Elementary hypothesis testing

- Introduction
- **Some distributions related to normal distribution**
- Types of **hypotheses**
- Types of errors
- Critical regions
- Example when variance is known
- Example when variance is not known
- **Multiple hypotheses testing**
- Power of tests

Introduction

Statistical hypotheses are in general different from scientific ones. Scientific hypotheses deal with the behavior of scientific subjects such as interactions between all particles in the universe. These hypotheses in general cannot be tested statistically. Statistical hypotheses deal with the behavior of observable random variables. These are ones that are testable by observing some set of random variables. They are usually based on the distribution(s) of observed random variables.

For example if we have observed two sets of random variables $\boldsymbol{x}=(x_1,x_2,,,,x_n)$ and $\boldsymbol{y}=(y_1,y_2,,,,y_m)$ then one natural question is: are the means of these two sets different? It is a statistically testable hypothesis. Another question may arise: do these two sets of random variables come from the population with the same variance? Or do these random variables come from the populations with the same distribution? These questions can be tested using observed samples.

Examples of statistical hypotheses

1. Differences between means: effects of two or more different treatments. When can I say that what I observe is significant?
2. Outlier detection: Does my dataset contain an error? What does it mean to have an erroneous observation?
3. Model comparison: do I have enough data to distinguish between two models? Which models can I compare? Which model should I select?

Hypothesis testing and probability distributions

Hypotheses are usually expressed using some statistics: functions of observed quantities. Since observations are considered as random variables then any function depending on observations is also random variable:

$$t = t(x_1, x_2, \ldots, x_n)$$

In particular mean, variance, median, parameters of linear models are functions of observations and therefore they are random variables. In principle if we know the distributions of observations then we can derive distributions of required statistics also. In practice it is rarely possible and we often use approximations.

For some simple cases it is possible to derive distributions of statistics. These include mean, variance, ration of variances, ratio of mean and variances – under assumptions that observations are independent, identically distributed with Gaussian distribution.

Distributions related with normal

Sample mean

Let us assume that we have drawn independently n values from the population with distribution $N(0,\sigma)$. It is our sample with independently and identically distributed (i.i.d.) random variables form the population with normal distribution.

The sum of the normal random variables has normal distribution – $N(0,\sqrt{n}\sigma)$. Then the sample mean has $N(0, \sigma/\sqrt{n})$

Small exercises:

What happens if the population distribution is $N(\mu,\sigma)$, i.e. mean is not 0. What is the distribution of the sum of the sample points? What is the distribution of the sample mean?

Let us complicate it a bit further: Let us remove the assumption that all means and variances are equal, i.e. i-th sample point comes form the population with normal distribution $N(\mu_i,\sigma_i)$. What is the distribution of sum of these random variables and what is the distribution of the sample mean?

Sample variance

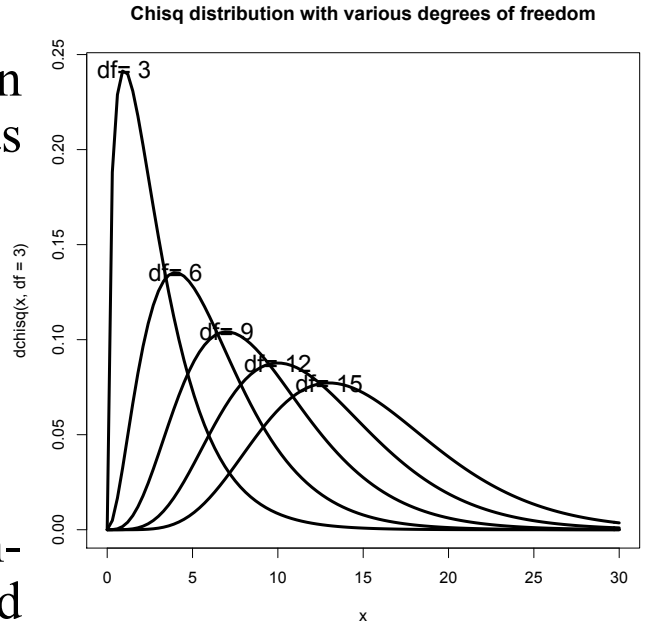
If the i.i.d. random variables have N(0,1) distribution then the sum of the squares of these random variables has χ^2 distribution with n degrees of freedom.

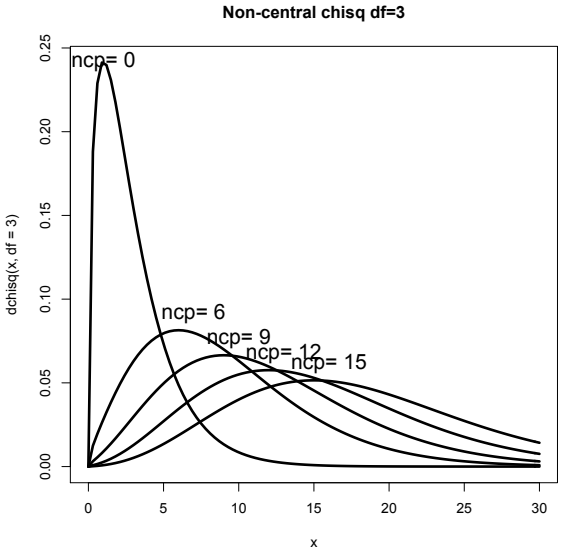
If i.i.d. random variables have N(0,σ) distribution then

$$y = \sum_{i=1}^{n} \frac{x_i^2}{\sigma^2}$$

has χ^2 distribution with n degrees of freedom.

If i.i.d. random variables have N(μ,σ) then y has non-central χ^2 distribution with n degrees of freedom and with non-centrality parameter – $n(\mu/\sigma)^2$





Sample variance

Variance of the iid random variables with N(μ,1):

$$\text{var} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is multiple of χ^2 distribution. More precise *(n-1)**var has χ^2 distribution with *n-1* degrees of freedom. Degrees of freedom is n-1 because we have one linear constraint. In general the number of degrees of freedom is reduced by the number of linear constraints.

Let us examine a very simple case. We have two random variables from N(μ,1). Then we can write:

$$(x_1 - \frac{x_1 + x_2}{2})^2 + (x_2 - \frac{x_1 + x_2}{2})^2 = \frac{(x_1 - x_2)^2}{2}$$

We see that 1*var is a square of a single random variable - x_1-x_2 that has normal distribution N(0,√2). So (x_1-x_2)/√2) has normal distribution N(0,1) and 1*var has χ^2 distribution with 1 degree of freedom.

