

ONE- AND TWO-SAMPLE T-TEST

APPLIED STATISTICS (STAT200)



In the last section you learned the logic of hypothesis testing, which was illustrated by means of the *z*-test.

- In the following we treat a similar test, the so-called *t*-test
- Recall that you have to know the standard deviation of the population for the *z*-test. This is not the case for the *t*-test

Therefore:

$$\sigma_x \begin{cases} \text{known} & \rightarrow & z - \text{test} \\ \text{unknown} & \rightarrow & t - \text{test} \end{cases}$$

ONE-SAMPLE T-TEST

SECTION 9.2 (DEVORE & BERK 2012)

The one-sample t -test is a parametric inferential procedure for a one-sample experiment, with unknown variance of the population.

- This context is by far more common than the case with known variance
- In most software packages, only the t -test is implemented

Example: A “home-and-gardening / housekeeping” magazine targeting women presents a test of housekeeping abilities and reports that the average score for women equals 75.

- Our question is: “How do men perform?”
- This shall be examined by letting a sample of men carry out the tests for assessing their abilities
- Afterwards, the mean of their scores \bar{x} (estimate for μ of men) is compared to the μ of the women (here 75)

In this setting, the statistical test can be set up as follows:

1. Formulate hypothesis: we do not expect a particular outcome
→ two-tailed test with
$$H_0: \mu = 75 \quad H_1: \mu \neq 75$$
2. Alpha: to be chosen, e.g. 5%
3. Check the [conditions of application / assumptions](#) of the test are satisfied.

The assumptions to be checked for 3. are:

- The sample of householding scores is random
- The scores follow the Gaussian / normal distribution
- The standard deviation of the score can be estimated by

$$S_x = \sqrt{S_x^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

How to carry out the testing procedure?

- To test the hypothesis, we follow the same logic described for the z-test
- However, now the critical values result from a t -distribution, as does our test statistic.

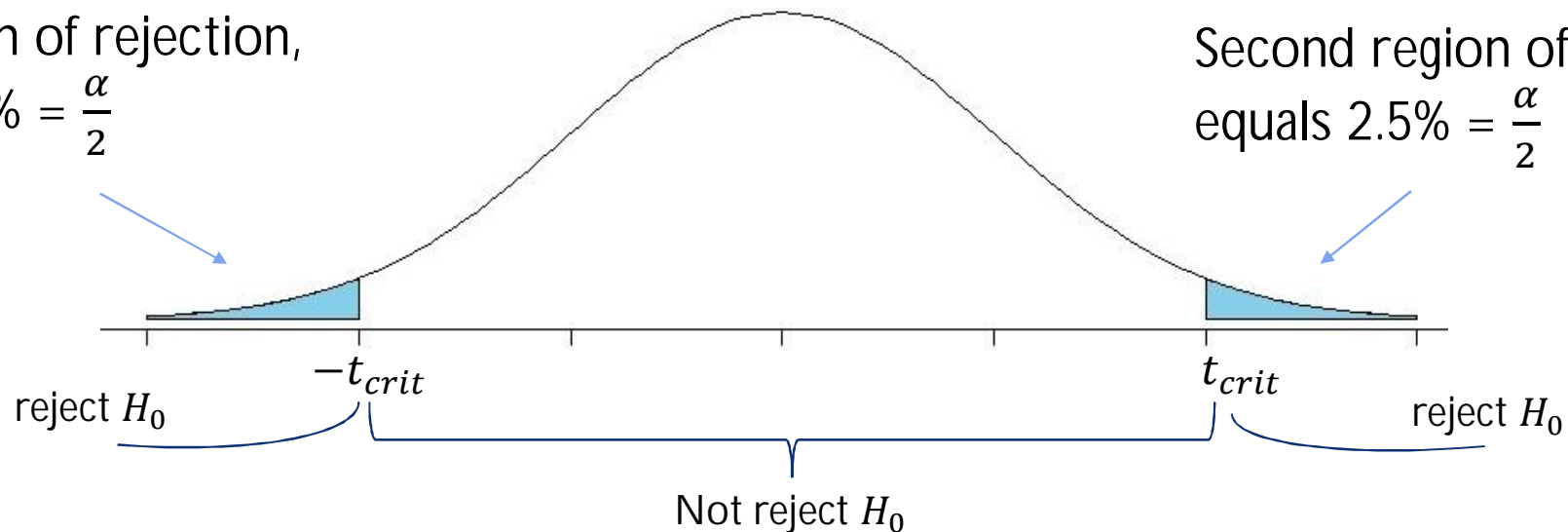
Therefore, we compare t_{obs} with t_{crit} , instead of z_{obs} with z_{crit} . The value of this new test statistic is calculated by

$$t_{obs} = \frac{\bar{x} - \mu}{S_{\bar{x}}} \quad \text{with} \quad S_{\bar{x}} = \sqrt{\frac{S_x^2}{n}}$$

