What is probable and what isn't?
What is different and what isn't?

Normal distribution, probability, significance level. F-test and Student's t-test.

Katalin Mády

Pázmány Summer Course in Linguistics for Linguistics Students, Piliscsaba

26 August, 2013

Normal distribution

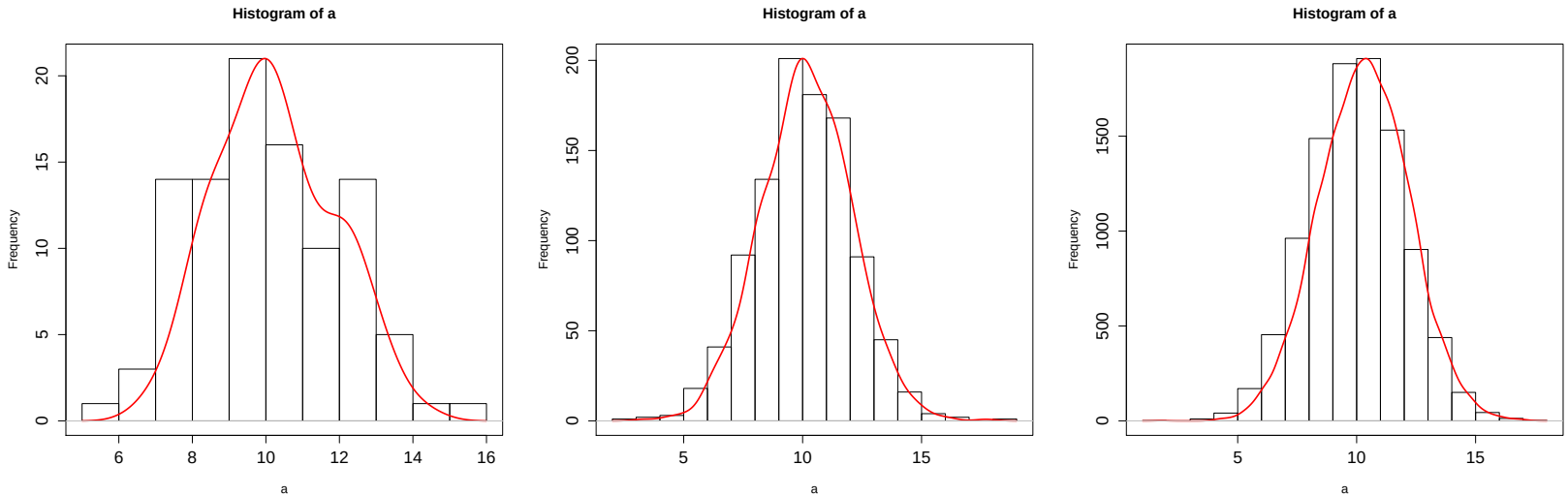
Most statistical tests require a data set with unimodal distribution, i.e. it has one single modus.

Tests for parametric data based on mean usually require data with normal distribution.

- Continuous variables,
- mode is found in the middle of the distribution and is equal with median and mean,
- the frequency of data decreases from the middle to both directions symmetrically,
- an approximate bell-shape (Gaussian curve).
- it is asymptotic: approximates 0 but does not reach it.

Normal distribution (N)

It depends on n (number of elements), and on k (number of classes or categories).



Histogram and density curve for normally distributed data set with 100, 1000 and 10000 elements.
Mean = 10, standard deviation = 2.
The higher n, the closer the density to normal distribution.

Where does normal distribution occur?

Most types of data are normally distributed if sample is large enough.

- IQ scores – but this is a deliberate definition.
- Heights and weights of human beings.
- Weights and lengths of fish caught in a large lake.
- Annual incomes of households.
- Grades on exams or in large classes.
- Climate (e.g. high and low temperatures over time).

Where does normal distribution not occur?

- Financial indicators and economic data.
- Price changes, interest rates, stock values, commodity prices, exchange rates.
- Lifetimes of humans.
- Lifetimes of mechanical and electronic devices.
- Waiting times (queuing).
- Safety data (e.g. car crash data).

Parameters for N

We have to distingush between population and sample. We deal with a sample but want to say something about the population.

Relevant measures:
μ: estimated mean of population.
σ: estimated standard deviation of population.
$\bar{x}$: mean of sample.

s: standard deviation of sample.

mean of sample:

$$\bar{x} = \frac{x_1+x_2+x_3+\cdots+x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Variance

Dispersion measure for normally distributed parametric data. Squared deviation from the mean: each value is subtracted from the mean, and the difference is squared.