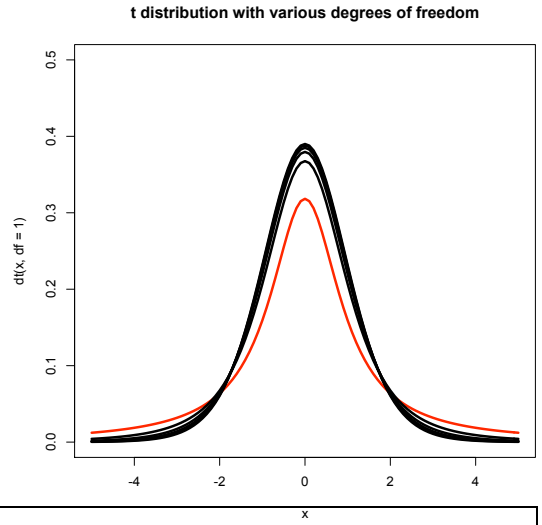
Ratio of the sample mean to the sample variance

Let us assume we have two random variables – x and y. x has $N(0,1)$ and y has χ^2 distribution with n degrees of freedom. Then the distribution of

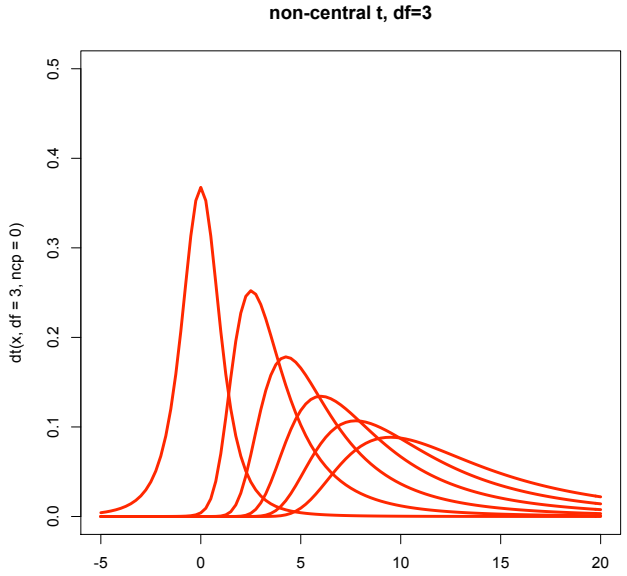
$$z = x / \sqrt{y / n}$$

is Student's t distribution with n degrees of freedom.

If the distribution of x is $N(\mu,1)$ then the distribution of ratio – z has a non-central t distribution with n degrees of freedom and with non-centrality parameter – μ.



T distribution with degrees of freedom 1,3,5,7,9,11



Non-central t distribution with 3 degrees of freedom, ncp=0,3,5,7,9,11

Ratio of mean and variance

Let us assume a sample has iid *n* random variables from *N(0,σ)*. Then the sample mean has *N(0, σ/√n)*. Therefore $\sqrt{n}\,\bar{x}/\sigma$ has N(0,1). Sample variance *(n-1) var/σ²* has χ^2 distribution with *n-1* degrees of freedom. Then:

$$z = \frac{\sqrt{n}\,\bar{x}/\sigma}{\sqrt{(n-1)\,\mathrm{var}/(\sigma^2(n-1))}} = \frac{\sqrt{n}\,\bar{x}}{\sqrt{\mathrm{var}}} = \frac{\sqrt{n}\,\bar{x}}{sd}$$

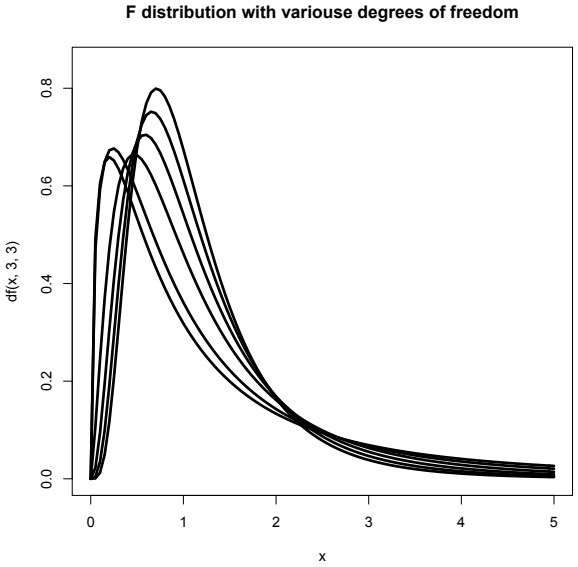
has t distribution with n-1 degrees of freedom. Attractive side of this random variable is that it does not depend on the population standard deviation – *σ*.

If the sample points are from *N(μ,σ)* then $\sqrt{n}\,\bar{x}/\sigma$ has *N(√nμ/σ,1)* and the ratio – *z* has non-central *t* distribution with non-centrality parameter *√nμ/σ*. This distribution depends on the population standard deviation. In practice it is replaced by the sample standard deviation.

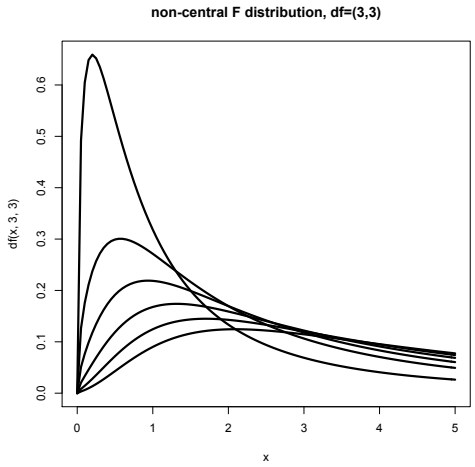
Ratio of variances of two independent samples

If we have two random variables – x and y with χ^2 distributions with degrees of freedom n and m respectively then $z = (x/n)/(y/m)$ has F distribution with (n,m) degrees of freedom.

If x has non-central χ^2 with non-centrality parameter λ then z has non-central F distribution with (n,m) degrees of freedom and non-centrality parameter λ.