- The value t_{obs} has to be compared with the critical value t_{crit} , which results from the t -distribution
- The t -distribution has a similar shape as the Gaussian distribution – they may be very close, but they are never identical (except asymptotically)

First region of rejection,
equals $2.5\% = \frac{\alpha}{2}$

Second region of rejection,
equals $2.5\% = \frac{\alpha}{2}$



Attention should be paid to the following detail:

- For the z-test, the values for z_{crit} resulted directly from a standard Gaussian distribution ($\mu=0$ and $\sigma=1$)
- There are many different versions of the t -distribution, each of which has a slightly different shape.

General rule is the following: the t -distribution to be used

- Is very close to a Gaussian distribution for large samples.
- Differs (substantially) from the Gaussian distribution for small samples.

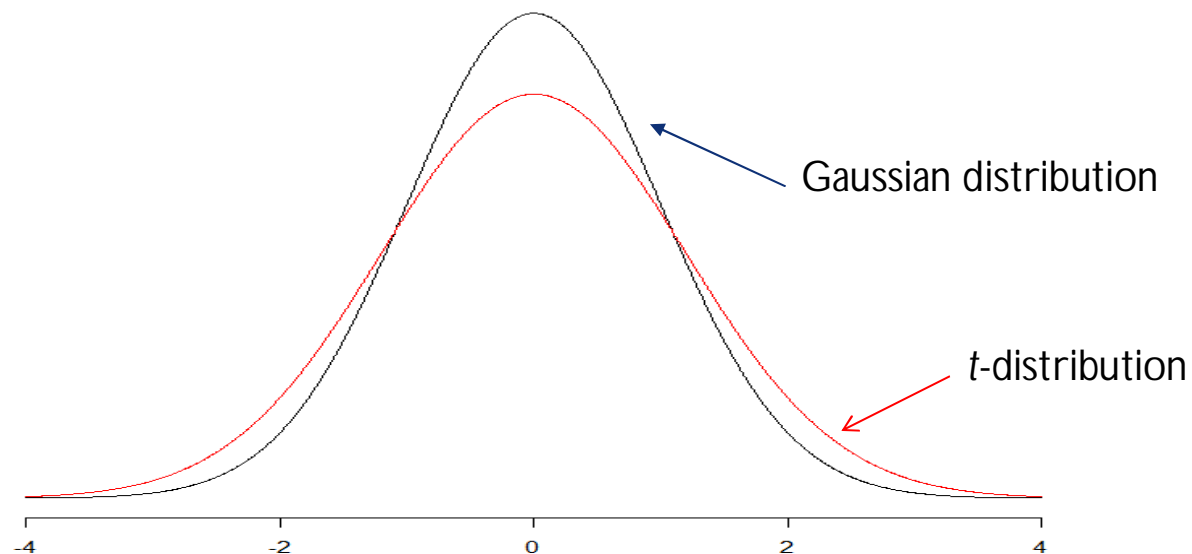
The estimator of the population variance S_x^2 may be subject to an important amount of variability for small samples – in contrast to the z-test, where the population variance is known.

Therefore the test statistic itself varies more than in the Gaussian (z-test) case.

- The shape of the t -distribution is determined by a parameter called **degrees of freedom**, (d.f. / df) which takes the value $n-1$ in this context
- The d.f. parameter is often abbreviated by ν
- Rough rule of thumb:
For $n-1$ $\begin{cases} > 120 \\ \text{in } [40, 120] \\ < 40 \end{cases}$ $\begin{cases} \text{the } t - \text{dist is almost identical to the Gaussian dist.} \\ \text{the } t - \text{dist differs from the Gaussian dist.} \\ \text{may differ strongly from the Gaussian dist.} \end{cases}$

Compared to the Gaussian distribution, the t -distribution has more probability mass in the tails, in particular for low values of ν .

Hence, for given values of α and identical sample sizes, the values of $\pm t_{crit}$ are further away from zero than the corresponding values of $\pm z_{crit}$.



