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 and their Sampling Distributions
   - Maximum Likelihood Estimation
2. Confidence Interval Estimation
   - Introduction
   - Some of the Most Used CI
   - Examples in `R`
3. Hypothesis Testing
   - Introduction
   - Parametric Tests

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.

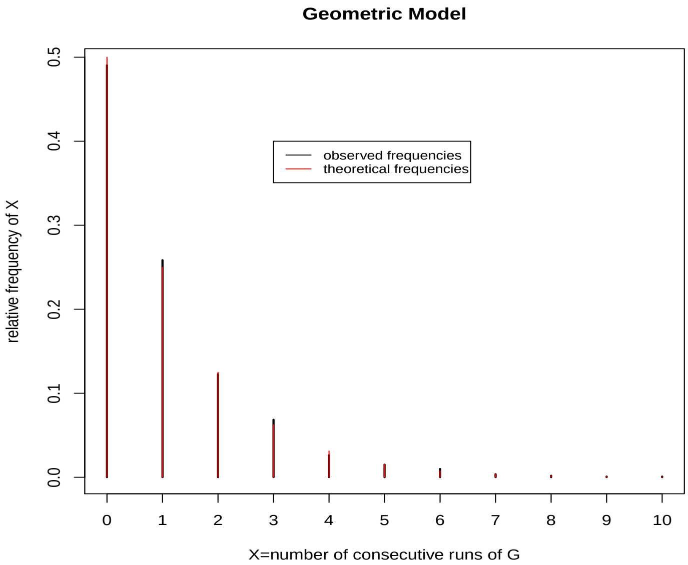
- Scanning the sequence from left to right and counting the number of G nucleotides before any one of the nucleotides A, C or T occurs, we will have a sequence of iid random variables $X_1, \ldots, X_n$. This is what we call a random sample.
- A *statistic* of interest may be $X_{max} = \max(X_1, \ldots,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

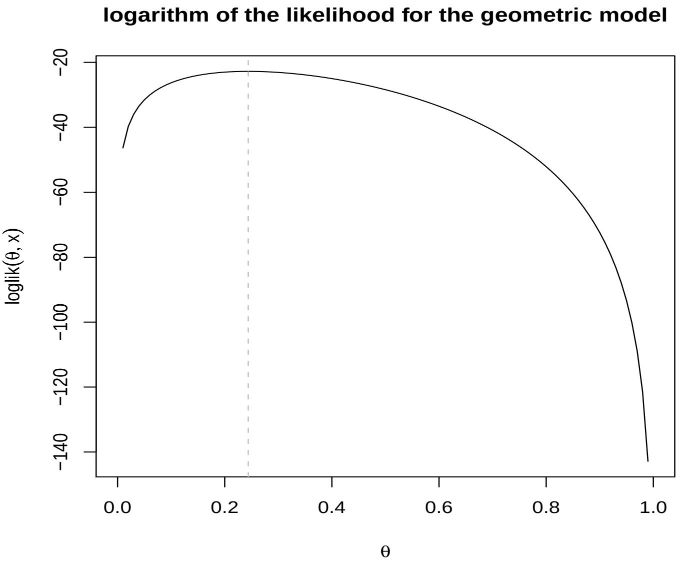
$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} \theta(1-\theta)^{x_i} = \theta^n \, (1-\theta)^{\sum_{i=1}^{n} x_i},$$

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

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

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

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

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

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

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

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

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

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) $\alpha = 0.10 -> z_{0.95} = 1.645$
$\alpha = 0.05 -> z_{0.975} = 1.960$
$\alpha = 0.01 -> z_{0.995} = 2.576$
b) $\alpha = 0.10 -> t_{n-1;0.95}$
$\alpha = 0.05 -> t_{n-1;0.975}$
$\alpha = 0.01 -> t_{n-1;0.995}$
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:
   $\alpha = 0.10$ –> $z_{0.95} = 1.645$
   $\alpha = 0.05$ –> $z_{0.975} = 1.960$
   $\alpha = 0.01$ –> $z_{0.995} = 2.576$.

$(1-\alpha)100\%$ 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:
   $\alpha = 0.10$ –> $\chi^2_{n-1;0.05}$ and $\chi^2_{n-1;0.95}$
   $\alpha = 0.05$ –> $\chi^2_{n-1;0.025}$ and $\chi^2_{n-1;0.975}$
   $\alpha = 0.01$ –> $\chi^2_{n-1;0.005}$ 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 $X_{A1}, \dots, X_{An_A}$ and $X_{B1}, \dots, X_{Bn_B}$.

$(1-\alpha)100\%$ **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 $\underset{\sim}{y} = (38.14, 39.07, 37.29, 41.20, 40.31, 39.07, 34.99, 36.82, 35.23, 37.97, 36.21, 45.13, 35.98, 36.55, 37.45, 40.23, 38.45, 45.01, 36.94, 42.09)$. 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
```