BioSys PhD | Earthsystems PhD

Statistics 1

Parametric Inference

Lisete Sousa
lmsousa@fc.ul.pt
Room: 6.4.25

Departamento de Estatística e Investigação Operacional
Faculdade de Ciências da Universidade de Lisboa

2017

Summary

1. Basics of Statistical Inference

The nature of statistical inference

- Statistics deals with data arising from any experiment which result is subject to some random mechanism.
- This means that any time the experiment is performed the result can be different.
- It is not known for certainty what the result will be, but it is known the set of its possible values.

- Experiments are performed in order to draw conclusions.
- However the scientist may want to generalize from that particular experiment to the class of all similar experiments.
- This is the field of inductive inference.
- In inductive inference uncertainty is always present.
- However uncertain inferences can be made, and the degree of uncertainty can be measured if the experiment is performed according with certain principles.

- The theory of Statistics provides techniques for making inductive inferences and for measuring the degree of uncertainty of such inferences.
- Uncertainty is measured in terms of probability.
- For that the result of the experiment is considered to be an observed value of some random variable (or random vector) with a known sample space (the set of possible values to be observed).
- Adequate probabilistic models which may govern the chance mechanism inherent to the observed data are built and relevant inferences are then drawn.

Statistics and their Sampling Distributions

Definition: Statistic

If we have a (hypothetical) sample $X_1, \ldots, X_n$ a statistic is any function of the data which does not depend on unknown parameters.

Usually (but not necessarily) we represent it by $T(X_1, \ldots, X_n)$, or simply T.

For a realized (observed) sample $x_1, \ldots, x_n$ we obtain a realized value of the statistic and represent it by the corresponding lower case letter.

A statistic is itself a random variable and it has a distribution -sampling distribution- which is obtained as a transformation from the proposed joint distribution of the sample $X_1, \ldots, X_n$.

This notion is very important since inference, from a classical point of view, is done with the help of adequately defined *statistics*.

The uncertainty of the inference is measured through the sampling distribution of the chosen Statistic.

Example 1: Long Repeats

- Consider a very very long sequence of DNA of length N and suppose that we are interested in one specific nucleotide, say G and ask whether there is significant evidence of long repeated sequences of this nucleotide.
- Suppose that, if the nucleotides occur at random in the sequence, the probability of the nucleotide G occurring at any site in the sequence is $1 - \theta$.
- We are interested in counting the number of G nucleotides before any one of the nucleotides A, C or T occurs (success).
- This number is the random variable of interest, let us say X, which can be modelled with a geometric distribution with probability of success θ.

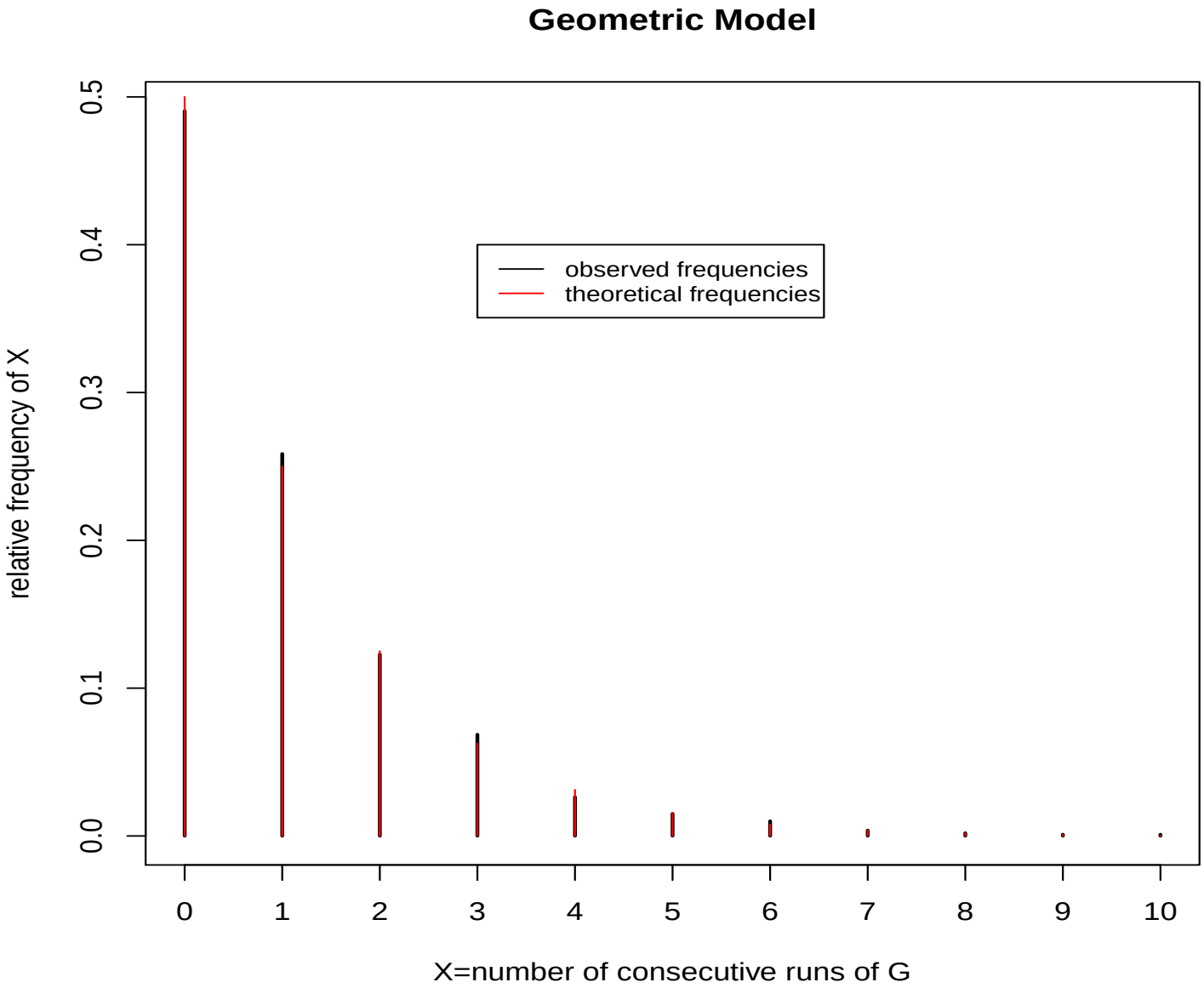
- Scanning the sequence from left to right and counting the number of G nucleotides before any one of the nucleotides A, C or T occurs, we will have a sequence of iid random variables $X_1, \ldots, X_n$. This is what we call a random sample.
- A *statistic* of interest may be $X_{max} = \max(X_1, \ldots, X_n)$.
- To be able to answer the question of interest we will have to be able to model this new random variable.
- The model for the *statistic* is called *sampling distribution of the statistic*