F distribution with degrees of freedom (3,3),(3,5),(5,7),(7,9), (9,11),(11,13)



Non-central F, df=(3,3), non-centrality parameter: 0,3,5,7,9,11

Ratio of variances of two independent samples

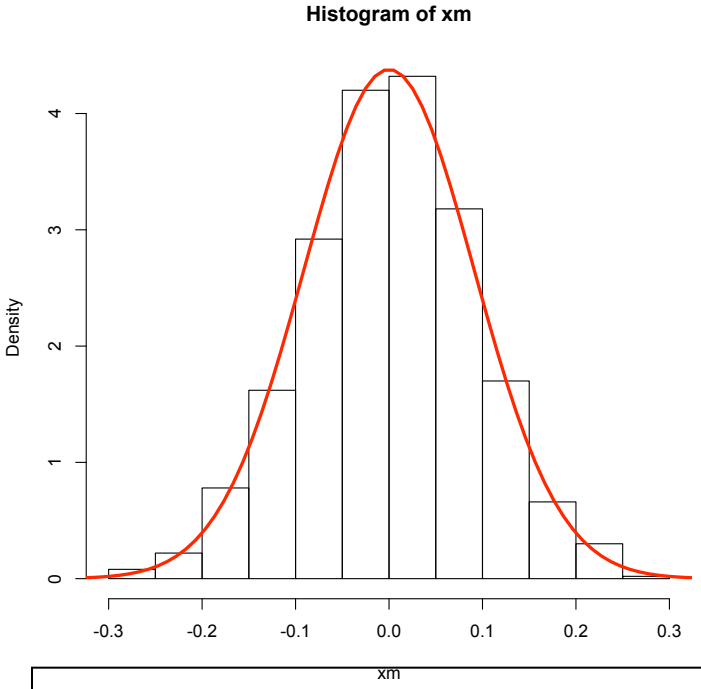
If we have two independent samples – x_1 and x_2 of sizes n_1 and n_2 from $N(\mu_1,\sigma)$ and $N(\mu_2,\sigma)$, then $(n_1\text{-}1)\text{var}_1/\sigma^2$ and $(n_2\text{-}1)\text{var}_2/\sigma^2$ have χ^2 distributions with $n_1\text{-}1$ and $n_2\text{-}1$ degrees of freedom respectively. Then the ratio:

$$z = \frac{((n_1 - 1)\,\text{var}_1/\sigma^2)/(n_1 - 1)}{((n_2 - 1)\,\text{var}_1/\sigma^2)/(n_2 - 1)} = \frac{\text{var}_1}{\text{var}_2}$$

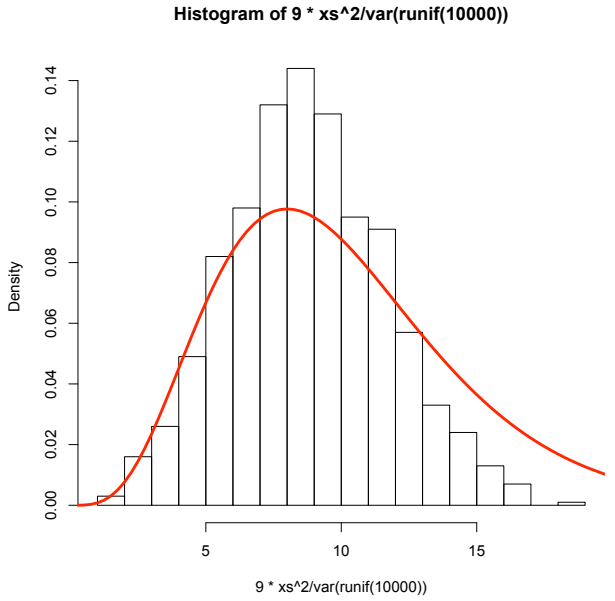
have F distribution with $(n_1\text{-}1, n_2\text{-}1)$ degrees of freedom. Again z does not depend on the unknown parameter σ.

Departure from normal distribution

Although χ^2, t, F distributions are derived from normal distribution they work well in many cases when population distributions are not normal. Let us take an example: a sample from uniform distribution in the interval (-0.5,0.5). The distribution of the sample mean is very well approximated by the normal distribution. Distribution of the sample variance is not approximated by χ^2 very well.

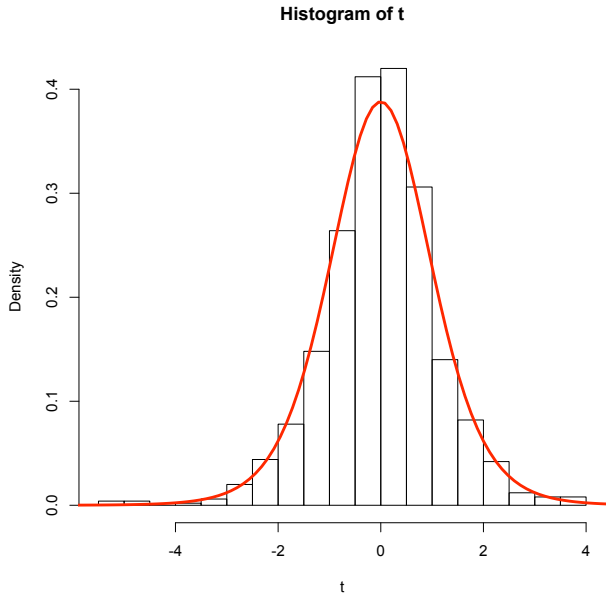


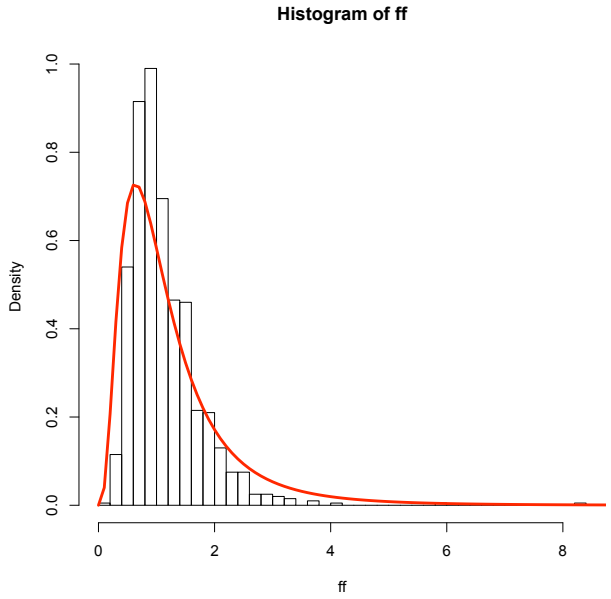
Histogram: distribution of sample mean from uniform distribution from (-0.5,0.5)
Red: normal distrobution curve with 0 mean and sd(runif(1000))/sqrt(10)



Hist: distribution of 9*var/var(runif(1000)) for sample from uniform distribution
Red curve: χ^2 distribution with 9 degrees of freedom.

Departures from normal distribution





Hist: Distribution of $\sqrt{10}$*xmean/sd
Red curve: t distribution with 9 degrees of freedom

Hist: distribution of ratio of variances of samples of sizes 10 from uniform population with distribution.
Red curve: F distribution with (9,9) degrees of freedom

χ^2 does not approximate for the sample size 10, however t distribution approximates very well. When the number of samples increases then both χ^2 and t approximates the corresponding distributions very well. Approximation by F distribution of ratio of variances distribution is good.

