Statistical Computing
Week 11: 13th December 2019
Hypothesis Tests 1

Hypothesis Tests 1: Contents

- General principle of hypothesis tests
- Null and alternative hypotheses
- Critical region
- Types of error
- Significance level and power
- Confidence intervals and hypothesis tests
- Two sample tests

This subject will be continued in the Lecture/Workshop on 10th January

General principle

A sample of 20 Berlin residents was obtained and the height of each person was measured.

The population mean is assumed to be 170 cm.

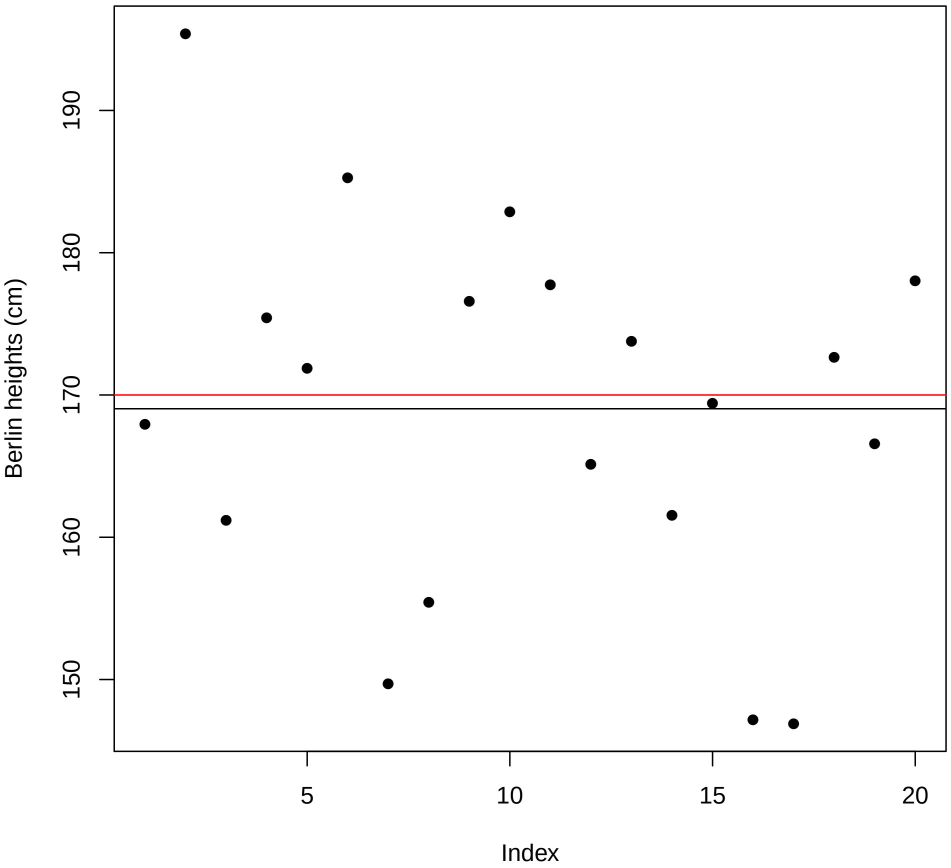
The sample has been *simulated*, i.e. created in R (fake data!)

Sample values:

```
167.9 195.4 161.2 175.4 171.9 185.3 149.7 155.4 176.6
182.9 177.7 165.1 173.8 161.6 169.4 147.2 146.9 172.7
166.6 178.0
```

Six-number summary

```
Min.   1st Qu.  Median   Mean  3rd Qu.   Max.
146.9   161.5   170.6   169.0   176.9   195.4
```



Note that the observed mean (sample mean, black line) is not equal to the population mean 170 cm (red line),...

The population mean and sample mean are two different things! They are almost always different values, but in this case they are similar.

We can believe that the population mean is 170 cm, when our sample mean is 169.0 cm.

What if the sample mean was 155 cm? Would we believe that the sample comes from a population with a mean of 170 cm? — Probably not.

Where is the cut off, when we no longer believe that the population mean is 170 cm?

To put this question into a statistical framework we set up a hypothesis test.