Implementing in R the Example 1:

```
DNA<-factor(c("A","C","G","T"))
p<-c(0.15,0.15,0.50,0.20)
N<-10000
#success happens when A,C,T occurs, hence the probability of success is 1-P(G)=0.5

data<-sample(DNA,N,replace=T,prob=p) #simulates a string of DNA of length N
    according to the specified probabilities in p

loc<-which(data!="G") #finds the locations where a success occurs

x<-c(loc[1]-1,diff(loc)-1) # calculates the number of runs of G

Mean_run<-sum(x)/length(x) #calculates the mean value of
     the number of runs of G; should be around to P(G)/(1-P(G))

table(x)/length(x) #gives the relative frequency of the number of runs

plot(table(x)/length(x),"h",xlab="X=number of consecutive runs of G",ylab="frequenc
of X")
lines(0:13,dgeom(0:13,0.5),col=2,"h") #p.m.f. of the geometric distribution
legend(3,0.4,legend=c("observed frequencies", "theoretical frequencies"),
col=1:2,lty=1,cex=0.8)
```

We obtain the following results

```
> data[1:30]
 [1] T G G T T C G T A T G T G T C G G G T C A T T T G G G G G T
Levels: A C G T
> loc[1:10]
 [1]  1  4  5  6  8  9 10 12 14 15
> x[1:10]
 [1] 0 2 0 0 1 0 0 1 1 0
> Mean_run
[1] 1.013693
> round(table(x)/length(x),5) #gives the relative frequency of the number of runs
x
      0       1       2       3       4       5
0.49054 0.25856 0.12284 0.06867 0.02638 0.01510
      6       7       8       9      10
0.01007 0.00383 0.00201 0.00101 0.00101
```


Geometric Model
observed frequencies
theoretical frequencies
relative frequency of X
0.0
0.1
0.2
0.3
0.4
0.5
0
1
2
3
4
5
6
7
8
9
10
X=number of consecutive runs of G

Example 2: Blood Group Data

- Blood was collected from a random sample of 2128 individuals and the frequency of the four phenotype classes was observed. This is a *Statistic*.

Phenotype	Observed counts
A	725
AB	72
B	258
0	1073

- From the sample we may be interested in estimating the phenotype frequencies of the different blood groups in a "target population".
- The phenotype frequencies (parameters) in the target population are supposed to be unknown and we use the data to obtain an estimated value for those parameters.

Maximum Likelihood Estimation

The notion of likelihood

Suppose that we have a random sample $(X_1, \ldots, X_n)$, i.e., X_i are independent identically distributed (iid) with some common distribution (p.m.f or p.d.f.) $f_X(x|\theta)$. Then the joint distribution is

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{n} f_X(x_i|\theta).$$

Given a particular value of θ (which is usually unknown) this is a function of $x_1, \ldots, x_n$. For each possible value of θ we have a different function.

When $x_1, \ldots, x_n$ is observed we may consider $f(x_1, \ldots, x_n|\theta)$ as only a function of θ. Then we call this function the likelihood of θ. It represents the likelihood of the different values of θ on the light of the observed data. We write then

$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f_X(x_i|\theta).$$

We do not need to have iid random variables to define the likelihood of a parameter. In general if $X_1, \ldots, X_n$ is a random vector with joint distribution $f(x_1, \ldots, x_n|\theta)$ then the *likelihood* of θ for a given observed vector $(x_1, .., x_n)$, is the function with domain Θ defined by

$$L(\theta|x_1, \ldots, x_n) = f(x_1, \ldots, x_n|\theta).$$

Example 1 (cont.): Long Repeats

For the observed sample $x_1, \ldots, x_n$ the likelihood is

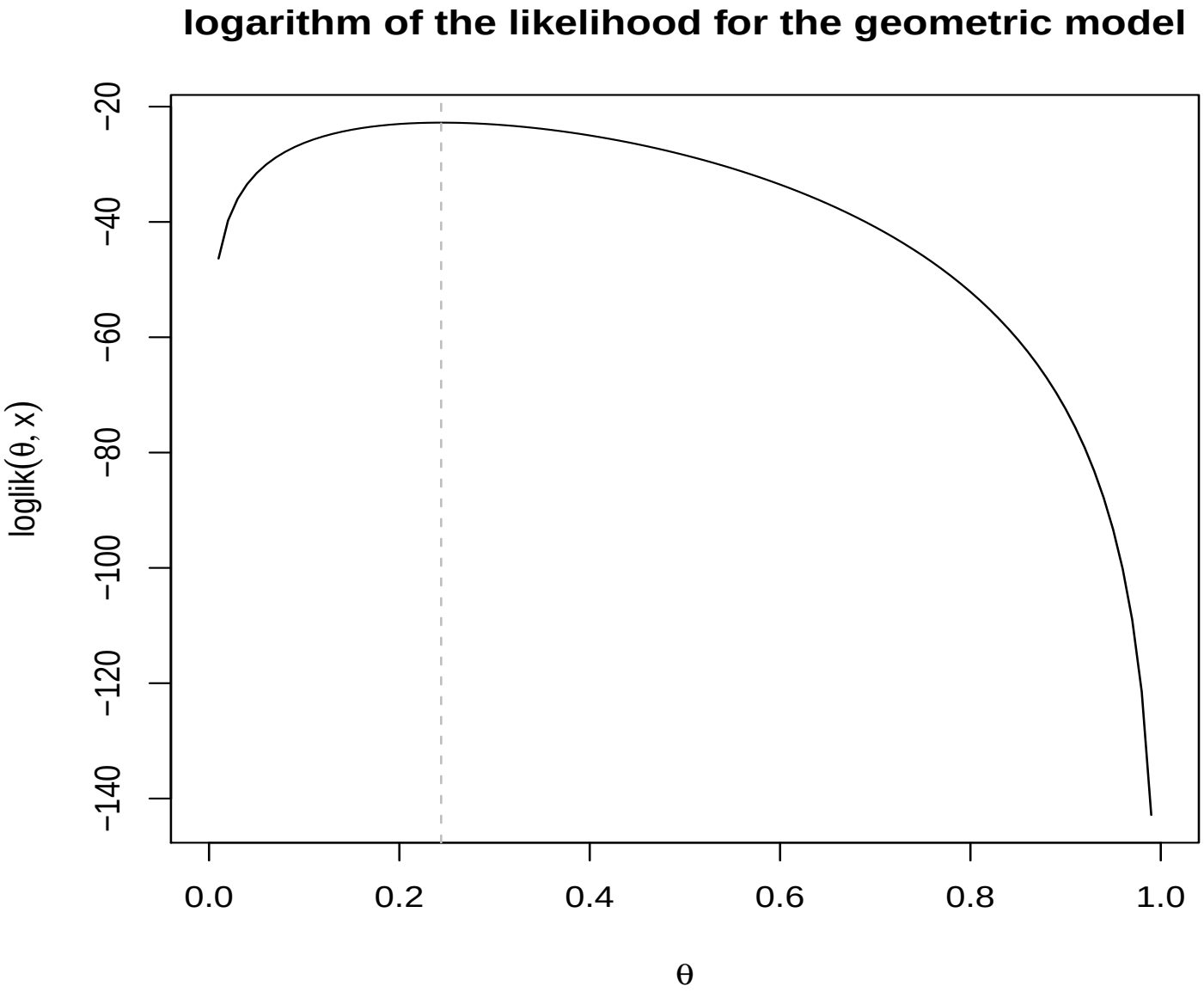
$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} \theta(1-\theta)^{x_i} = \theta^n (1-\theta)^{\sum_{i=1}^{n} x_i},$$