Elements of hypothesis testing

Types of hypotheses

Hypotheses in general can be divided into two categories: a) parametric and b) non-parametric. Parametric hypotheses concern with situations when the distribution of the population is known. Parametric hypotheses depend on the value of one or several parameters of this distribution. Non-parametric hypotheses concern with situations when none of the parameters of the distribution is specified in the statement of the hypothesis. For example hypothesis that two sets of random variables come from the same distribution is non-parametric one.

Parametric hypotheses can also be divided into two families: 1) Simple hypotheses are those when all parameters of the distribution are specified. For example hypothesis that set of random variables comes from a population with normal distribution with known variance and known mean is a simple hypothesis 2) Composite hypotheses are those when some parameters of the distribution are specified and others remain unspecified. For example hypothesis that set of random variables comes from a population with normal distribution with a given mean value but unknown variance is a composite hypothesis.

Errors in hypothesis testing

Hypothesis is usually not tested alone. It is tested against some alternative one. Hypothesis being tested is called the null-hypothesis and denoted by H_0 and alternative hypothesis is denoted H_1. Subscripts may be different and may reflect the nature of the alternative hypothesis. Null-hypothesis gets "benefit of doubt". There are two possible conclusions: reject null-hypothesis or not-reject null-hypothesis. H_0 is only rejected if the sample data contains sufficiently strong evidence that it is not true. Usually testing of a hypothesis comes to verification of some test statistic (a function of the sample points). If this value belongs to some region w hypothesis is rejected.. This region is called critical region. The region complementary to the critical region that is equal to W-w is called acceptance region. By rejecting or accepting hypothesis we can make two types of errors:

Type I error: Reject H_0 if it is true

Type II error: Accept H_0 when it is false.

Type I errors usually considered to be more serious than type II errors.

Type I errors define significance levels and Type II errors define power of the test. In ideal world we would like to minimize both of these errors.

Power of a test

The probability of Type I error is equal to the size of the critical region, α. The probability of the type II error is a function of the alternative hypothesis (say H_1). This probability usually denoted by β. Using notation of probability we can write:

$$P(x \in w \mid H_0) = \alpha$$

$$P(x \in W - w \mid H_1) = \beta \text{ or } P(x \in w \mid H_1) = 1 - \beta$$

Where x is the sample points, w is the critical region and W-w is the acceptance region. If the sample points belong to the critical region then we reject the null-hypothesis. Above equations are nothing else than Type I and Type II errors written using probabilistic language.

Complementary probability of Type II error, 1-β is called the power of the test of the null hypothesis against the alternative hypothesis. β is the probability of accepting null-hypothesis if alternative hypothesis is true and 1-β is the probability of rejecting H_0 if H_1 is true.

Power of a test is the function of α, the alternative hypothesis - H_1 and probability distributions conditional on H_0 and H_1.

Critical regions and power

The table shows schematically relation between relevant probabilities under null and alternative hypothesis.

	do not reject	reject
Null hypothesis is true	$1-\alpha$	α (Type I error)
Null hypothesis is false	β (Type II error)	$1-\beta$

Critical region

Let us assume that we want to test if one parameter of the population is equal to a given value against alternative hypothesis. Then we can write (for example):

$$H_0 : \theta = \theta_0 \quad \text{against} \quad H_1 : \theta < \theta_0$$

Test statistic is usually a point estimation for θ or somehow related to it. If critical region defined by this hypothesis is an interval $(-\infty; c_u]$ then c_u is called the critical value. It defines upper limit of the critical interval. All values of the statistic to the left of c_u lead to rejection of the null-hypothesis. If the value of the test statistic is to the right of c_u this leads to not-rejecting the hypothesis. This type of hypothesis is called left one-sided hypothesis. Problem of the hypothesis testing is: either for a given significance level find critical value - c_u or for a given sample statistic find the observed significance level (p-value).

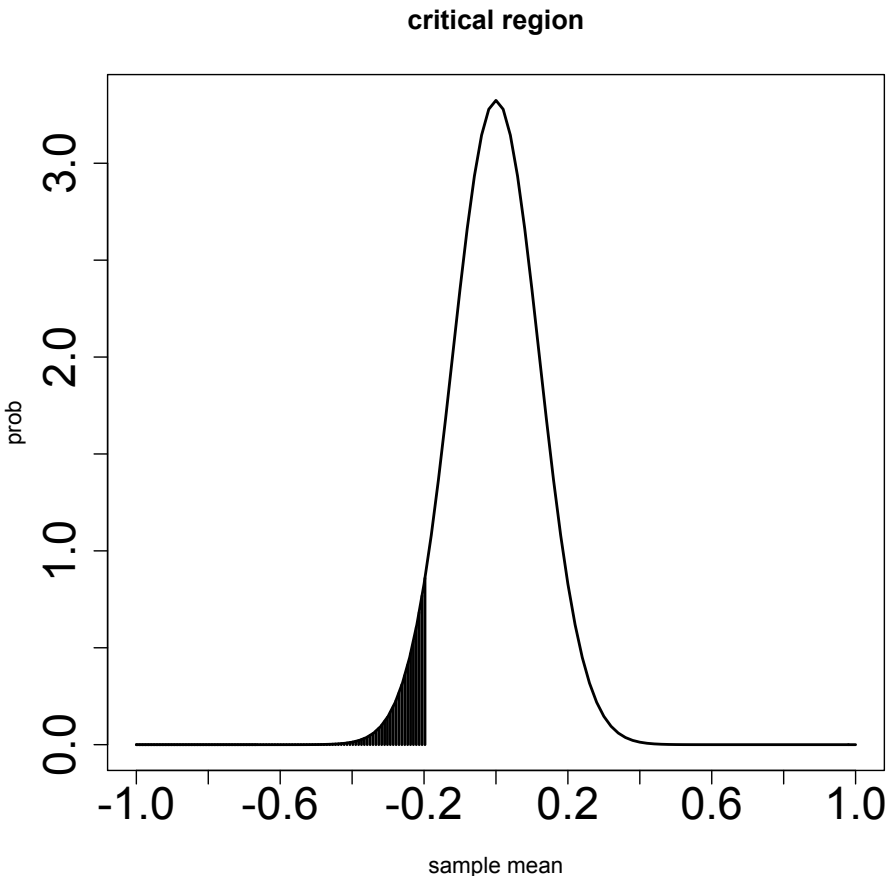
Example of tests: Normal distribution

We have a sample of size 10 from normal distribution. The sample mean is -0.3. Let us assume that the population standard deviation is 0.39. We want to test hypothesis:

H_0: $\mu=0$ against alternative hypothesis H_1: $\mu<\mu_0$

Under H_0 the sample mean has normal distribution with 0 mean and $0.39/\sqrt{10}\approx 0.12$ standard deviation. Let us set significance level at 0.05. To find critical region we need to solve $P(x<c_u)=0.05$. We do it using R command qnorm(0.05,sd=$0.39/10^{1/2}$) and it is -0.203.

