# Applied Statistics

## for Computer Science BSc, Exam

# Probability Theory and Mathematical Statistics

## for Computer Science Engineering BSc, Term grade

**István Fazekas**

**University of Debrecen**

**2020/21 fall**

This work was supported by the construction EFOP-3.4.3-16-2016-00021. The project was supported by the European Union, co-financed by the European Social Fund.

# Main topics

1. Probability theory
2. Statistics

Mathematical tools: combinatorics, calculus

Computer tool: Matlab

Book:

Yates, Goodman:

Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers

# Lecture 11

# Testing statistical hypotheses

# The z-test

**Example.** We want to check if the weight of a bar of chocolate is equal to 100 grams. We weight $n = 16$ bars. The results are

$$x_1 = 100.3,\ x_2 = 99.8, \dots, x_{16} = 99.9$$

So it is our sample realization.
**Notation.**
Let $m$ denote the theoretical value of the weight. It is the expectation of the generic random variable $X$, i.e. $m = \mathbb{E}X$.
$X$ is the generic random variable of the underlying population.

$$X_1, X_2, \dots, X_n$$

is the sample from this population.

## The z-test

We denote the prescribed value of $m$ by $m_0$. In our example $m_0 = 100$.

$$H_0 \quad : \quad m = m_0$$

is called the null hypothesis.
The hypothesis

$$H_1 \quad : \quad m \neq m_0$$

is called the alternative hypothesis. It is a two-sided alternative hypothesis.
We should decide if $H_0$ or $H_1$ is true.
In the case of $z$-test it is assumed that the sample comes from a normally distributed population with known variance. That is $X \sim \mathcal{N}(m, \sigma^2)$, where $m$ is unknown, but $\sigma$ is known.

# The z-test

Calculate the empirical mean. We know, that

$$\bar{X} \sim \mathcal{N}(m, \sigma^2/n).$$

So we standardize it, then we obtain a standard normal random variable

$$\frac{\bar{X} - m}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1).$$

So the test statistic in $z$-test is

$$z = \frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}}$$

We emphasize that here we use $m_0$.
So $z$ is standard normal if and only if $m = m_0$, i.e. the the null hypothesis $H_0$ is true.

# The z-test

A realistic decision is the following. We reject $H_0$ if the value of $|z|$ is too large, larger than a critical value.
How to find the critical value?
First assume that a small number is preliminary given: $\alpha = 0.1, 0.05, 0.01, ....$ This $\alpha$ is the error of first kind.
It means that we reject $H_0$ with probability $\alpha$ when $H_0$ is true. And we accept $H_0$ with probability $1-\alpha$. So we have

$$\begin{aligned} 1-\alpha \quad &= P\left(-z_{\alpha/2} < \tfrac{\overline{X}-m_0}{\sigma}\sqrt{n} < z_{\alpha/2} \,|H_0\right) \\ &= \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 2\Phi(z_{\alpha/2}) - 1, \end{aligned}$$

where $\Phi(x)$ is the standard normal CDF . That is $z_{\alpha/2}$ is the number for which

$$1-\frac{\alpha}{2} = \Phi(z_{\alpha/2}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_{\alpha/2}} e^{-\frac{x^2}{2}} dx.$$

# The z-test

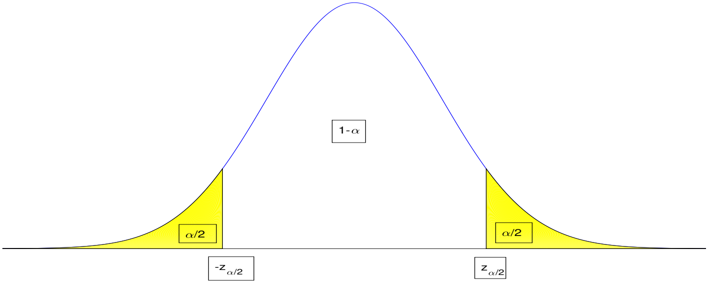


Figure: How to find the critical value for two-sided $z$-test? We cut both tails of the standard normal PDF

# The z-test

**Example.** In the previous example we had $m_0 = 100$ and $n = 16$.
Assume that $\sigma = 0.1$.
We calculated $\bar{x} = 99.9$.
Therefore

$$z = \frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}} = \frac{99.9 - 100}{\frac{0.1}{\sqrt{16}}} = -4.$$

Choose the significance level as 99% that is $\alpha = 0.01$
As $\Phi(2.58) = 0.995$, so $z_{0.005} = 2.58$.
As $|-4| > 2.58$, we reject the null-hypothesis.