Note:

1. The appropriate t_{crit} for the one-sample t -test comes from the t -distribution with $n-1$ **degrees of freedom** (d.f. / df), where n is the number of observations in the sample.
2. The test statistic of the t -test may also be re-arranged for obtaining a **confidence interval** of the (unknown) mean μ :

$$\bar{x} - t_{crit} \cdot s_{\bar{x}} \leq \mu \leq \bar{x} + t_{crit} \cdot s_{\bar{x}}$$

Confidence interval for μ :

$$\bar{x} - t_{crit} \cdot s_{\bar{x}} \leq \mu \leq \bar{x} + t_{crit} \cdot s_{\bar{x}}$$

The size/range of this interval is determined by

- the estimated standard error of the mean ($s_{\bar{x}}$) and
- t_{crit} , where t_{crit} itself depends on the value of the level (criterion) α .

Then, one can say the following:

For a given value of α , we are $1-\alpha$ % **confident** that our interval contains the unknown parameter μ (estimated by \bar{x}).

The t -test is available directly in R and easy to use

- Use the function `t.test`
- Confidence intervals, test statistic, and p-values computed
- Only one argument necessary: sample values
- Two-sided test with hypothesis of zero mean by default
- Type `?t.test` for help

If you are not sure how a test works, try a simple **example**:

```
x <- rnorm(10000, mean = 5)
```

```
t.test(x)
```

```
t.test(x, mu = 5)
```

ONE-SAMPLE T-TEST: IMPLEMENTATION IN R

15

Statistic Degrees of freedom p-value

One Sample t-test

```
data: x
t = -0.64573, df = 9999, p-value = 0.5185
alternative hypothesis: true mean is not equal to 5
95 percent confidence interval:
 4.973619 5.013307
sample estimates:
mean of x
 4.993463
```

Confidence interval

Sample mean

The diagram illustrates the output of a one-sample t-test in R. It shows the test statistic (t = -0.64573), degrees of freedom (df = 9999), and p-value (p-value = 0.5185). The alternative hypothesis is stated as 'true mean is not equal to 5'. The 95 percent confidence interval is shown as 4.973619 to 5.013307. The sample estimates are shown as the mean of x, which is 4.993463. Arrows point from the labels 'Statistic', 'Degrees of freedom', 'p-value', 'Confidence interval', and 'Sample mean' to their respective values in the output.

TWO-SAMPLE T-TEST

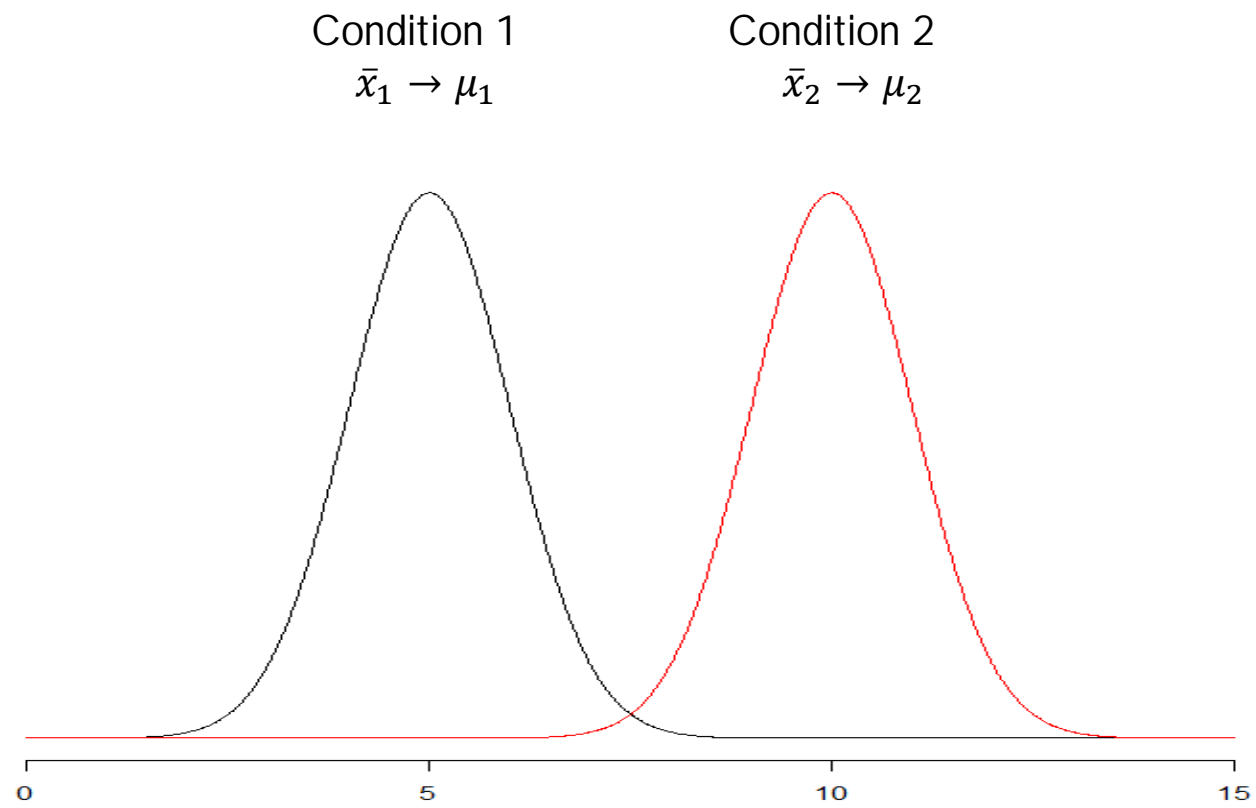
SECTION 10.2, AND 10.3 (DEVORE & BERK 2012)