for $\theta \in [0, 1]$. Suppose that we observed the following data

$$2, 5, 3, 1, 0, 4, 6, 5, 3, 2.$$

Then the likelihood is

$$L(\theta|2, 5, 3, 1, 0, 4, 6, 5, 3, 2) = \theta^{10}(1-\theta)^{31}, \quad \theta \in [0, 1].$$


logarithm of the likelihood for the geometric model
loglik(θ, x)
−20
−40
−60
−80
−100
−120
−140
0.0
0.2
0.4
0.6
0.8
1.0
θ

Example 2 (cont.): Blood Group Data

For the observed data $n_A = 1073, n_{AB} = 258, n_B = 72, n_0 = 725$ the likelihood is

$$L(p_A, p_B, p_{AB}, p_0 | n_A = 1073, n_B = 72, n_{AB} = 258, n_0 = 725) =$$
$$= \frac{2128!}{1073!72!258!725!} p_A^{1073} p_B^{72} p_{AB}^{258} p_0^{725}$$

for (p_A, p_B, p_{AB}, p_0) such that $p_i \geq 0, \sum p_i = 1,\ i = A, B, AB, 0.$

Maximum Likelihood Estimation Method

The maximum likelihood estimate (m.l.e.) of a parameter is the value of the parameter (as a function of the data) which maximizes the likelihood of the parameter under the proposed parametric model for the data.

The usual way of obtaining the value for which a function attains a maximum is through differentiation.

Since the likelihood appears, in general, as a product of terms, it is easier to go trough the maximization of the logarithm (natural logarithm) of the likelihood (we call it log-likelihood). As the logarithm is an increasing function, both the likelihood and its log-likelihood attain the maximum at the same point.

Example 1 (cont.): Long Repeats

Although here the inference problem of interest was not to obtain an estimate to the probability of success, we will apply the maximum likelihood method to obtain an estimate for it.

For the observed sample $x_1, \ldots, x_n$ the likelihood is

$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} \theta(1-\theta)^{x_i} = \theta^n (1-\theta)^{\sum_{i=1}^{n} x_i},$$

for $\theta \in [0, 1]$.

Hence the log-likelihood is

$$\log L(\theta|x_1, \ldots, x_n) = n \log \theta + \log(1-\theta) \sum_{i=1}^{n} x_i.$$

Differentiating with respect to θ and equating to zero we get as m.l.e. for θ

$$\hat{\theta} = \frac{n}{n + \sum_{i=1}^{n} x_i}.$$

With the observed data

$$2, 5, 3, 1, 0, 4, 6, 5, 3, 2,$$

we get

$$\hat{\theta} = \frac{10}{41} \approx 0.24.$$

Example 2 (cont.): Blood Group Data

There were two different estimation problems in this example.

First we wanted to estimate the probabilities for each blood group, namely p_A, p_B, p_{AB}, p_0.

Here we have to remember that we have to impose the condition that $p_A + p_B + p_{AB} + p_0 = 1$.

The best way to deal with the problem is to substitute in the likelihood, e.g., p_0 by $1 - p_A - p_B - p_{AB}$.

Then we have to differentiate the log-likelihood with respect to each parameter p_A, p_B, p_{AB} and equate to zero. We get a system of three equations which solution is

$$\begin{aligned} \hat{p}_A &= \frac{n_A}{n} \\ \hat{p}_B &= \frac{n_B}{n} \\ \hat{p}_{AB} &= \frac{n_{AB}}{n}. \end{aligned}$$

Consequently

$$\hat{p}_0 = 1 - \hat{p}_A - \hat{p}_B - \hat{p}_{AB} = \frac{n_0}{n}.$$

Again this is an expected result. The frequencies of the blood groups in the population are estimated by their relative frequencies in the sample.

For our data set $n_A = 725, n_{AB} = 72, n_B = 258, n_0 = 1073$ we obtain

$$p_A \approx 0.34, p_{AB} \approx 0.04, p_B \approx 0.12, p_0 \approx 0.5.$$

Example 2 (variation):

Another problem of interest here is the estimation of the probabilities of the occurrence of the alleles $A, B, 0$ in the population, namely p_A^*, p_B^*, p_0^*.