# The z-test

The region of acceptance is

$$C_0 = \left\{ (x_1, \dots, x_n) : \left| \frac{\overline{X} - m_0}{\sigma} \sqrt{n} \right| < z_{\alpha/2} \right\},$$

If the sample falls into $C_0$, then we accept $H_0$.
The critical region is

$$C_1 = \left\{ (x_1, \dots, x_n) : \left| \frac{\overline{X} - m_0}{\sigma} \sqrt{n} \right| \geq z_{\alpha/2} \right\}.$$

If the sample falls into $C_1$, then we reject $H_0$.

## The one sided z-test

**Example.**

Assume that a student is on slimming diet, so he/she rejects a bar of chocolate if it is too large. Find the appropriate version of the z-test!

Now

$$H_0 \quad : \quad m = m_0$$

is the null hypothesis.

$$H_1 \quad : \quad m > m_0$$

is the alternative hypothesis. It is a one-sided alternative hypothesis.

We should decide if $H_0$ or $H_1$ is true.

Again, it is assumed that the sample comes from a normally distributed population with known variance. That is $X \sim \mathcal{N}(m, \sigma^2)$, where $m$ is unknown, but $\sigma$ is known.

# The one sided z-test

We reject $H_0$ if the test statistics

$$z = \frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}}$$

is too large, i.e. $z > z_\alpha$.

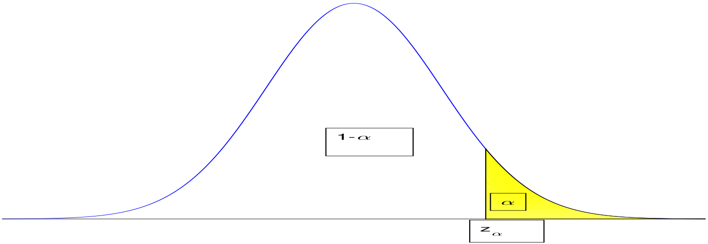


Figure: We cut the right tail of the standard normal PDF to find the critical value for this one-sided $z$-test

# The one sided z-test

**Example.** In the previous example let $m_0 = 100$ and $n = 25$.
Assume that $\sigma = 0.4$.
We calculated $\bar{x} = 100.1$.
Therefore

$$z = \frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}} = \frac{100.1 - 100}{\frac{0.4}{\sqrt{25}}} = 1.25.$$

Choose the significance level as 99% that is $\alpha = 0.01$
As $\Phi(2.32) = 0.99$, so $z_{0.01} = 2.32$.
As $1.25 < 2.32$, we accept the null-hypothesis.

## The one sided z-test (the other side)

**Example.**

Assume that a student likes chocolate very much, so he/she rejects a bar of chocolate if it is too small. Find the appropriate version of the z-test!

Now

$$H_0 \quad : \quad m = m_0$$

is the null hypothesis.

$$H_1 \quad : \quad m < m_0$$

is the alternative hypothesis. It is a one-sided alternative hypothesis.

We should decide if $H_0$ or $H_1$ is true.

Again, it is assumed that the sample comes from a normally distributed population with known variance. That is $X \sim \mathcal{N}(m, \sigma^2)$, where $m$ is unknown, but $\sigma$ is known.

## The one sided z-test (the other side)

We reject $H_0$ if the test statistics $z = \frac{\bar{X}-m_0}{\frac{\sigma}{\sqrt{n}}}$ is too small, i.e. $z < -z_\alpha$.

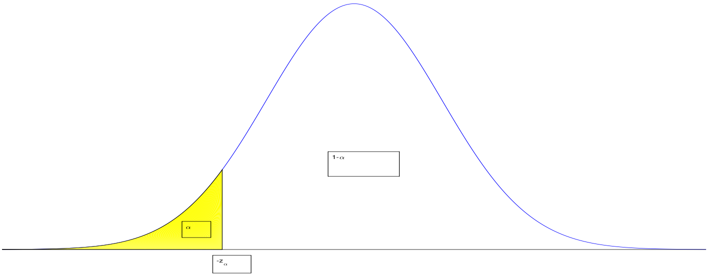


Figure: How to find the critical value for one-sided $z$-test? We should cut the left tail of the standard normal PDF

# The one sided z-test (the other side)

**Example.** In the previous example let $m_0 = 100$ and $n = 36$.
Assume that $\sigma = 0.8$.
We calculated $\bar{x} = 99.7$.
Therefore

$$z = \frac{\bar{X} - m_0}{\frac{\sigma}{\sqrt{n}}} = \frac{99.7 - 100}{\frac{0.8}{\sqrt{36}}} = -2.25.$$

Choose the significance level as 99% that is $\alpha = 0.01$
As $\Phi(2.32) = 0.99$, so $z_{0.01} = 2.32$.
As $-2.32 < -2.25$, we accept the null-hypothesis.