The broad principles of a hypothesis test are

- Set up two hypothesis, exactly one of which must be correct.
- Of the two hypotheses, one (the *null hypothesis*) has specific assumptions which we can model.
  H_0: The population mean is 170 cm
- The second hypothesis is called *alternative hypothesis*.
  H_1: The population mean is **not** 170 cm
- We compare our data with what we would expect, if the *null hypothesis* were true.
- If our data are very “unexpected” we don’t believe the null hypothesis, we reject it and “choo” the alternative hypothesis.
- If we don’t reject the null hypothesis, either *the null hypothesis is true* or *it is untrue but we have not collected enough evidence yet to prove it*.

One sample t-test

The relevant test is called the *t*-test. The type of *t*-Test in this example is a One sample *t*-test. The *t*-test is a *test of location*.

```
> t.test(Berlin,mu=170)

        One Sample t-test
data:  Berlin
t = -0.33816, df = 19, p-value = 0.739
alternative hypothesis: true mean is not equal to 170
95 percent confidence interval:
 163.0398 175.0240
sample estimates:
mean of x
 169.0319
```

The *p*-Value is much greater than 0.05.
We have no evidence to reject the null hypothesis.

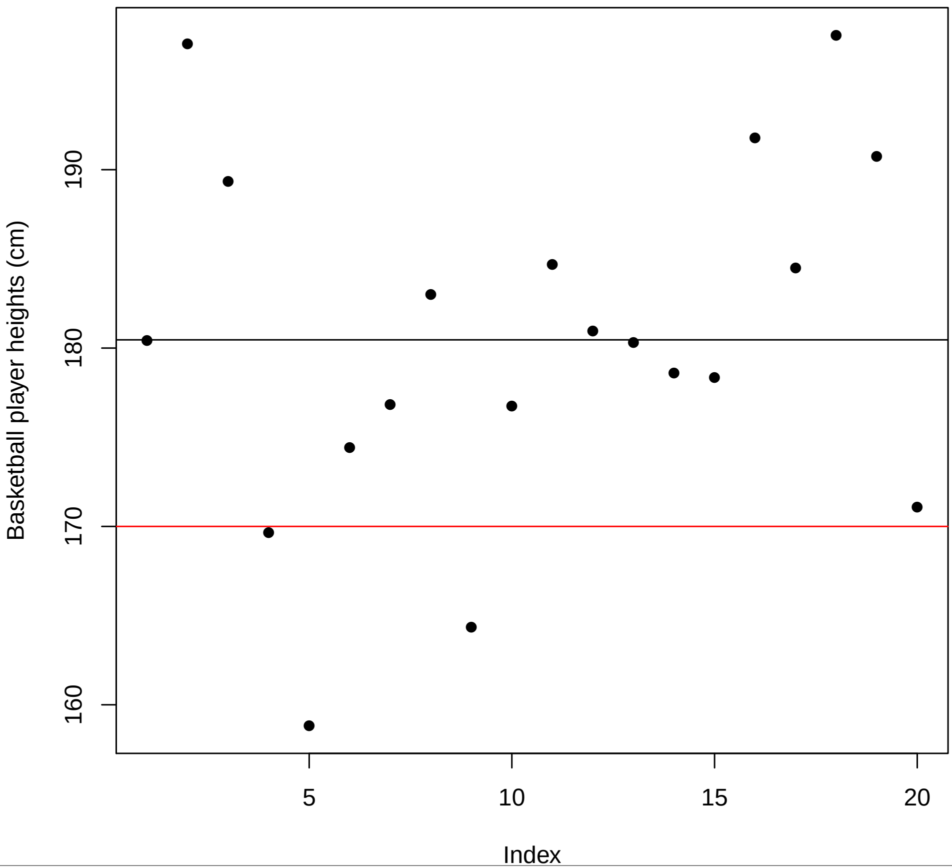
Suppose I had taken my sample exclusively from basketball players in Berlin.

The sample is not representative of *all Berlin residents*. It is representative of basketball players.

Does the basketball player population have a mean height of 170cm?

Looking at the graph on the next slide, it seems unlikely.

Heights of 20 basketball players



The *t*-test for this sample is

```
> t.test(Basketball,mu=170)

        One Sample t-test

data:  Basketball
t = 4.6489, df = 19, p-value = 0.000175
alternative hypothesis: true mean is not equal to 170
95 percent confidence interval:
 175.7514 185.1709
sample estimates:
mean of x
 180.4611
```

The *p*-Value is 0.00018, much *smaller* than 0.05.