Why squared? We are not interested in the direction of the deviation, i.e. whether the difference to the mean is negative or positive. Squared values are necessarily positive.

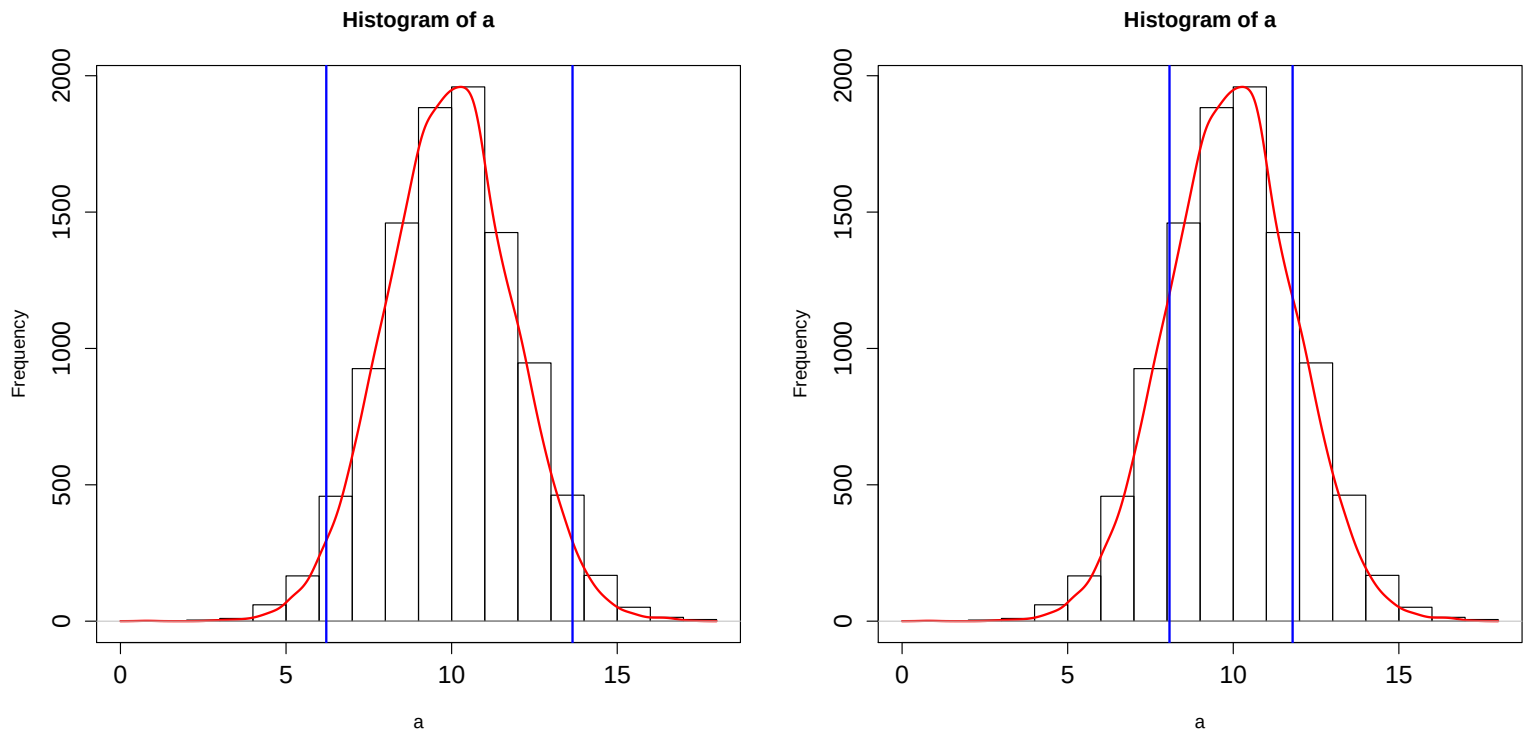
$$variance = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Why does the denominator contain $n-1$ instead of n? It expresses that we deal with a sample and not the population, and that sample s is only an approximation of population σ.

Standard deviation (SD or s)

The square root of variance: it can be interpreted as a measure of the dispersion (width of distribution).

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$



Left: variation (6, 14), right: standard deviation (8, 12).

Relevance of normal distribution for inferential statistics

Assumption: the means of n samples $\bar{X}$ are normally distributed around the mean of the population μ with a standard deviation σ. Their distribution is $\mathrm{N}(\mu, \sigma/\sqrt{n})$

where the standard deviation equals the standard error:

$$se = \frac{s}{\sqrt{n}}$$

The standard deviation expresses the distance of single data points around the sample mean.
The standard error expresses the distance of sample means around the hypothesised population mean.