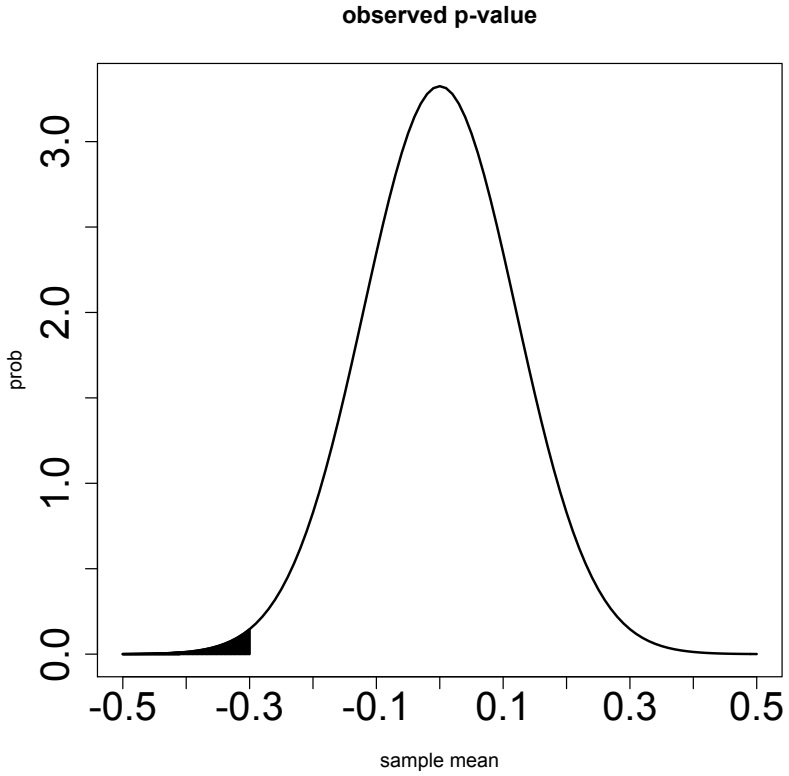
The size of the shaded are is 0.05 i.e. if H_0 is true the probability that an observed mean will belong to this area is 0.05. Critical value is -0.203. Since the observed mean, -0.3 is less than this value we reject H_0 at a significance level 0.05.



Example of tests: Normal distribution

Since the observed mean -0.3 now we can calculate the probability of observing -0.3 or less if the H_0 is true. It can be done using R command – pnorm(-0.3,sd=0.39/$10^{1/2}$). That is equal to 0.007. This value is called observed p-value and quoted by stat packages. It can be interpreted as follows: If we would have 1000 samples of sizes 10 from the normal distribution N(0,0.39) then only around 7 times the sample mean would be -0.3 or less.

Black aree is the observed critical region and the size of this region is 0.007

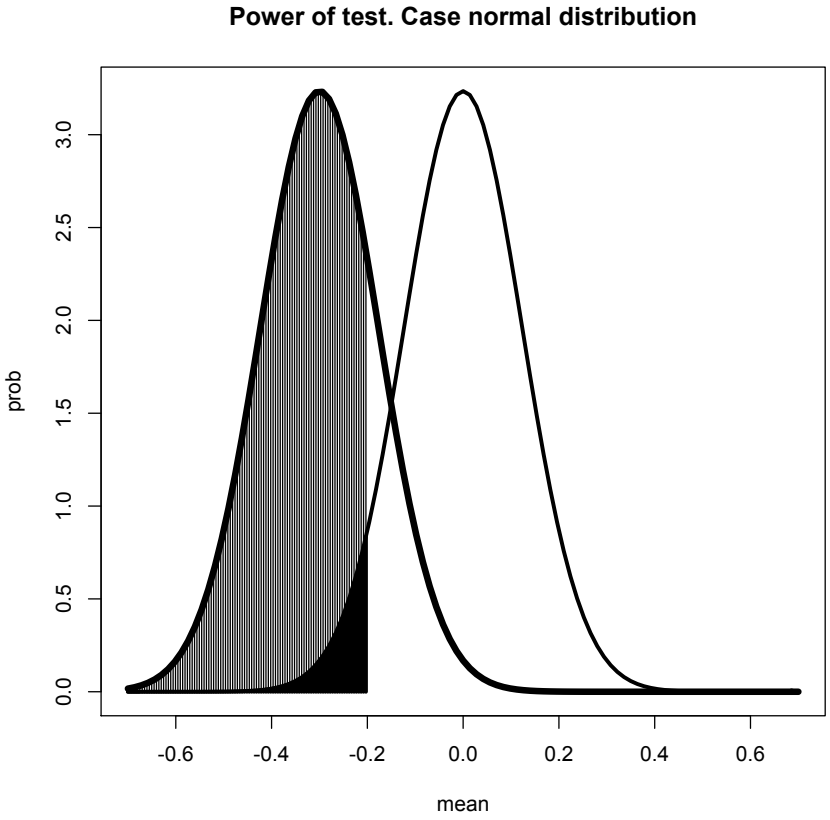


Example of tests: Normal distribution

To calculate the power of the test we need to specify an alternative hypothesis. Let us assume that alternative hypothesis is specified as H_1:μ=-0.3.

We have already critical region for a given significance level 0.05. That is equal -0.203. We want to calculate the probability of rejecting H_0 if H_1 is true. If H_1 is true then the distribution of the sample mean will be $N(-0.3,0.39/10^{1/2})$. Power of the test can be calculated using R command - *pnorm(-0.203,mean=-0.3,sd=0.39/10^{1/2})=0.78*

Thin line – distribution under null-hypothesis
Thick line – distribution under H1
Dark area – size of the critical region under H_0. Size of this area is the significance level (0.05).
Shaded area – probability of rejecting H_0 when H1 is true. The size of this area is 0.78 and therefore power of the test is 0.783



Composite hypothesis

In the above example we assumed that the population variance is known. It was a simple hypothesis (all parameters of the normal distribution have been specified). But in real life it is unusual to know the population variance. For example if the population variance is not known the hypothesis becomes composite (hypothesis defines the population mean but population variance is not known). In this case variance is calculated from the sample and it replaces the population variance. Then *t* distribution with n-1 degrees of freedom is used. When n (>50) is large then as it can be expected normal distribution very well approximates *t* distribution.