We have good evidence to reject the null hypothesis.

The population mean height for Berliner basketball players is not 170 cm.

Let μ be the *true* expected value of the population (population mean).
μ is unknown.

Null hypothesis H_0: The population mean μ is equal to a specific value μ_0.
In our example μ_0 is 170 and is called the "null value".

Alternative hypothesis H_1: The population mean is **not** equal to μ_0

Exactly one of the two hypotheses is true.

In mathematical notation

$$H_0 : \mu = 170 = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0 (\mu \neq 170)$$

We will assume that the data comes from an iid random sample $X_1, X_2, \ldots, X_n$. The distribution of each X_i is normal with population mean $E(X) = \mu$ and population variance $\mathrm{Var}(X) = \sigma^2$.

$$X_i \sim N(\mu, \sigma^2)$$

We start off by assuming the null hypothesis is true, that $\mu = \mu_0$. If the data suggests that this is very unlikely, we will reject the null hypothesis and choose the alternative hypothesis.

Under the null hypothesis assumptions the following random variable follows a t-distribution:

Test statistic: $$t_{stat} = \frac{\overline{X} - \mu_0}{S_X}\sqrt{n} \sim t_{n-1}.$$

t_{n-1} is the t-distribution with $n-1$ degrees of freedom.
The test statistic t_{stat} is compared it to a **critical value** to see if the data are unlikely if the null hypothesis is true.

Critical region

Assuming H_0 to be true, the probability that t_{stat} lies in the central region is $1-\alpha = 95\%$. The probability that t_{stat} is extreme and lies in side regions is $\alpha = 5\%$ (the critical region).

The critical value t_{cr} is by definition the 0.975-quantile of the t_{n-1} distribution, and is found using the R command `qt(0.975,n-1)`

This leads to the decision rule:

If $|t_{stat}| \leqslant t_{cr}$, we „accept" the null hypothesis

If $|t_{stat}| > t_{cr}$, the null hypothesis is rejected, we do not believe that $\mu = \mu_0$.

On the last slide, we assumed that $\alpha = 5\%$. This is called the significance level and $\alpha = 5\%$ is the usual value by convention. Other common significance levels are 10% and 1%.

Example

From Slide 6, the R output is

```
> t.test(Berlin,mu=170)

One Sample t-test

data:  Berlin
t = -0.33816, df = 19, p-value = 0.739
alternative hypothesis: true mean is not equal to 170
95 percent confidence interval:
163.0398 175.0240
sample estimates:
mean of x
169.0319
```

The test statistic is $t_{stat} = -0.33816$.

There are 19 degrees of freedom: the critical value is $T_{cr} = 2.09$.

$|t_{stat}| = 0.33816 < T_{cr} = 2.09$. We "accept" the null hypothesis.

Instead of a critical value, statistical software outputs a *p*-value.

The exact definition of the p-value is: the smallest significance level required to reject the null hypothesis?

In the example the smallest α level to be able to reject the null hypothesis is 73%, which is way too large.

The p-value based decision rule is:
If p-value $> \alpha$, we „accept" the null hypothesis
If p-value $\leqslant \alpha$, the null hypothesis is rejected, we do not believe that $\mu = \mu_0$.

In our example the p-value is much greater than 0.05 we "accept" the null hypothesis.

"accept" the null hypothesis is written in quotes, because we must be careful not to infer too much. The correct interpretation is "we do not have enough evidence to reject the null hypothesis"
The problem is that we cannot know if H_0 really is true or that we have not collected enough information, because the sample size n is too small.

Types of Error

In a hypothesis test we cannot completely avoid the possibility that we make the wrong conclusion. If we could then we would not need to do the test! There are two possible types of error:

Type 1 error The null hypothesis is true, but we reject it.

Type 2 error The alternative hypothesis is true, but we do not reject H_0.

Another way of expressing this is

False positive: Type 1 error We have found an "effect" when there is none there.

False negative: Type 2 error We have failed to identify the genuine effect.