Importance: mean and SD of a single sample allow for assumptions about unknown values of population.

Difference between SD of sample and population

Let's assume we are throwing 100 times with two dice and add their values. What is the most frequent sum we will receive, and what are the least frequent ones?

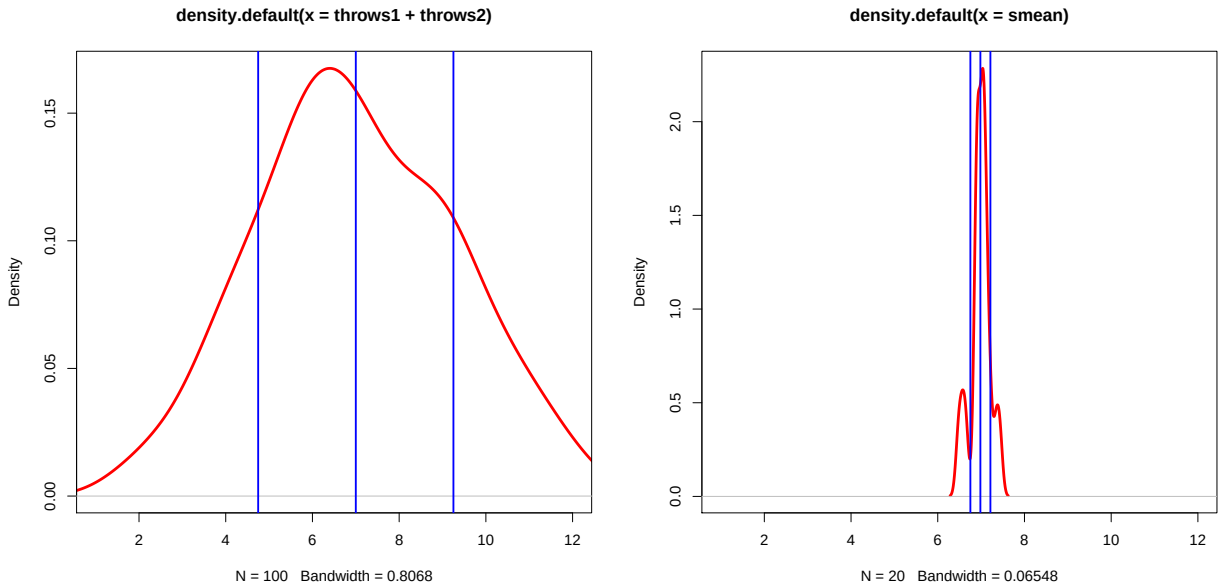
Difference between SD of sample and population

Let's assume we are throwing 100 times with two dice and add their values. What is the most frequent sum we will receive, and what are the least frequent ones?

Since 7 is the sum of the most combinations (1+6, 2+5, 3+4, 4+3, 5+2, 1+6), this is expected to be the most frequent sum. Both 2 and 12 are sums of one combination (1+1, 6+6), so these are expected to be the least frequent sums.

Distribution of sums and means

Left: density of 100 throws, right: density of 20 means of 100 throws.



Sample mean $= 7$, mean of means $= 6.98$.
Sample SD $= 2.25$, SD of means $= 0.23$
Standard error based on sample: $2.25/\sqrt{100} = 0.225$
– a close approximation to SD of 20 means.

z transformation

SD is dependent on the number of elements and on mean $\Rightarrow$ the distribution of samples with different sizes and different means are not comparable.

Solution: standardising to a z-score. For each value:

$$z_i = \frac{x_i - \mu}{\sigma}$$

i.e. each value is subtracted from the population mean (= sample mean) and divided by the standard deviation of the population.

Remember: SD of population has n in its denominator, whereas sample SD has $n - 1$.

Standard normal distribution

Normal distribution is described as $N(\mu, \sigma)$.

mean: $x_i = \mu = \bar{x}$

after z-transformation:

$$z_i = \frac{\mu - \mu}{\sigma} = 0$$

standard deviation: $\sigma = \mu + \sigma$

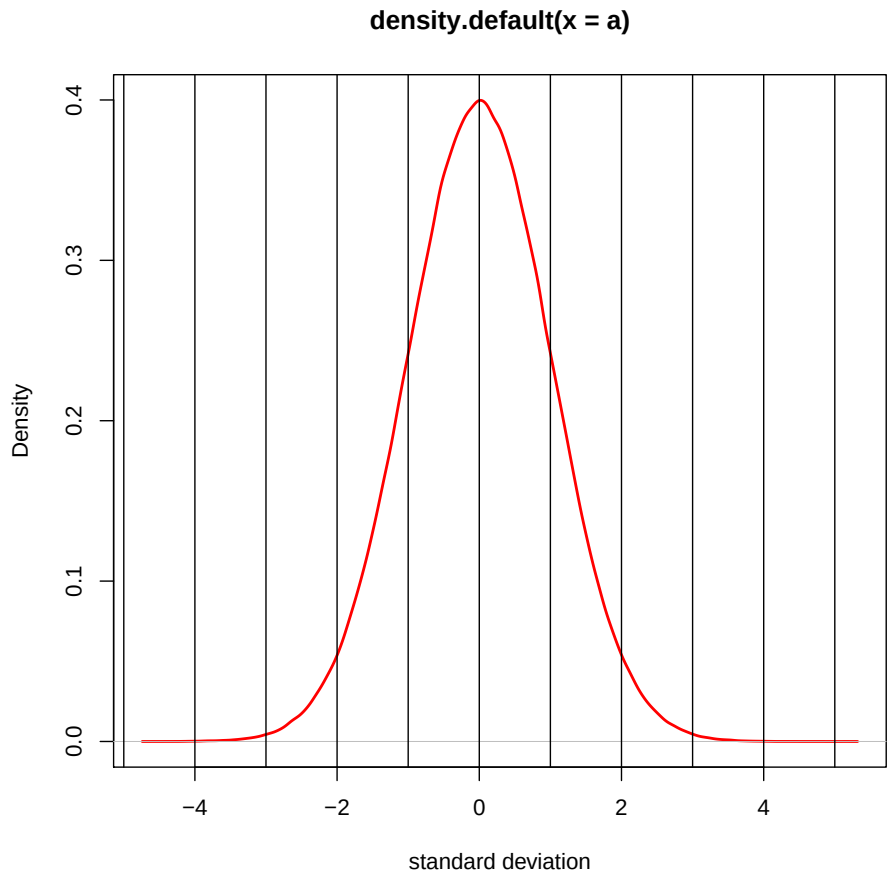
$$z_i = \frac{(\mu + \sigma) - \mu}{\sigma} = 1$$

description of standard normal distribution: $N(0, 1)$

Standard normal distribution

Density function:

x-axis: standard deviation, range: $\sigma = -\infty \cdots + \infty$.