If we have two samples from the population with equal but unknown variances then tests of differences between two means comes to *t* distribution with (n_1+n_2-2) degrees of freedom. Where n_1 is the size of the first sample and n_2 is the size of the second sample. When variances of the populations are different then approximation to t distribution (Welch approximation) is used and in this case degrees of freedom could be non-integer and less than n_1+n_2-2.

If the variances for both population would be known then test statistics for differences between two means has a normal distribution.

Composite hypothesis.

If we do not know the population variance then we cannot use normal distribution since the distribution of the sample mean – $N(\mu_0,\sigma/n^{1/2})$ depends on the population variance. We need another statistic that has no unknown parameters.

We know: $u=n^{1/2}*(mean-\mu_0)/\sigma$ has $N(0,1)$, *and* $v=(n-1)var/\sigma^2$ has χ^2 distribution with *n-1* degrees of freedom then $u/(v/(n-1))^{1/2} = n^{1/2}(mean-\mu_0)/(var)^{1/2}=n^{1/2}(mean-\mu_0)/sd$ has *t* distribution with *n-1* degrees of freedom and it does not depend on unknown parameter – population variance. Now we have a statistic that is fully specified and we can use it to design tests.

To calculate power we specify the alternative hypothesis, e.g. H_1: $\mu=\mu_1$. We need the distribution of the statistic under *H1* also. This distribution becomes non-central t distribution with n-1 degrees of freedom and with non-centrality parameter $n^{1/2}(\mu_1-\mu_0)/\sigma$, where σ is unknown population standard deviation. In practice it is replaced by the sample standard deviation.

Example

Let us use the same example but this time we do not know the population variance. This time we test H_0 against two sided alternative hypothesis.

$$H0: \mu=0 \text{ against } H1: \mu \neq 0$$

Again the sample mean is -0.30 and the sample standard deviation is 0.39, the sample size is 10. If H_0 is true then $t=10^{1/2}*mean/sd$ has the t distribution with 9 degrees of freedom. Observed value is ts=-2.43. For a given significant level (α=0.05) let us calculate critical region. To find left and right critical values we need to to solve the equation:

$$P(t<t_l) + P(t>t_r) = 0.05$$

There can be many solution of this equation. Since the probability distribution is symmetric we want to make this region also symmetric (in general we would like to minimise the Euclidean measure of the acceptance region), i.e.

$$P(t<t_l) = P(t>t_r) \rightarrow 2P(t<t_l))=0.05 \rightarrow P(t<t_l)=0.025$$

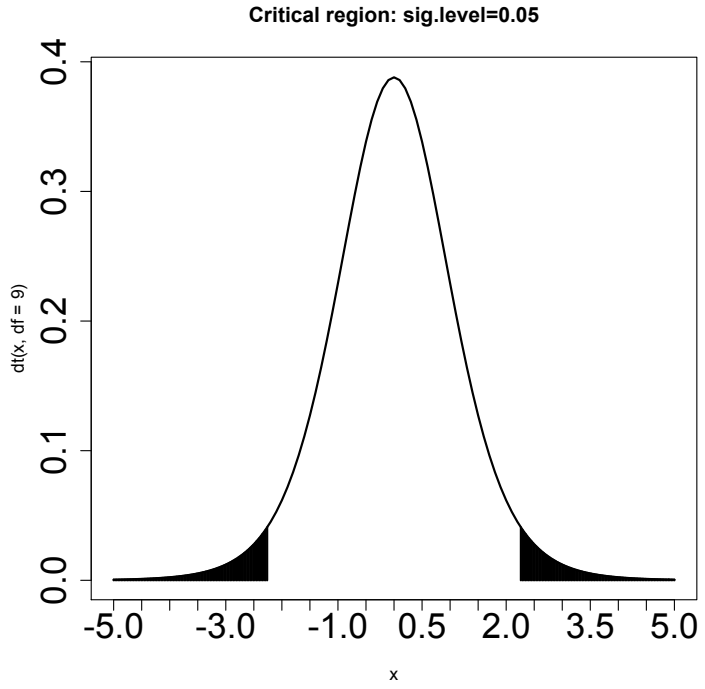
It can be solved using R command *qt(0.025,df=9)* for the left critical point and *qt(0.975,df=9)* for the right critical point. These are: -2.262 and 2.262.

Observed critical region *is* $(-\infty,-abs(ts)) \cup (abs(ts), \infty)$, Observed p.value is calculated using *pt(-abs(ts),df=9)+1-pt(abs(ts),df=9)*. In this case p.value is 0.038. We could reject null-hypothesis at sig.level 0.05 but we could not reject at sig.level 0.01. Once we have the critical values for t distribution we can calculate them for our sample. For example confidence interval is (say 95% confidence interval):

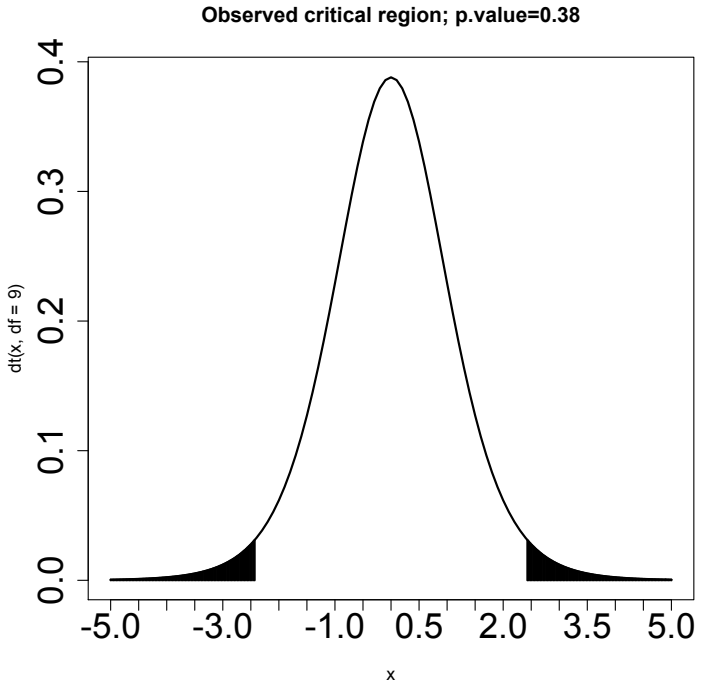
$$(\bar{x} - qt(0.025, df = 9) * sd / \sqrt{n}, \bar{x} + qt(0.975, df = 9)) = (-0.58, -0.21)$$

It does not contain 0 so we can say with 95% confidence that mean is not 0.