## The two-sample z-test

We compare the expectations of two samples.
Let $X \sim \mathcal{N}(m_1, \sigma_1^2)$, and $Y \sim \mathcal{N}(m_2, \sigma_2^2)$, where $\sigma_1$ and $\sigma_2$ are known. Let

$$X_1, X_2, \ldots, X_{n_1}; \text{ and } \; Y_1, Y_2, \ldots, Y_{n_2}$$

be independent samples from populations $\mathcal{N}(m_1, \sigma_1^2)$, resp. $\mathcal{N}(m_2, \sigma_2^2)$. Let

$$H_0 : \quad m_1 = m_2$$

be the null hypothesis. Then the test statistic

$$z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has standard normal distribution if and only if $H_0$ is true.
The possible alternative hypotheses are $H_1 : m_1 \neq m_2$ or $H_1 : m_1 > m_2$ or $H_1 : m_1 < m_2$.
The decision is the same as in the one-sample case.

## Type 1 and type 2 errors

| | We accept $H_0$ | We reject $H_0$ |
|---|---|---|
| $H_0$ is true | Good decision | Bad decision<br>Type 1 error, $\alpha$ |
| $H_0$ is false | Bad decision<br>Type 2 error | Good decision |

It is not possible to minimize both type 1 error and type 2 error at the same time.

## The $p$-value

1. The significance level $\alpha$ for a study is chosen before data collection, and it is typically set to $\alpha = 0.1$ (i.e. 10%) or $\alpha = 0.05$ (i.e. 5%) or much lower, depending on the field of study. Sometimes $1 - \alpha$ is given, that is significance level 90% means that $\alpha = 0.1$. The preliminarily given $\alpha$ means, that we want to decide as follows. We reject the null hypotheses with probability $\alpha$, given that the null hypothesis is true.
2. Statistical computer programs often do not require the preliminarily given $\alpha$, but they offer $p$-value. They calculate the value of the test statistic, $z = 1.85$, say. The $p$-value is the probability of obtaining a result at least as extreme as $z = 1.85$, given that the null hypothesis is true.
3. Of course, the above two points of view are comparable. We should reject $H_0$, when $p \leq \alpha$.

# The $p$-value

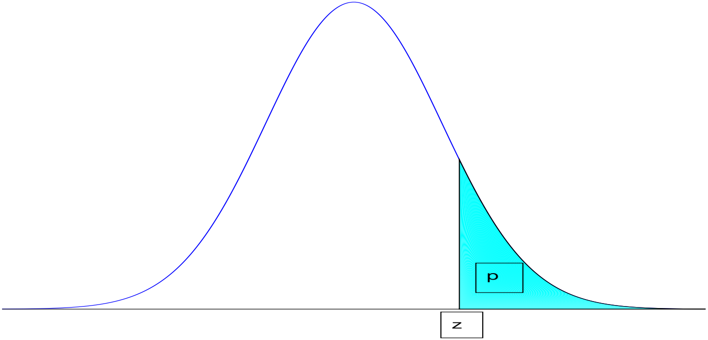


Figure: For one sided $z$-test we obtained $z = 1.2$. Then the $p$-value is the area on the right tail of the standard normal PDF

## The chi-square distribution

Let $X_1, \dots, X_n$ be independent with standard normal distribution. Then the distribution of

$$Y_n = X_1^2 + \cdots + X_n^2$$

is called chi-square distribution with degree of freedom $n$.
Notation: $Y_n \sim \chi_n^2$.
If $X_1 \sim \mathcal{N}(0, 1)$, then $\mathbb{E} X_1^2 = 1$, and $\mathrm{Var} X_1^2 = 2$, so

$$\mathbb{E} Y_n = n, \quad \mathrm{Var}\, Y_n = 2n.$$

If $Y_m$ and $Y_n$ are independent with $Y_m \sim \chi_m^2$ and $Y_n \sim \chi_n^2$, then $Y_m + Y_n \sim \chi_{m+n}^2$.

## The PDF of the chi-square distribution

The PDF of the $\chi^2_n$ distribution is

$$f(x) = \begin{cases} \dfrac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}} 2^{-\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}, & x > 0\,, \\ 0\,, & x \leq 0\,. \end{cases}$$

Here

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du$$

is the Gamma function ($\alpha > 0$).