According to the genetic model we have

Genotype	Phenotype	Observed frequency	Probability
AA	A	n_A	$(p_A^*)^2$
$A0$	A		$2p_A^*p_0^*$
AB	AB	n_{AB}	$2p_A^*p_B^*$
BB	B	n_B	$(p_B^*)^2$
$B0$	B		$2p_B^*p_0^*$
00	0	n_0	$(p_0^*)^2$

The likelihood in this case is

$$\begin{aligned} & L(p_A^*, p_B^* | n_A, n_B, n_{AB}, n_0) \\ = & \frac{n!}{n_A! n_B! n_{AB}! n_0!} \times [p_A^*(2 - p_A^* - 2p_B^*)]^{n_A} \\ \times & [p_B^*(2 - p_B^* - 2p_A^*)]^{n_B} [2p_A^* p_B^*]^{n_{AB}} [1 - p_A^* - p_B^*]^{2n_0} \end{aligned}$$

Differentiating the log-likelihood with respect to p_A^* and p_B^* we get the two equations

$$\begin{aligned} \frac{\partial \log L(p_A^*, p_B^* | n_A, n_B, n_{AB}, n_0)}{\partial p_A^*} = & \frac{n_{AB}}{p_A^*} + \frac{n_A(2 - 2p_A^* - 2p_B^*)}{p_A^*(2 - p_A^* - 2p_B^*)} - \\ - & \frac{2n_B}{2 - 2p_A^* - p_B^*} - \frac{2n_0}{1 - p_A^* - p_B^*} \end{aligned}$$

$$\frac{\partial \log L(p_A^*, p_B^* | n_A, n_B, n_{AB}, n_0)}{\partial p_B^*} = \frac{n_{AB}}{p_B^*} + \frac{n_B(2 - 2p_A^* - 2p_B^*)}{p_B^*(2 - 2p_A^* - p_B^*)} - \frac{2n_A}{2 - p_A^* - 2p_B^*} - \frac{2n_0}{1 - p_A^* - p_B^*}$$

There is no explicit solution for this system of two equations. The solution has to be obtained by iterative methods, such as Newton-Raphson or EM algorithm.

For our data set $n_A = 725, n_{AB} = 72, n_B = 258, n_0 = 1073$ we obtain

$$p_A^* \approx 0.21, p_B^* \approx 0.08, p_0^* \approx 0.71.$$

The result was obtained by using function `maxNR` (Newton-Raphson) from package `maxLik`:

```
> na<-725
> nb<-258
> nab<-72
> n0<-1073
> n<-na+nb+nab+n0
> f<-function(p){(na+nab)*log(p[1])+na*log(2-p[1]-2*p[2])+
+ (nb+nab)*log(p[2])+nb*log(2-p[2]-2*p[1])+nab*log(2)+
+ 2*n0*log(1-p[1]-p[2])}
> summary(maxNR(f,start=c(0.6,0.3))
```

```
--------------------------------------------
Newton-Raphson maximisation
Number of iterations:  7
Return code:  1
gradient close to zero.  May be a solution
Function value:  -2303.550
Estimates:
     estimate      gradient
[1,] 0.20913065            0
[2,] 0.08080101            0
--------------------------------------------
```

2. Confidence Interval Estimation

Introduction

- Instead of giving a point estimator for a parameter we may instead give an interval estimator which contains the true value of the parameter with a certain probability.
- A **confidence interval** for a parameter is an interval of numbers within which we expect the true value of the population parameter to be contained. The endpoints of the interval are computed based on sample information.
- If confidence intervals are constructed across many separate data analyses of repeated (and possibly different) experiments, the proportion of such intervals that contain the true value of the parameter will match the confidence level.

How to construct a confidence interval (CI)?

Suppose $X_1, \ldots, X_n$ random variables independent identically distributed.

1. Identify the parameter of interest;
2. Determine the confidence level $(1 - \alpha)100\%$;
   Note: if not specified, set the confidence to 95%
3. Check the assumptions;
4. Identify the required formula for the CI;
5. Identify the descriptive statistics needed, from the sample $x_1, \ldots, x_n$;
6. Find the required critical value (probability quantile);
7. Compute de CI based on formula in step 4.

Some of the Most Used CI

$(1-\alpha)100\%$ **CI for the mean** μ

1. Parameter: μ (expected value).
2. Confidence level: $(1-\alpha)100\%$.
3. a) Normal population, σ known;
   b) Normal population, σ unknown;
   c) Population not normal, but $n \geq 30$.
4. Formula for the CI:
   a) $\bar{x} \pm z_{critical} \frac{\sigma}{\sqrt{n}}$
   b) $\bar{x} \pm t_{critical} \frac{s}{\sqrt{n}}$
   c) $\bar{x} \pm z_{critical} \frac{s}{\sqrt{n}}$ (approximate)

5 Descriptive statistics needed:
sample mean $\bar{x}$;
standard deviation s;
sample size n.

6 Critical value:
a) $\alpha = 0.10 \text{ -> } z_{0.95} = 1.645$
$\alpha = 0.05 \text{ -> } z_{0.975} = 1.960$
$\alpha = 0.01 \text{ -> } z_{0.995} = 2.576$
b) $\alpha = 0.10 \text{ -> } t_{n-1;0.95}$
$\alpha = 0.05 \text{ -> } t_{n-1;0.975}$
$\alpha = 0.01 \text{ -> } t_{n-1;0.995}$
c) Same as in a).

$(1-\alpha)100\%$ CI for the proportion p

1. Parameter: p (proportion).
2. Confidence level: $(1-\alpha)100\%$.
3. Assumptions: $np \geq 10$ and $n(1-p) \geq 10$.
4. Formula for the CI:
   $\hat{p} \pm z_{critical}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ (approximate)
5. Descriptive statistics needed:
   sample proportion $\hat{p}$;
   sample size n.
6. Critical value:
   $\alpha = 0.10 \text{ -> } z_{0.95} = 1.645$
   $\alpha = 0.05 \text{ -> } z_{0.975} = 1.960$
   $\alpha = 0.01 \text{ -> } z_{0.995} = 2.576$.

$(1-\alpha)100\%$ CI for the variance σ^2

1. Parameter: σ^2.
2. Confidence level: $(1-\alpha)100\%$.
3. Assumptions: normal population.
4. Formula for the CI:
$$\left(\frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}; \frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}}\right)$$
5. Descriptive statistics needed:
sample standard deviation s;
sample size n.
6. Critical value:
$\alpha = 0.10 \text{ -> } \chi^2_{n-1;0.05}$ and $\chi^2_{n-1;0.95}$
$\alpha = 0.05 \text{ -> } \chi^2_{n-1;0.025}$ and $\chi^2_{n-1;0.975}$
$\alpha = 0.01 \text{ -> } \chi^2_{n-1;0.005}$ and $\chi^2_{n-1;0.995}$

Consider now two random variables, X_A and X_B from normal populations A and B, with parameters (μ_A, σ_A) and (μ_B, σ_B), respectively; and two random samples from each population $X_{A1}, \ldots, X_{An_A}$ and $X_{B1}, \ldots, X_{Bn_B}$.