The **two-sample t-test** is a straight forward extension of the one-sample t-test. It is used for comparing the means of two (non-overlapping) samples.

This situation occurs relatively often in practice, because the true mean used in a one-sample setting (e.g. IQ = 100 of the population without taking the pill) is rarely known. However, more common is the situation of measuring/observing **two samples that are subject to two different conditions**, e.g.:

- Male / female
- Takes medication / does not take medication
- Plays sport / does not play sport

and a common **quantitative target variable** (heart rate, blood pressure,..) is observed.



There are two ways to carry out a two-sample t-test. They depend on the experimental setup:

- Independent samples t-test
- Related samples t-test

Independent-samples t-test: Here, the two samples are independent, i.e., we randomly select participants for a sample, without regard to who else has been selected for either sample. Formally, the conditions for applying this test are:

1. The two samples are independent
2. The distribution of the population underlying both samples is (approximately) Gaussian
3. The two populations represented by our samples have homogeneous variance, i.e., the variances of the populations represented are equal
4. The number of observations in each sample should not be massively unequal

Remarks:

- You will know if you meet (most) of the assumptions by specific tests or previously published research
- Variance homogeneity is not absolutely necessary - but we won't treat the non-homogeneous case here
- The default command carrying out a t -test in R uses a modified procedure assuming unequal variances by default. This is the so-called [Welch test](#). Other statistical software first test for variance equality, and then carry out either the t -test or Welch's test

First, the **hypothesis** of the test has to be formulated.

If we do not predict a particular direction (increase/decrease) of the scores, we have a two tailed test, and can formulate

$$H_0: \quad \mu_1 = \mu_2 \quad \Leftrightarrow \quad \mu_1 - \mu_2 = 0$$

and

$$H_1: \quad \mu_1 \neq \mu_2 \quad \Leftrightarrow \quad \mu_1 - \mu_2 \neq 0$$

- Note that our hypotheses do not contain a specific value of μ . Therefore, testing H_0 is equivalent to testing for a zero difference between our two samples
- Naturally, extensions exist for testing, e.g., for a difference of at least 2,3,4,... (or any other value).

The **test statistic** in the independent sample setting is calculated in several steps.

1. Estimate $s_{x_1}^2$ and $s_{x_2}^2$ by the usual estimator, i.e., $S_x^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
2. Calculate the **pooled variance** of the two samples as “weighted average” of the two sample variances:

$$s_{pool}^2 = \frac{(n_1 - 1)s_{x_1}^2 + (n_2 - 1)s_{x_2}^2}{(n_1 - 1) + (n_2 - 1)},$$

where n_1 and n_2 are the sample sizes of X_1 and X_2 , respectively

2. (cont.) The resulting s_{pool}^2 is an estimator of the variance homogeneity. The estimator s_{pool}^2 serves for calculating the standard error of the mean difference:

$$s_{\bar{x}_1 - \bar{x}_2} = s_{pool}^2 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

3. The value for t_{obs} follows by

$$t_{obs} = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_{\bar{x}_1 - \bar{x}_2}},$$

which has then to be compared to the critical values of $\pm t_{crit}$.

These critical values of t for the independent samples t -test results from a t -distribution with $(n_1 - 1) + (n_2 - 1)$ degrees of freedom

As in the case of an one sample t-test, a confidence interval for the difference between two means $\mu_1 - \mu_2$ can be derived as well:

$$(\bar{x}_1 - \bar{x}_2) - t_{crit} \cdot s_{\bar{x}_1 - \bar{x}_2} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + t_{crit} \cdot s_{\bar{x}_1 - \bar{x}_2}$$

Recall how to interpret this confidence interval!