# The PDF of the chi-square distribution

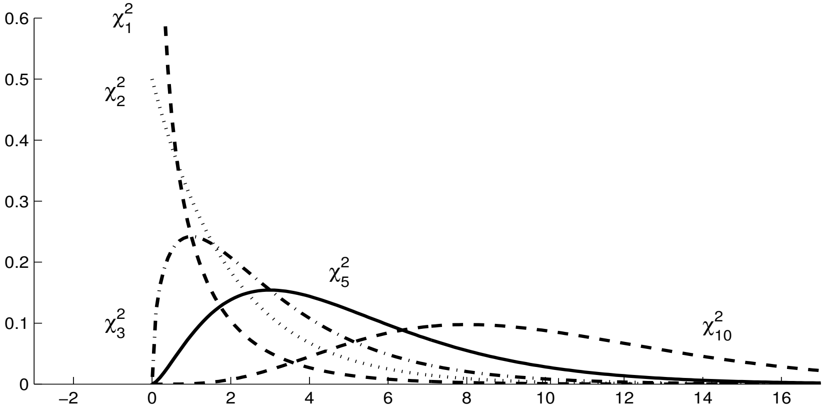


Figure: The PDF of the chi-square distribution for degrees of freedom 1, 2, 3, 5, 10.

## The chi-square distribution converges to the normal distribution

If $Y_n \sim \chi^2_n$, then its standardized version converges to standard normal distribution

$$\frac{Y_n - n}{\sqrt{2n}} \implies \mathcal{N}(0,1) \quad \text{as} \quad n \to \infty.$$

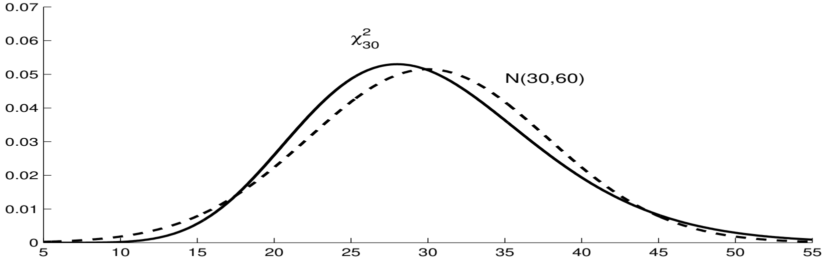


Figure: The PDF's of $\chi^2_{30}$ and $\mathcal{N}(30,60)$.

# The Student distribution

Let $X$ and $Y$ be independent with $X \sim \mathcal{N}(0,1)$ and $Y \sim \chi_n^2$. Then the distribution of

$$\frac{X}{\sqrt{Y/n}}$$

is called Student's t-distribution with degree of freedom $n$.
The PDF of the t-distribution with degree of freedom $n$ is

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi}\sqrt{n}\,\Gamma\left(\frac{n}{2}\right)\left(1+\frac{x^2}{n}\right)^{\frac{n+1}{2}}},$$

for $x \in \mathbb{R}$.

# The PDF of the t-distribution

If $n \to \infty$, then the t-distribution with degree of freedom $n$ converges to the standard normal distribution.

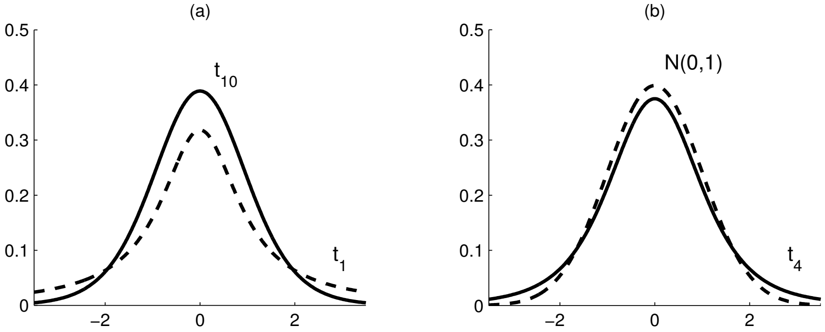


Figure: The PDF's of $t_1$ and $t_{10}$ (left) and the PDF's of $t_4$ and the standard normal PDF (right).