$(1-\alpha)100\%$ CI for the difference between means:

σ_A and σ_B known, $\bar{x}_A - \bar{x}_B \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$

$\sigma_A = \sigma_B = \sigma$ unknown, $\bar{x}_A - \bar{x}_B \pm t_{n_A+n_B-2;1-\alpha/2} s_p \sqrt{\frac{n_A+n_B}{n_A n_B}}$

and $s_p^2 = \frac{(n_A-1)s_A^2+(n_B-1)s_B^2}{n_A+n_B-2}$, with s_A^2 and s_B^2 the variances of samples A and B, respectively.

Examples in R

Example 3: CI for the expected value (μ)

(Normal population)

Consider a sample of 20 observations

$\underset{\sim}{x} = (32.81, 37.04, 37.21, 31.15, 26.97, 26.58, 31.85, 30.09, 28.63, 25.12, 31.67, 28.26, 28.57, 37.39, 30.55, 32.98, 24.52, 28.28, 27.37, 26.35)$.

Suppose we want to find the 99% confidence interval for μ. Since the variance σ^2 is unknown, the CI is given by

$$\overline{x} \pm t_{n-1;1-\alpha/2} \frac{s}{\sqrt{n}} .$$

For the data consider in the Example, the 99% CI for μ is,

$$(27.692 \ , \ 32.647)$$

This interval can be calculated easily in R by using the function t.test:

```
> x <- c(32.81,37.04,37.21,31.15,26.97,26.58,31.85,
+ 30.09,28.63,25.12,31.67,28.26,28.57,37.39,30.55,
+ 32.98,24.52,28.28,27.37,26.35)
> t.test(x,alternative="two.sided",conf.level=0.99)$conf.int
```

```
[1] 27.69154 32.64746
attr(,"conf.level")
```

```
[1] 0.99
```

Example 4: CI for the ratio of two variances (σ_x^2/σ_y^2)

(Two Normal and independent populations)

Consider another sample of 20 observations $y = (38.14, 39.07, 37.29, 41.20, 40.31, 39.07, 34.99, 36.82, 35.23, 37.97, 36.21, 45.13, 35.98, 36.55, 37.45, 40.23, 38.45, 45.01, 36.94, 42.09)$. Now, we want to find the 95% confidence interval for σ_x^2/σ_y^2, which is given by

$$\left(\frac{s_x^2 F_{n_y-1,n_x-1;\alpha/2}}{s_y^2}, \frac{s_x^2 F_{n_y-1,n_x-1;1-\alpha/2}}{s_y^2} \right).$$

For the data in this Example and Example 3, the 95% CI for σ_x^2/σ_y^2 is,

$$(0.709 \ , \ 4.523)$$

This interval can be calculated easily in R by using the function `var.test`:

```
> y <- c(38.14,39.07,37.29,41.20,40.31,39.07,34.99,
+ 36.82,35.23,37.97,36.21,45.13,35.98,36.55,37.45,
+ 40.23, 38.45, 45.01, 36.94, 42.09))
> var.test(x,y)$conf.int
```

```
[1] 0.7086474 4.5232640
attr(,"conf.level")
```

```
[1] 0.95
```

Example 5: CI for the difference of expected values ($\mu_x - \mu_y$)

(Two Normal and independent populations)

From Example 4, we can consider the populations' variances to be equal (we will see further, why) at a significance level of 0.05. Since that variance is unknown, the 95% confidence interval for $\mu_x - \mu_y$ (expression in frame 38) is,

$$(-10.726\ ,\ -6.348)$$

This interval can be calculated easily in R by using the function t.test:

```
> y <- c(38.14,39.07,37.29,41.20,40.31,39.07,34.99,
+ 36.82,35.23,37.97,36.21,45.13,35.98,36.55,37.45,
+ 40.23, 38.45, 45.01, 36.94, 42.09))
>
t.test(x,y,alternative="two.sided",var.equal=T,paired=F)$conf.int
```

```
[1] -10.725979 -6.348021
attr(,"conf.level")
```

```
[1] 0.95
```

3. Hypothesis Testing

Introduction

FAQ

- What is statistical hypothesis?
- What is the null hypothesis?
- What does a p-value mean?
- What does it mean to say that a finding is *statistically significant*?
- What does it mean "It was assumed a 5% significance level"?
- A **hypothesis test** is a procedure for determining if an assertion about a characteristic of a population is reasonable.

For Example:

Someone says that the average weight of the mice used in experiments in FCUL's laboratories is 20g.

How would you decide whether this statement is true?

- find all the mice in the laboratories of FCUL and weight them all, or
- find out the average weight of mice at a small number of randomly chosen laboratories and compare the average weight to 20g.

Suppose your sample average was 19g. Is this 1g difference a significant result, or is the original assertion incorrect?

Terminology

- The null hypothesis, H_0, is usually a hypothesis of agreement with conditions presumed to be true. A null hypothesis is either rejected or not rejected.
  - $H_0 : \mu = 20g$

- The alternative hypothesis, H_1, represents the statement that the researcher wants to prove. (It determines the rejection area).
  - $H_1 : \mu > 20g$
  - $H_1 : \mu < 20g$
  - $H_1 : \mu \neq 20g$

- When H_0 is not rejected we say that "the data do not give enough evidence to reject the null hypothesis".
- When the H_0 is rejected we say that "the data at hand are not compatible with the null hypothesis stated", but are supportive of some other hypothesis which is the one which is established as the alternative hypothesis.
- The significance level alpha (α)
  - $P[\textit{Type I error}] = P[\textit{Rej } H_0 | H_0 \textit{ is True}]$
  - $\alpha = 0.05$: the probability of incorrectly rejecting the null hypothesis when it is actually true is 0.05.
  - If you need more protection from this error, then choose a lower value of α.

Error Types

It is important to bear in mind the following:

The fact that a hypothesis is not rejected does not mean that it is true. We can only say that the hypothesis is supported by the available data.

When we formulate a hypothesis testing problem there are two types of errors we can commit.

- When we reject a true null hypothesis, we say that we commit a type I error. The probability of committing a type I error is represented by α.
- When we fail to reject a false null hypothesis then we say that a type II error is committed. The probability of committing a type II error is represented by β.