Example: Critical region



Critical region corresponding to 0.05 significance level

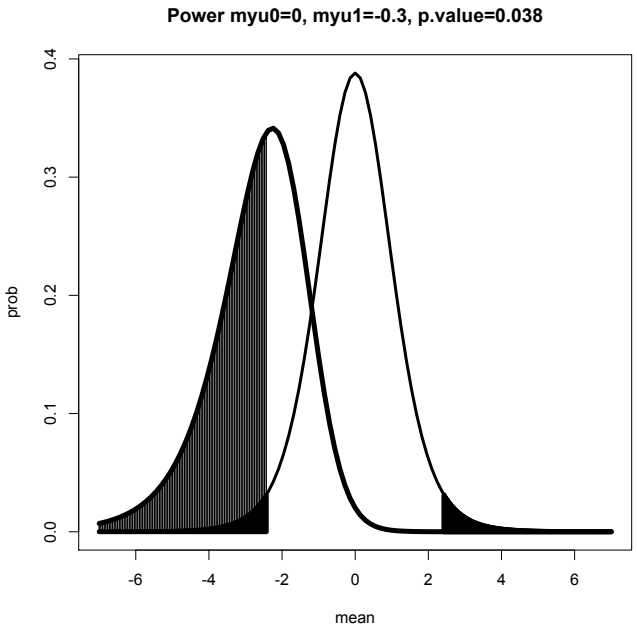


Observed critical region with the size of 0.038 (p.value)

Power of test

Power of the test depends on the significance level, null hypothesis and the alternative hypothesis. All they must be fully specified. Let us specify alternative hypothesis, for example. H_1:μ=*-0.3*. Recall that observed value of *t* is -2.43, standard deviation is 0.39, the sample size is 10. Since H_0:μ=*0* to calculate the power we need non-central t distribution with degrees of freedom 9 and non-centrality parameter $ncpp = n^{1/2}(\mu_1-\mu_0)/sd$=$10^{1/2}$(-0.3)/0.39=-2.43. We know the observed critical region (rejection area). The power of the test can be calculated using R command:

pt(-abs(ts),df=9,ncp=ncpp)+1-pt(abs(ts),df=9,ncp=ncpp)=0.525.



Thin line: the distribution under H_0
Thick line: the distribution under H_1
Black area: Critical region. Size of this region gives p.value
Dashed region: Power of the test. Size of this region gives probability of rejecting null hypothesis if the alternative is true. For current case the power is 0.525

Power of test

For a simple sample mean test power is calculated using the following formula (for two sided test)

$pw = pt(qt(\alpha/2,df=n-1),df=n-1,ncp=n^{1/2}\ \Delta/\sigma)+1-= pt(qt(1-\alpha/2,df=n-1),df=n-1,ncp=n^{1/2}\ \Delta/\sigma)$

Where α is significance level, n is the sample size, σ is population standard deviation, Δ is desired effect: i.e. difference between means under H_1 and H_0

In this equation there are four unknowns. We need to set up three of them to find the fourth.

Power of test

Power of a test can be used before as well as after experimental data have been collected. Before the experiment it is performed to find out the sample size and measurement accuracy (standard deviation) to detect a given effect. It can be used as a part of the design of an experiment.

After the experiment it uses the sample size, effect (e.g. observed difference between means), standard deviation and calculates the power of the test, i.e. probability of rejecting null hypothesis when alternative is true.

For example if we want to detect difference between means equal to 1 (delta) in paired design with power equal 0.8 at a significance level 0.05 in one sided test then we need around 8 observations.

It is done in R using the command
power.t.test(delta=1,sd=1,power=0.8,type='paired',alt='one.sided')

The result of R function:

```
Paired t test power calculation

              n = 7.7276
          delta = 1
             sd = 1
      sig.level = 0.05
          power = 0.8
    alternative = one.sided
```

Likelihood ratio test

Likelihood ratio test is one of the techniques to calculate test statistics. Let us assume that we have a sample of size n ($\boldsymbol{x}=(x_1,,,,x_n)$) and we want to estimate a parameter vector $\boldsymbol{\theta}=(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2)$. Both $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ can also be vectors. We want to test null-hypothesis against alternative one:

$$H_0 : \theta_1 = \theta_{10} \text{ against } H_1 : \theta_1 \neq \theta_{10}$$

Let us assume that likelihood function is $\boldsymbol{L(x|\ \theta)}$. Then likelihood ratio test works as follows: 1) Maximise the likelihood function under null-hypothesis (i.e. fix parameter(s) $\boldsymbol{\theta}_1$ equal to $\boldsymbol{\theta}_{10}$, find the value of likelihood at the maximum, 2) maximise the likelihood under alternative hypothesis (I.e. unconditional maximisation), find the value of the likelihood at the maximum, then find the ratio:

$$w = L(x \mid \theta_{10}, \hat{\theta}_2) / L(x \mid \hat{\hat{\theta}}_1, \hat{\hat{\theta}}_2)$$

$\hat{\theta}_1$ is the value of the paramater after constrained ($\theta_1 = \theta_{10}$) maximisation

$\hat{\hat{\theta}}_1, \hat{\hat{\theta}}_2$ are the values of the both parameters after unconstrained maximisation

w is the likelihood ratio statistic. Tests carried out using this statistic are called likelihood ratio tests. In this case it is clear that:

$$0 \leq w \leq 1$$

If the value of w is small then null-hypothesis is rejected. If $g(w)$ is the the density of the distribution for w then critical region can be calculated using:

$$\int_0^{c_\alpha} g(w)dw = \alpha$$

R commands for tests

t.test - one, two-sample and paired t-test

pairwise.t.test – pairwise test for mutiple samples with or without corrections to multiple hypothesis

var.test - test for equality of variances

power.t.test - to calculate power of t-test

Some other tests. These are nonparametric tests and are less sensitive to distribution assumptions and outliers

wilcox.test - test for differences between means (works for one, two sample and paired cases)

ks.test - Kolmogorov-Smirnov test for equality of distributions

Multiple hypotheses testing

Multiple comparison

- One comparison - use t-test or equivalent
- Few comparisons - use Bonferroni
- Many comparisons - Tukey's honest significant differences, Holm, Scheffe or others

Bonferroni correction

If there is only one comparison then we can use t-test or intervals based on t distribution. However if the number of tests increases then probability that significant effect will be observed when there is no significant effect becomes very large. It can be calculated using $1-(1-\alpha)^n$, where α is significance level and n is the number of comparisons. For example if the significance level is 0.05 and the number of comparisons (tests) is 10 then the probability that at least one significant effect will be detected by chance is $1-(1-0.05)^{10}=0.40$. Bonferroni suggested using α/n instead of α for designing simultaneous confidence intervals. It means that the intervals will be calculated using

$$\mu_i-\mu_j \pm t^{\alpha/(2n)}(\text{se of comparison})^{0.5}$$

Clearly when n becomes very large these intervals will become very conservative. Bonferroni correction is recommended when only few effects are compared. pairwise.t.test in R can do Bonferroni correction.