- As before, use the function `t.test`
- Arguments: sample values for both samples, null hypothesis if desired

Example:

```
y <- rnorm(10000, mean = 10)
```

```
t.test(x, y)
```

```
t.test(x, y, alternative = "greater")
```

The related-samples (or paired) t-test is applied when two measurements (observations) are recorded for the same subjects in two different conditions. Therefore, the samples representing each condition are related through the paired observations. Typical situations are

- Patients' heartbeat / IQ /... before and after treatment
- Quantity of a product bought before and after exposure to a commercial
- Any other score recorded on identical individuals in two conditions

The logic of a paired t-test is simple: perform a one-sample t-test on the differences of the scores calculated for each individual.

Example: Neck pain before (X_1) and after (X_2) therapy, measured by a score from 0 (no pain) to 100 (extremely painful).

Patient	Before	After	D
Anne	72	30	42
Sam	88	52	36
John	47	5	42
...

Then, the testing procedure follows the common steps:

1. The **experimental** hypothesis that therapy has an effect on the pain score leads to the formulation of the **statistical hypothesis** and alternative:

$$H_0: \mu_D = 0, \quad H_1: \mu_D \neq 0$$

2. The well-known quantities for performing the one-sample t-test have to be calculated:

$$S_D^2 = \frac{1}{n-1} \sum_{i=1}^n (D_i - \bar{D})^2, \quad S_{\bar{D}} = \sqrt{\frac{s_D^2}{n}}, \quad t_{obs} = \frac{\bar{D} - \mu_D}{S_{\bar{D}}},$$

where n denotes the number of paired observations.

3. Then, the critical values $\pm t_{crit}$ can be calculated from a t -distribution with $n-1$ degrees of freedom.

Note: a paired t -test is intrinsically more powerful than an independent-samples t -test, because the variability between subjects is removed by taking the difference of scores (e.g., John and Anna, having identical values of D , but very different pair scores). This aspect should be remembered when designing a study.

- As before, use the function `t.test`
- Arguments: - sample values for both samples
- `paired = TRUE`
- null hypothesis if desired

Example:

```
y <- x + 4 + rnorm(10000, sd = 0.1)
```

```
t.test(x, y, paired = TRUE)
```

```
t.test(x, y, paired = TRUE, alternative = "less")
```

Concluding remarks for this section:

- Recall that the two-sample t -test requires **variance equality** in the case of independent samples. If you wish, you may investigate this further by testing for variance (in)equality.
A suitable approach for testing for variance equality is, e.g., an F -test having test statistic $F = \frac{S_{x_1}^2}{S_{x_2}^2}$ with $n-1$ and $m-1$ d.f.
- All t -tests presented require **normality of the sample(s)**. This is something that should be tested for.

We will see more details on F -tests in the ANOVA part of the lecture.

Several tests are also available for testing for **normality**. A very powerful test is the test of **Shapiro-Wilk**:

- Principle idea of the test is to evaluate the ratio of two variance estimators for obtaining the test statistic

$$W = \frac{b^2}{(n-1)s^2},$$

where b^2 determines the “expected” variance if the sample was normally distributed, and s^2 corresponds to the effective variance of the sample

- W takes values between 0 and 1. For W close to 1, the two estimators provide similar results, and normality is plausible. For smaller values of W (attention, often still quite close to 1), the normality hypothesis can be rejected

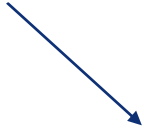
- Critical values W_{crit} are generally determined by Monte Carlo methods
- The Shapiro-Wilk test works for sample sizes up to 5000 (currently). For greater sample sizes, a good alternative is the [Anderson–Darling test](#)

Remark: The [null hypothesis](#) of the Shapiro-Wilk (and many other tests for normality) is that the [sample follows a Gaussian distribution](#). Hence, one assumes normality if the null hypothesis is not rejected.

This [is not in line with the usual principle](#) that the experimental hypothesis (normality) constitutes the alternative hypothesis of our test.

However, it is basically the only way here: under the null hypothesis of normality, a test statistic can be derived. This is not possible in a convenient way if the null hypothesis was “the sample is not normal”.

"require" does basically the same as "load", it loads an R-package.
This package needs to be installed before.



```
require(nortest)
```

```
x <- rnorm(1000)
```

```
shapiro.test(x)
```



Carry out the Shapiro-Wilk test

```
ad.test(x)
```



Carry out the Anderson-Darling test