Error types and Test power

	Condition of null hypothesis	
Possible action	True	False
Fail to reject H_0	Correct $(1-\alpha)$	Type II error (β)
Reject H_0	Type I error (α)	Correct $(1-\beta)$

- Type I Error (with probability α): calling genes as differentially expressed when they are NOT
- Type II Error (with probability β): NOT calling genes as differentially expressed when they ARE
- Power of a test: $1-\beta$. The odds of confirming our theory correctly

Error type I vs. Error type II

What we would expect is to have a lower alpha (α) and higher power (1-β), but:

- the lower the α, the lower the power; the higher the α, the higher the power.
- the lower the α, the less likely it is that you will make a Type I Error (i.e., reject the null when it is true)
- the lower the α, the more "rigorous" the test.

Running a Hypothesis Test

Steps in hypothesis testing

- Determine the null and alternative hypothesis, using mathematical expressions if applicable.
- Select a significance level (α).
- Take a random sample from the population of interest.
- Calculate a test statistic from the sample that provides information about the null hypothesis.
- Decision (by classical definition or with p-value).
- Conclusion.

Decision

The decision as to whether H_0 is rejected or not rejected is made on the basis of data using the result of a *test statistic*, say $T(X_1, \ldots, X_n)$, or for short, T.

A good test statistic should be such that the probability of committing a type I error is as small as possible.

How to proceed once a test statistic T is chosen?

Rejection regions

The set of possible values the test statistic T can take is divided into two regions

- The acceptance region $\mathcal{A}$; observed values of the test statistic T, falling in this region lead to non-rejection of the null hypothesis.
- The rejection region $\mathcal{R}$; observed values of the test statistic T falling in this region lead to the rejection of the null hypothesis.
- The alternative hypothesis tells the tale (1-tailed vs 2-tailed tests)

This is accomplished either with the knowledge of the exact sampling distribution of the test statistic, under the null hypothesis, or with the help of asymptotic theory.

The rejection regions relatively to a significance level α, are usually of one of the types,

$$\mathcal{R}_\alpha = \{t : t > t_{1-\alpha}\}, \qquad \mathcal{R}_\alpha = \{t : t < t_\alpha\},$$

$$\mathcal{R}_\alpha = \{t : t < t_1 \quad \text{or} \quad t > t_2\},$$

Usually, for reasons of symmetry, we choose equal probabilities on the tails and hence, we choose t_1 and t_2 such that

$$P(T < t_1|H_0) = P(T > t_2|H_0) = \frac{\alpha}{2}.$$

In a hypothesis testing problem as it was stated, the significance level α, which is a measure of the uncertainty associated with our inference, is fixed beforehand. The usual values chosen for α are 0.10, 0.05, 0.01. The size of the rejection region depends on this value. If for a certain test statistic with rejection region α, call it $\mathcal{R}_\alpha$ we have the relation:

$$\mathcal{R}_{\alpha_1} \supset \mathcal{R}_{\alpha_2} \Leftrightarrow \alpha_1 > \alpha_2.$$

P-value

- A p-value is a measure of how much evidence we have against the null hypothesis.
- The smaller the p-value, the more evidence we have against H_0
- The p-value measures consistency by calculating the probability of observing the results from your sample of data or a sample with results more extreme, assuming the null hypothesis is true. The smaller the p-value, the greater the inconsistency.
- A large p-value should not automatically be constructed as evidence in support of the null hypothesis. Perhaps the failure to reject the null hypothesis was caused by an inadequate sample size.
- You should also be cautious about a small p-value, but for different reasons. In some situations, the sample size is so large that even differences that are trivial from a biological perspective can still achieve statistical significance.

Notes:

1. You should not interpret the p-value as the probability that the null hypothesis is true. Such an interpretation is problematic because a hypothesis is not a random event that can have a probability.

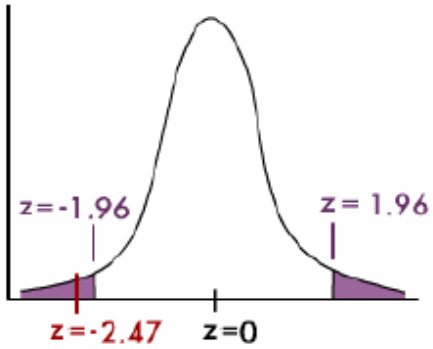
2. Bayesian statistics provides an alternative framework that allows you to assign probabilities to hypotheses and to modify these probabilities on the basis of the data that you collect.

If for some specific data we observe t_{obs} as the value for the test statistic, then the p-value is the probability of observing a value for the test statistic as "extreme" as t_{obs}. For each of the type of the rejection regions considered above we have:

$$\begin{aligned} p\text{-value} &= P(T > t_{obs}|H_0\text{true}), \\ p\text{-value} &= P(T < t_{obs}|H_0\text{true}), \\ p\text{-value} &= 2\ P(T > \text{abs}(t_{obs})|H_0\text{true}) \end{aligned}$$

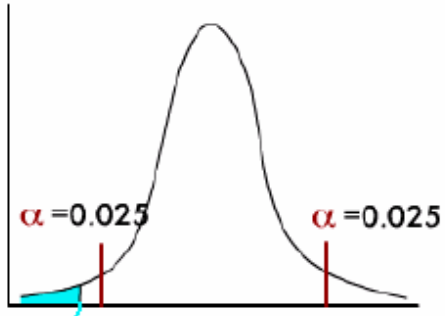
where $\text{abs}(t_{obs})$ means the absolute value of t_{obs}.

The Classical Approach



Conclusion: since the z value of the test statistic (-2.47) is less than the critical value of z=-1.96, we reject the null hypothesis.

The *P*-Value Approach



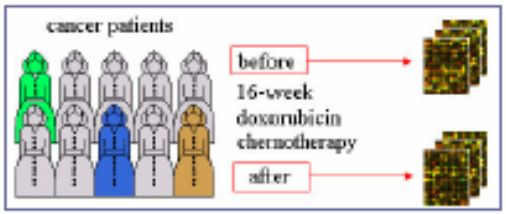
P -value = 0.0068 times 2 (for a 2-sided test) = 0.0136

Conclusion: since the *P* -value of 0.0136 is less than the significance level of α=0.05, we reject the null hypothesis.

Confidence intervals and p-values

- The two statistical concepts are complementary.
- Taken in isolation, p-values provide a measure of the statistical plausibility of a result.
- With a defined level of significance, p-values allow a decision about the rejection or maintenance of a previously formulated null hypothesis in confirmatory studies.
- Confidence intervals provide an adequately plausible range for the true value related to the measurement of the point estimate.
- Confidence intervals give an estimate of the precision with which a statistic estimates a population value.
- If the alternative hypothesis is unilateral **it is not possible** to compare the results with confidence intervals.