If Bonferoni correction is used then p values are multiplied by the number of comparisons (Note that of we are testing effects of I levels of factor then the number of comparisons is $I(I-1)/2$

Bonferroni correction: Example

Let us take the example dataset - poisons and try Bonferroni correction for each factor:

pairwise.t.test(poison,treat,"none",data=poisons)

```
      1       2
2 0.32844 -
3 3.3e-05 0.00074
```

pairwise.t.test(poison,treat,"bonferroni",data=poisons)

```
  1      2
2 0.9853 -
3 1e-04  0.0022
```

As it is seen each *p*-value is multiplied by the number of comparisons 3*2/2 = 3. If the corresponding adjusted *p*-value becomes more than one then it is truncated to one.

It says that there are significant differences between effects of poisons 1 and 3 and between 2 and 3. Difference between effects of poisons 1 and 2 is not significant.

Note: Command in R - pairwise.t.test can be used for one way anova only.

Holm correction

Another correction for multiple tests – Holm's correction is less conservative than Bonferroni correction. It is a modification of Bonferroni correction. It works in a sequential manner.

Let us say we need to make n comparisons and significant level is α. Then we calculate *p* values for all of them and sort them in ascending order and apply the following procedure:

1) set $i = 1$

2) If $p_i < \alpha/(n-i+1)$ then it is significant, otherwise it is not.

3) If a comparison number *i* is significant then increment *i* by one and if $i \leq n$ go to the step 2

The number of significant effects will be equal to *i* where the procedure stops.

When reporting *p*-values Holm correction works similar to Bonferroni but in a sequential manner. If we have m comparisons then the smallest p value is multiplied by *m*, the second smallest is multiplied by *m-1, j*-th comparison is multiplied by *m-j+1*

Holm correction: example

Let us take the example - the data set poisons and try Holm correction for each factor:

pairwise.t.test(poison,treat,"none",data=poisons)

```
      1       2
2 0.32844 -
3 3.3e-05 0.00074
```

pairwise.t.test(poison,treat,"holm",data=poisons) # this correction is the default in R

```
  1      2
2 0.3284 -
3 1e-04  0.0015
```

The smallest is multiplied by 3 the second by 2 and the largest by 1

It says that there is significant differences between effects of poisons 1 and 3 and between 2 and 3. Difference between effects of poisons 1 and 2 is not significant.

Tukey's honest significant difference

This test is used to calculate simultaneous confidence intervals for differences of all effects.

Tukey's range distribution. If we have a random sample of size N from normal distribution then the distribution of stundentised range - $(max_i(x_i)-min_i(x_i))/sd$ is called Tukey's distribution.

Let us say we want to test if $\mu_i - \mu_j$ is 0. For simultaneous $100\alpha\%$ confidence intervals we need to calculate for all pairs lower and upper limits of the interval using:

$$difference \pm q_{l,\nu}\, sd\, (1/J_i+1/J_j)^{0.5} /\sqrt{2}$$

Where q is the α-quantile of Tukey's distribution, J_i and J_j are the numbers of observations used to calculate μ_i and μ_j, *sd* is the standard deviation, l is the number of levels to be compared and ν is the degree of freedom used to calculate *sd*.

Tukey's honest significant difference

R command to perform this test is *TukeyHSD*. It takes an object derived using aov as an input and gives confidence intervals for all possible differences. For example for poison data (if you want to use this command you should use *aov* for analysis)

lm1 = aov(time~poison+treat,data=poisons)

TukeyHSD(lm1)

```
$poison
          diff         lwr         upr     p adj
2-1 -0.073125 -0.2089936  0.0627436 0.3989657  # insignifacnt
3-1 -0.341250 -0.4771186 -0.2053814 0.0000008  # significant
3-2 -0.268125 -0.4039936 -0.1322564 0.0000606  # significant


$treat
            diff          lwr          upr     p adj
B-A  0.36250000  0.18976135  0.53523865 0.0000083 #sginificant
C-A  0.07833333 -0.09440532  0.25107198 0.6221729 #insgnific
D-A  0.22000000  0.04726135  0.39273865 0.0076661 #significant
C-B -0.28416667 -0.45690532 -0.11142802 0.0004090 #significant
D-B -0.14250000 -0.31523865  0.03023865 0.1380432 #insignific
D-C  0.14166667 -0.03107198  0.31440532 0.1416151 #insignific
```

Scheffe's simultaneous confidence intervals

If we have a parameter vector β then a linear combination $c^T\beta$ is estimatable if there exists a vector a so that $E(a^Ty) = c^T\beta$.

Scheffe's theorem states that simultaneous 100(1-α)% confidence interval for all estimatable ψ is:

$$\psi \pm (q\, F_{q,n-r}(\alpha))^{1/2}\, (var(\psi)^{1/2}$$

q is the dimension of the space of all possible contrasts, r is the rank of X (design matrix), n is the number of observations. It can also be applied for regression surface confidence intervals

$$x^T\beta \pm (q\, F_{q,n-r}(\alpha))^{1/2}\, (var(x^T(X^TX)^{-1}x))^{1/2}$$

More tests

Mann-Whitney-Wilcoxon test – non-parametric test for median

Kolmogorov-Smirnoff test – test if distribution of two random variables are same

Grubbs test – tests for outliers (more in R package – outliers)

Conclusions

1) Hypotheses are designed to be rejected
2) Testing hypotheses is reduced to analysis of few statistics – functions of observations
3) Many tests are reduced to analysis of means and variances of distributions
4) When multiple hypotheses are tested then one need to add corrections

Further reading

Full exposition of hypothesis testing and other statistical tests can be found in:

Stuart, A., Ord, JK, and Arnold, S. (1991) Kendall's advanced Theory of statistics. Volume 2A. Classical Inference and the Linear models. Arnold publisher, London, Sydney, Auckland

Box, GEP, Hunter, WG, Hunter, JS (1978) Statistics for experimenters

Peter Dalgaard, (2008) Introductory statistics with R