# Sample from normally distributed population

**Theorem.** Let

$$X_1, X_2, \dots, X_n$$

be a sample from $\mathcal{N}(m, \sigma^2)$.
Then $\bar{X}$ and $s_n^{*2}$ are independent with

$$\bar{X} \sim \mathcal{N}\left(m, \frac{\sigma^2}{n}\right), \qquad \frac{(n-1)s_n^{*2}}{\sigma^2} \sim \chi^2_{n-1}.$$

Therefore

$$\frac{\bar{X} - m}{\frac{s_n^*}{\sqrt{n}}}$$

is of Student distribution with degree of freedom $n - 1$.

## Student's t-test

The $t$-test is similar to the $z$-test. The major difference of the two tests is that in the case of $t$-test **the variance is unknown**.
So let

$$X_1, X_2, \dots, X_n$$

be a sample from population $\mathcal{N}(m, \sigma^2)$, where $m$ and $\sigma$ are unknown.
The null hypothesis is

$$H_0 \quad : \quad m = m_0,$$

where $m_0$ is a fixed number.
We start with the two-sided alternative hypothesis

$$H_1 \quad : \quad m \neq m_0$$

# The t-test

Calculate the empirical mean and the corrected empirical variance from the sample. From our previous theorem we know, that

$$\frac{\bar{X} - m}{\frac{s_n^*}{\sqrt{n}}}$$

has $t$-distribution with degree of freedom $n - 1$. So the test statistic of $t$-test is

$$t = \frac{\bar{X} - m_0}{\frac{s_n^*}{\sqrt{n}}}$$

We emphasize that here we use $m_0$.
So $t$ has $t_{n-1}$ distribution if and only if $m = m_0$, i.e. the the null hypothesis $H_0$ is true.

# The two-sided t-test

Using significance level $\alpha$, we reject $H_0$ if $|t| \geq t_{\alpha/2}$,
where the critical value $t_{\alpha/2}$ satisfies $F(t_{\alpha/2}) = 1 - \frac{\alpha}{2}$
and $F$ is the CDF of $t$-distribution with degree of freedom $n-1$.

# The two-sided t-test

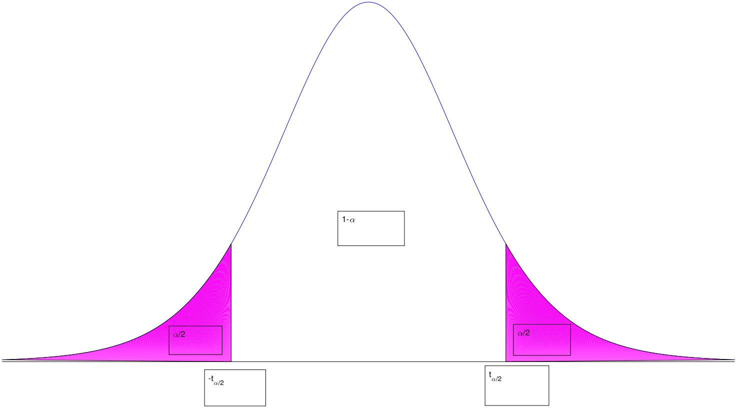


Figure: How to find the critical value for two-sided $t$-test? We cut both tails of the $t_{n-1}$ PDF

## The two-sided t-test

**Example.** We want to test if the weight of a certain kind of apple is 150 gram. So

$$H_0 \quad : \quad m = 150, \qquad H_1 \quad : \quad m \neq 150$$

We collect 100 apples and weight them. Then calculate the empirical mean and the corrected empirical variance from the sample. We obtain $\bar{X} = 146$, and $s_n^* = 25$. Therefore

$$t = \frac{146 - 150}{\frac{25}{\sqrt{100}}} = -1.6$$

Use significance level 95%. From the table of $t_{99}$ distribution, we obtain that $t_{\alpha/2} = 1.98$.
As $|t| < t_{\alpha/2}$, so we accept $H_0$.

## t-tests

**One-sided $t$-tests**
There are two versions of the one sided $t$-tests: the alternative hypothesis can be either $H_1 \ : \ m < m_0$ or $H_1 \ : \ m > m_0$. To handle these cases we should modify the two-sided $t$-test like we done in the case of $z$-test.
**Two-sample $t$-tests** Let

$$X_1, X_2, \ldots, X_{n_1}, \quad and \quad Y_1, Y_2, \ldots, Y_{n_2},$$

be two independent samples from population $\mathcal{N}(m_1, \sigma_1^2)$, resp. $\mathcal{N}(m_2, \sigma_2^2)$, where $\sigma_1$ and $\sigma_2$ are unknown.
The null hypothesis is

$$H_0 \ : \ m_1 = m_2.$$