Types of hypothesis tests



Dependent samples

Independent samples

Comparison	Two Groups		More than two Groups
Hypothesis Testing	**Paired data**	**Unpaired data**	**Complex data**
Parametric (variance equal)	One sample t-test	Two-sample t-test	One-Way Analysis of Variance (ANOVA)
Parametric (variance not equal)	Welch t-test		Welch ANOVA
Non-Parametric	Wilcoxon Signed-Rank Test	Wilcoxon Rank-Sum Test (Mann-Whitney U Test)	Kruskal-Wallis Test

Parametric Tests

One-sample t-test

- The one-sample t-test compares the mean score of a sample to a known value. Usually, the known value is a population mean.
- Assumption: the variable is normally distributed.
- $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$
- Test statistic: under H_0, $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} \frown N(0,1)$ or $T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} \frown t_{n-1}$, if σ is known or unknown, respectively.
- Reject H_0 if $|z_{obs}| > z_{1-\alpha/2}$ or if $|t_{obs}| > t_{n-1;1-\alpha/2}$, respectively
- p-value$=2 \times P(|Z| > z_{obs})$ or p-value$=2 \times P(|T| > t_{obs})$, respectively

NOTES:

1. The distribution of the data being tested is normal.
   - For **paired** t-test, it is the distribution of the subtracted data that must be normal. In R use the argument paired=TRUE.
   - For **unpaired** t-test, the distribution of both data sets must be normal. In R use the (default) argument paired=FALSE.
   - Plots: Histogram, Density Plot, QQ Plot.
   - Test for Normality: Kolmogorov-Smirnov test, Shapiro-Wilk test.
2. Homoscedasticity: the variances of both populations are equal.
   - If the two populations have **equal** variances, then the two-sample t-test may be used. Variance ($\sigma^2 = \sigma_A^2 = \sigma_B^2$) is estimated by s_p^2 (see frame 38). In R use the argument var.equal=TRUE.
   - If the two populations have **unequal** variances, then use the two-sample unequal variances t-test (Welch's t-test). In this case, σ_A^2 and σ_B^2 are estimated by s_A^2 and s_B^2, respectively, and the degrees of freedom are given according to Welch's modification. In R use the (default) argument var.equal=FALSE.
   - Test for equality of the two variances: variance ratio F-test.

Comparing more than two groups: one-way ANOVA

Used to compare the means of more than two independent groups. Instead of a t-statistic, ANOVA uses a F statistic and its $p - value$ to evaluate the null hypothesis that all of several population means are equal.

In **one-way ANOVA** we classify the populations of interest according to a single categorical explanatory variable that we call a factor.

Assumptions:

1. The distribution of the means by group are normal with equal variances.
   - Bartlett's test (1937)
   - Levene's test (Levene 1960)
   - O'Brien (1979)

2. Sample sizes between groups do not have to be equal, but large differences in sample sizes by group may effect the outcome of the multiple comparisons tests.

(1) The hypotheses for the comparison of independent groups are:

$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ (means of the all groups are equal)

H_1 : not all of μ_i are equal

(2) The ANOVA table:

To assess whether several populations all have the same mean, we compare the variation **among** the means of several groups with the variation **within** groups. Because we are comparing variation, the method is called analysis of variance.

Variation is expressed by **sums of squares**. Each sum of squares is the sum of the squares of a set of deviations that expresses a source of variation.

Source of variation	SS	d.f.	Mean squared	F-ratio
Among groups	SSA	$k-1$	$MSA = \frac{SSA}{k-1}$	$F = \frac{MSA}{MSE}$
Within groups	SSE	$N-k$	$MSE = \frac{SSE}{N-k}$	
Total	$SST = SSA + SSE$	$N-1$	$MST = \frac{SST}{N-1}$	

Where:

1. k is the number of groups;
2. n_i $(i = 1, ..., k)$ is the size of group i;
3. N is the total sample size: $N = n_1 + ... + n_k$;

4. $SSA = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{\bar{x}})^2$;

5. $SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$;

When H_0 is true, the statistic F has the $F_{(k-1,N-k)}$ distribution.

(3) Decision:

When H_0 is true, the statistic tends to be small. We reject H_0 in favour of H_1 if the observed value of the statistic F, i.e. F_0, is sufficiently large.

Thus, we reject H_0 for a significance level α, if

$$p - value = P(F_{(k-1,N-k)} > F_0) < \alpha.$$

Example 7:

The following samples refer to the weights (kg) of cows, according to a certain treatment:

A	30.28	27.58	27.91	29.33				
B	34.26	32.55	21.78	25.59	35.08	26.86		
C	39.47	30.15	33.40	27.38	30.39	25.85	29.11	26.22
D	33.54	30.40	29.60	28.82	30.70	30.83	33.84	

Consider $\alpha = 0.01$.

We start by testing the equality of the variances using Bartlett's test:

$$H_0 : \sigma_A^2 = \sigma_B^2 = \sigma_C^2 = \sigma_D^2$$

The p-value obtained is 0.02994 and so, we do not reject the null hypothesis at the significance level of 0.01.

```
> a<-c(30.28,27.58,27.91,29.33)
> b<-c(34.26,32.55,21.78,25.59,35.08,26.86)
> c<-c(39.47,30.15,33.40,27.38,30.39,25.85,29.11,26.22)
> d<-c(33.54,30.40,29.60,28.82,30.70,30.83,33.84)
> observ<-c(a,b,c,d)
> treatm<-factor(rep(c("a","b","c","d"),c(4,6,8,7)))
> bartlett.test(observ~treatm)

        Bartlett test of homogeneity of variances
data:  observ and treatm
Bartlett's K-squared = 8.952, df = 3, p-value = 0.02994
```

Thus, we may proceed with the ANOVA:

Hypotheses: $H_0 : \mu_A = \mu_B = \mu_C = \mu_D$ (treatment effects are equal)

Decision: Since $p - value = P(F_{(3,21)} > 0.41) = 0.75 > 0.01$, for the usual levels of significance, do not reject H_0. This sample does not give evidence for differences between treatment effects.

```
> aov(observ~treatm)

Call:
   aov(formula = observ ~ treatm)
Terms:
                    treatm Residuals
Sum of Squares   17.49895 311.30299
Deg. of Freedom         3        21
Residual standard error: 3.850189
Estimated effects may be unbalanced

> summary(aov(observ treatm))

            Df Sum Sq Mean Sq F value Pr(>F)
treatm       3   17.5   5.833   0.393  0.759
Residuals   21  311.3  14.824
```

Comparing Two Proportions

Suppose that we have two DNA sequences and we want to test the hypothesis that the nucleotide A appears in both sequences with the same frequency.