	H_0 accepted	H_0 rejected
H_0 is true	OK	Type 1 error ($p=\alpha$)
H_0 is true	Type 2 error ($p=\beta$)	OK

The critical region was set up so that if H_0 is true then the probability that we reject it is α. This is exactly the Type 1 error.

β is the type 2 error. If we choose a very small value of α then β becomes large and vice versa. This is one reason why α is usually set at 5% or 1% rather than 0.001%.

Significance level

Until now we have specified that the type 1 error (area in the critical region) is 5%.

This is called the **significance level** and is usually written as α.

5% is a standard value for the significance level. 1% is also very common, and is used when it is important to avoid a type 1 error.

- A Significance level of 5% or 1% is nothing more than a convention.
- A *p*-value of 0.049 should lead to a very similar interpretation as a *p*-value of 0.051.
- Unfortunately, many scientists and editors insist on a strict critical region of $p \leqslant 0.05$, before the results are considered „significant“.

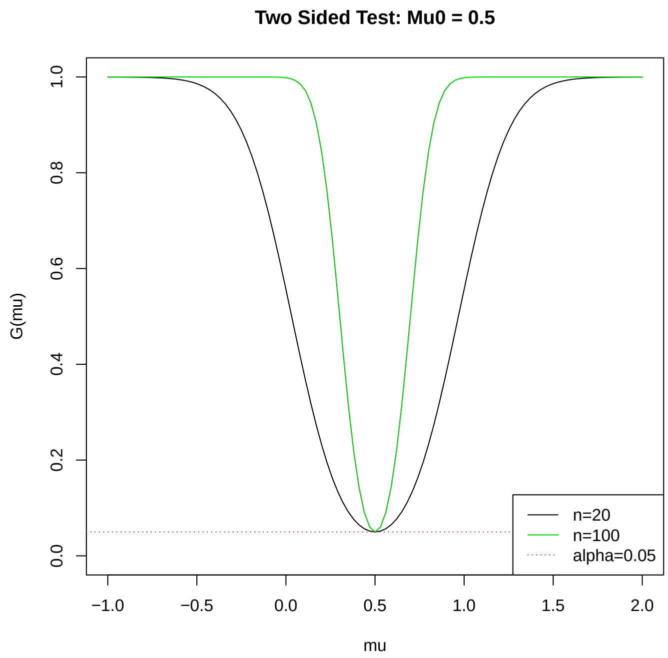
Power function

The **power** of a test is the probability that, if the alternative is true, then we will reject the null hypothesis. The power is $1 - \beta$ one minus the probaility of a type 2 error. It is a measure of the sensitivity of the test.

The value of the power of the test depends on the true value of the population mean, it is a function of μ, $G(\mu)$.

If μ is very different from μ_0 then it is easy to conclude that there is a difference and $G(\mu)$ is large.

If μ is similar to μ_0 then it is unlikely we can detect the difference and $G(\mu)$ is small (approximately α).



The power of a two sided t-test for varying μ (x-axis). The null value μ_0 is 0.5 For larger n the power of the test is greater for all values of $\mu \neq \mu_0$.

Connection between hypothesis tests and confidence intervals

Returning to the `t.test()` output:

```
95 percent confidence interval:
163.0398 175.0240
```

The confidence interval with 95% ($1-\alpha$) confidence level is directly linked to the result of the two sided test with $\alpha = 5\%$ significance level.

The test's null value is $\mu_0 = 170$. The null value is in the confidence interval $[163.0, 175.0]$. The null hypothesis is not rejected.

In general the connection is:
The $1-\alpha$ confidence interval contains the null value, if and only if the test with significance level α *does not* reject the null hypothesis.

Relaxing the assumptions

Previous assumption: the data come from an iid random sample with a normal distribution.

Because of the Central Limit Theorem we can relax this assumption as long as $n > 30$.

Provided the random sample is iid, then under the null hypothesis

$$t_{stat} = \frac{\overline{X} - \mu_0}{S_X}\sqrt{n} \overset{a}{\sim} N(0, 1).$$

The t-test can be applied to any random sample regardless of it's distribution, as long as the sample is iid.

Two sample t-test

We have assumed so far that there is one sample and we compare the sample mean to a fixed null value.

