This statistical test requires that the population is symmetric about the median, so that the number of differences on either side of the median is the same, i.e. the same number of $X_i - M_0$, where M_0 is the median of the null hypothesis, the theoretical one. The new variable to be included in the test is the previous difference denoted by D_i, and the order rank of it will be obtained. The statistic V, in this case will be calculated as:

$$V = min(W+, W-)$$

Where $W+$ and $W-$ are the sum of the ranks with positive sign, and the sum of the ranks with negative sign, respectively.

The null hypothesis will be rejected under the criteria described above for the w_α set. This test is less powerful than the *t*-test (less likely to reject the null hypothesis when it is false), but because it is based on the median it is more robust. If the sample size is greater than 25, the V statistic is normally distributed, and a Z value would be calculated based on the value of V obtained.

`Example`: Considering the above problem from the perspective of paired samples.

The hypothesis testing to be carried out will be the same. In the table of expressions obtained, one more column should be added, the column for the difference:

Gene	Expression - t1	Expression t2	Difference
g1	2670.171	1588.39	1081.7810
g2	2322.195	1377.756	944.4390
g3	887.6829	638.3171	249.3656
g4	915.1707	518.0488	397.1219
g5	19858.68	4784.585	15074.1000
g6	14586.05	3644.561	10941.4900
g7	44259.61	25297.95	18961.6600
g8	34081.98	18560.51	15521.4700
g9	2381.634	1557.293	824.3410

We then assess the rank for the difference column

```
d <- t1-t2
d.ranked <- rank(d)
d.ranked
```

```
## [1] 5 4 1 2 7 6 9 8 3
```

There are no negative values, such that $W-$ value is equal to 0 and W will be the minimun at the addition $W+$.

```
V <- sum(d.ranked)
V
```

```
## [1] 45
```

Finally if we get the rejection criterion as $W \leq n(m+n+1) - w_\alpha$, we can conclude

```
length(t2)*(length(t2)+length(t1)+1)-w.alpha
```

```
## [1] 149
```

V is lower than 149, therefore the null hypothesis is rejected.

The difference found must be passed to the `wilcox.test` function, and μ=0, which represents that the difference between each pair of observations is symmetrically distributed around 0.

```
wilcox.test(d,mu = 0)
```

```
##
##  Wilcoxon signed rank exact test
##
## data:  d
## V = 45, p-value = 0.003906
## alternative hypothesis: true location is not equal to 0
```

The result thus confirms that there are significant differences before and after applying the drug on the same samples.

Mann-Whitney test

Also known as the Mann-Whitney U test, it is equivalent to the Wilcoxon rank sum test, and is the alternative par excellence to the t-test for independent samples, although the comparison made is

of medians and non-means. It requires homogeneity of variances to be applied. It is one of the most powerful tests available to determine the differences between two samples. This test is based on the same premise as the Wilcoxon rank sum test, including the particularity that by combining the data of the two variables into a single variable and ordering, the aim is to determine the number of times a data item that was contained in sample Y is ahead of a data item contained in sample X. The hypothesis testing is the same as in the previous cases: the statistic U is the minimum sum of ranks found for each sample, W_1 and W_2

$$W_1 = \sum_{i=1}^{m} R_i - \frac{n_1(n_1+1)}{2}$$

$$W_2 = \sum_{i=1}^{n} R_i - \frac{n_2(n_2+1)}{2}$$

$$U = min(W_1, W_2)$$

The criteria for rejecting the null hypothesis are the same as those seen above for the Wilcoxon rank sum test, since these two tests are practically analogous.

In fact, the R function for the Mann-Whitney U is also `wilcox.test` with the paired t-tests at FALSE.

`Example:` The A gene expression data, shown in the table below, have been obtained for cancer patients and control patients. The null hypothesis is that the two groups have the same mean expression level. It is assumed that both groups are independent, use a significance level of 5% to test this hypothesis.

Control	Sick
0.55	0.342
0.51	0.794
0.888	0.465
0.98	0.249
0.514	0.335
0.645	0.4991
0.376	0.295
0.778	0.796
0.089	0.66

The problem does not provide any information about the distribution of either the small sample sizes or the populations of the two groups of subjects, so the t-test cannot be applied. We proceed to use the Mann-Whitney test, which does not require knowledge of the distribution of the samples.

```r
control <- c(0.55,0.51,0.888,0.98,0.514,0.645,0.376,0.778,0.089)
sick <- c(0.342,0.794,0.465,0.249,0.335,0.499,0.295,0.796,0.66)
wilcox.test(sick,control,alternative = "two.sided",paired = FALSE)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  sick and control
## W = 28, p-value = 0.2973
## alternative hypothesis: true location shift is not equal to 0
```

A p-value above 0.05 is obtained, therefore the null hypothesis is accepted, which means that the gene studied has the same level of expression in patients and controls.

Exercises

1. The amount of iron in lentils is 3.3 mg/100g. The amount of iron in a package obtained from the supermarket was measured in a laboratory and the result was 2.7 mg/100g. It was accepted that the difference in iron content was not significant, but later studies showed that it was. It poses the hypothesis contrast of the initial situation and the type of error that was made in accepting the difference as non-significant. ¿Which type of error are researchers performing here?
2. The mean systolic blood pressure is assumed to be 115. It is hypothesized that the pressure is lower than 115 in those who consume a small amount of dark chocolate each day. Select 100 people who consume dark chocolate at random from the population, measure their blood pressure, and measure the mean value of 111 and the standard deviation of 32. Considering that the blood pressure of the population follows a normal distribution, determine whether including a small amount of dark chocolate in the diet helps to lower blood pressure with an error of 0.1. Perform the statistical study of hypothesis testing using two different criteria.
3. We want to determine what proportion of people are smokers. It is hypothesized that the proportion of the population is less than 0.2. To ratify the hypothesis, 150 people are interviewed and it is found that 27 smoke regularly. Evaluate the null hypothesis with 5% significance.
4. Using the `cabbages` dataset from the `MASS` package, determine if the `c39` and `c52` crops have different vitamin C content, knowing that the population distribution is normal and that the population variances are equal. Use $alpha = 0.05$.

Bibliography

Altman, Douglas G, and J Martin Bland. 2009. "Parametric v Non-Parametric Methods for Data Analysis." *Bmj* 338.

Asghari Jafarabadi, Mohammad, and Momeneh Mohammadi. 2015. "Statistical Series: Common Non-Parametric Methods." *Iranian Journal of Diabetes and Metabolism* 14 (3): 145–62.

D'Agostino, Ralph B. 2017. "Tests for the Normal Distribution." In *Goodness-of-Fit Techniques*, 367–420. Routledge.

Turner, Robin, Ari Samaranayaka, and Claire Cameron. 2020. "Parametric Vs Nonparametric Statistical Methods: Which Is Better, and Why?" *New Zealand Medical Student Journal*, no. 30: 61–62.