If we call p_1 the probability of occurrence of nucleotide A in the first sequence and p_2 the probability of the occurrence in the second sequence, then we want to test

$$H_0 : p_1 = p_2 \quad \textit{versus} \quad H_1 : p_1 \neq p_2.$$

Assume that the sequences were independently generated and let X_1 and X_2 be the number of A nucleotides in subsequence of size n_1 and n_2 from the first and second sequences, respectively.

Again if n_1 and n_2 are large, $\frac{X_1}{n_1}$ and $\frac{X_2}{n_2}$ are approximately normal distributed.

Under the null hypothesis $p_1 = p_2$ they both have expected value p, (the common value of p_1, p_2) and variances $\frac{p(1-p)}{n_1}$ and $\frac{p(1-p)}{n_2}$, respectively.

Since they are independent then $\bar{X}_1 - \bar{X}_2$ is also approximately normally distributed (under H_0) with expected value 0 and variance $\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_2}$.

Hence we can use as test statistic

$$Z = \frac{\frac{X_1}{n_1} - \frac{X_2}{n_2}}{\hat{\sigma}_{\hat{p}_1 - \hat{p}_2}},$$

which under the null hypothesis follows approximately a standard normal distribution.

Here $\hat{\sigma}_{\hat{p}_1 - \hat{p}_2}$ is an estimate of standard deviation of $\frac{X_1}{n_1} - \frac{X_2}{n_2}$.

To perform the test we can do as follows:

1. Compute $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$, where x_1 and x_2 are observed values of X_1 and X_2 respectively.
2. Compute $\hat{p} = \frac{x_1+x_2}{n_1+n_2}$ as estimate for the common p under the null hypothesis.
3. Compute $\hat{\sigma}_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n_1} + \frac{\hat{p}(1-\hat{p})}{n_2}}$
4. Compute the observed value of the test statistic Z as

$$z_{obs} = \frac{\frac{x_1}{n_1} - \frac{x_2}{n_2}}{\hat{\sigma}_{\hat{p}_1-\hat{p}_2}}$$

5. If this value is outside the interval $[z_{\frac{\alpha}{2}}, z_{1-\frac{\alpha}{2}}]$ then reject the hypothesis $p_1 = p_2$ at the significance level α. Otherwise do not reject the equality of proportions.

Example 8:

As an example, suppose that for $n_1 = n_2 = 100$ we obtained $x_1 = 25, x_2 = 27$.

Then $\hat{p}_1 = 0.25$, $\hat{p}_2 = 0.27$, $\hat{p} = 0.26$, $\hat{\sigma}_{\hat{p}_1 - \hat{p}_2} = 0.06203$ and $z_{obs} = -0.322$ which is inside $[-1.96, 1.96]$.

Hence we would not reject the null hypothesis $H_0 : p_1 = p_2$ at the 5% significance level.

Approximate calculations using R functions

As n_1 and n_2 are large an approximate value of z_{obs} can be obtained by using functions `t.test` and `z.test`, however, we have to be careful about how to introduce the data:

t.test

```
> x<-c(rep(1,25),rep(0,75)); y<-c(rep(1,27),rep(0,73))
> t.test(x,y)

Welch Two Sample t-test

data:  x and y
t = -0.3209, df = 197.877, p-value = 0.7486
alternative hypothesis:true difference in means is not equal to 0
95 percent confidence interval:
-0.1429135 0.1029135
sample estimates:
mean of x mean of y
     0.25      0.27
```

Note that, in this case, $\hat{\sigma}_{\frac{x_1}{n_1}-\frac{x_2}{n_2}}$ is equal to $\sqrt{\frac{\hat{\sigma_1^2}}{n_1}+\frac{\hat{\sigma_2^2}}{n_2}}$, which is similar to $\hat{\sigma}_{\hat{p}_1-\hat{p}_2} = 0.06203$:

```
> sd<-sqrt((var(x1)+var(x2))/100); sd
[1] 0.06232855
```

z.test

```
> library(BSDA)
> z.test(x,y,sigma.x=sqrt(var(x)),sigma.y=sqrt(var(y)))

        Two-sample z-Test

data:  x and y
z = -0.3209, p-value = 0.7483
alternative hypothesis:true difference in means is not equal to 0
95 percent confidence interval:
 -0.1421617  0.1021617
sample estimates:
mean of x mean of y
     0.25      0.27
```

Performing a Chi-Square test for homogeneity

It is also possible two compare two proportions by using R function `prop.test`, but here the considered Statistic is Chi-square distributed.

```
> suc<-c(sum(x1),sum(x2)) # vector - total successes per group
> trials<-c(100,100)      # vector - total trials per group
> prop.test(suc,trials)
> # or
> fail<-100-suc                 # vector - total failures per group
> x<-matrix(c(suc,fail),2,2) # contingency table
> prop.test(x)

   2-sample test for equality of proportions with continuity correction
data:  x
X-squared = 0.026, df = 1, p-value = 0.8719
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.1515494  0.1115494
sample estimates:
prop 1 prop 2
  0.25   0.27
```

R Functions - Summary

In case of populations Normally distributed, with mean μ and variance σ^2, the following R functions may be used for inferences on the parameters:

H_0	R function	R package	Assumptions
$\mu = \mu_0$ or $\mu_1 - \mu_2 = \mu_0$	`t.test` `z.test`	`stats` `BSDA`	unknown variance known variance
$\mu_1 = \ldots = \mu_k$	`aov` or `anova`	`stats`	equal, unknown var
$\sigma^2 = \sigma_0^2$ $\sigma_1^2/\sigma_2^2 = \sigma_0^2$	`var.test`	`stats`	

NOTE: In order to apply these tests, it is important to check normality. The most appropriate R function to test normality is `shapiro.test` (Shapiro-Wilk test).

Acknowledgements:

Carina Silva-Fortes (ESTeSL – IPL), for allowing the use of some material produced by both of us in previous courses.

Bibliography:

Keen, K. J. (2010). *Graphics for Statistics and Data Analysis with R*. Chapman & Hall.

Lumley, T. (2010). *Complex Surveys: a Guide to Analysis Using R*. Wiley.

Maindonald, J. (2010). *Data Analysis and Graphics Using R: an Example-Based Approach*. Cambridge University Press.