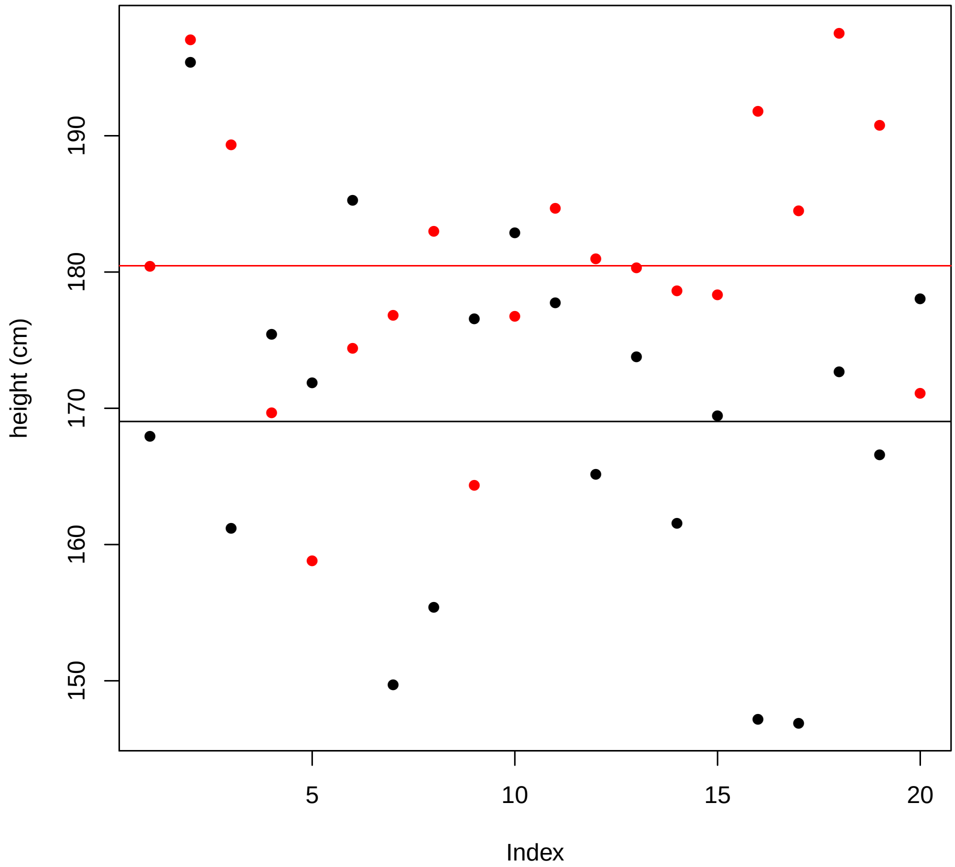
In many cases it is not so important to test the sample mean agains a value but to compare the mean between two groups. Often this will be a test group and a control group.

Back to the basketball player example: We can compare mean height of basketball players with the general Berlin population.

The null hypothesis is the two population means are the same.
The alternative hypothesis is the two population means are different.

Details and formulae for the two sample test are given in Worksheet 11, as Homework reading.

Heights of Berliners and basketball players




```
> t.test(Berlin,Basketball)

        Welch Two Sample t-test

data:  Berlin and Basketball
t = -3.1387, df = 35.991, p-value = 0.00338
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -18.814355  -4.044081
sample estimates:
mean of x mean of y
 169.0319  180.4611
```

The *p*-Value is 0.00338, again *smaller* than 0.05.
We have good evidence to reject the null hypothesis.

The population mean height for Berlin basketball players is different from population mean height for general Berliners.

Notice that the confidence interval does not contain 0. This means that the 95% confidence interval for the *difference between the two population means* is significantly different from 0.

Summary

Here is a Summary of what we have covered so far:

- A hypothesis test has a null hypothesis and an alternative hypothesis.
- We assume that the null hypothesis is true unless there is compelling evidence that this is unlikely. In which case we reject the null hypothesis.
- The significance level $\alpha(= 0.05)$ should be chosen in advance. α is the probability that we will make the wrong decision if the null hypothesis is actually true.
- The test result is decided by: comparing the test statistic with a critical value **or** by finding the p value.
- *p*-value method: if $p \leqslant \alpha$ then the null hypothesis is rejected
- A one sample *t*-Test compares the population mean with a “null value”
- A two sample *t*-Test compares the population means of two groups.