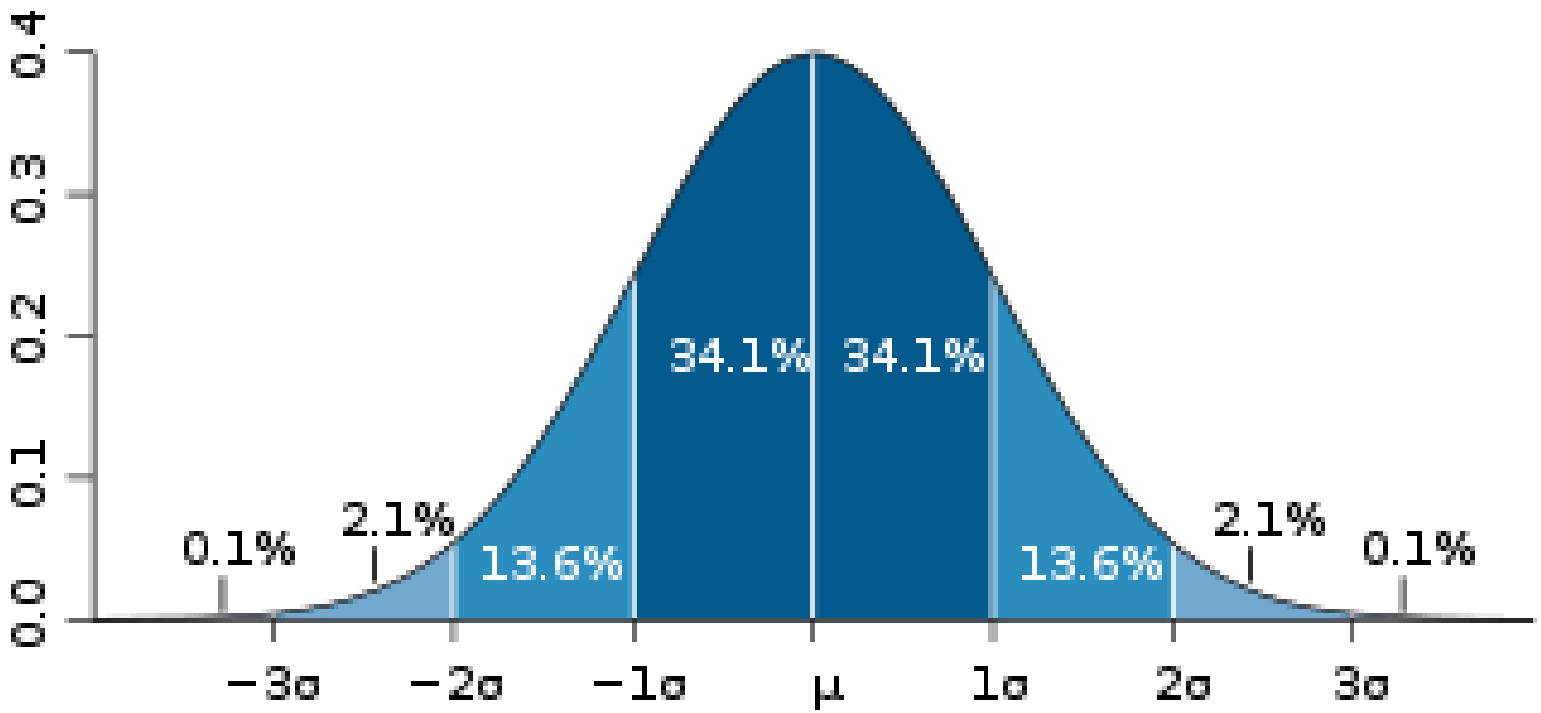


Density function

Characteristics:

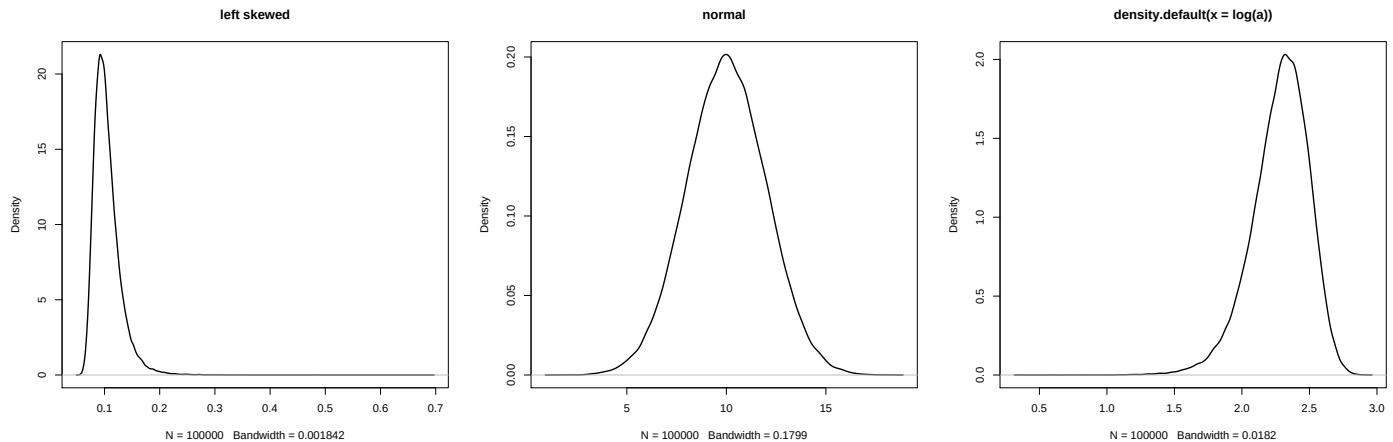
- Area between x-axis and density function equals 1.
- 50% of all cases is located left of mean.
- Range between $\sigma = -1 \cdots + 1$ includes 68.27% of all cases.
- Range between $\sigma = -2 \cdots + 2$ includes 95.45% of all cases.
- Range between $\sigma = -3 \cdots + 3$ includes 99.73% of all cases.

Standard normal distribution



Normality tests

Tests: Kolmogorov-Smirnov (in SPSS) or Wilk-Shapiro test (in R).
R-funkció: `shapiro.test(vektor)`



If distribution is unimodal, but not symmetrical: transformation methods such as

$x = log(x)$, $x = 1/x$, $x = \sqrt{x}$ etc. might result in a normal distribution.

Probability

Probability in everyday life

Meaning: an estimation based on your general experience (how often did a certain event happen in the given circumstances). E.g. "It's probably going to rain" (because it often does if the clouds are grey).

Probability never means a confidental knowledge about the truth! Sometimes it is not going to rain even if the clouds are grey.

So, everyday life does not tell us much about probability...

How probable is it that is is going to rain tomorrow?

How probable is it that is is going to rain tomorrow?

70%: possibility that weather conditions are identical on adjacent days.

Probability in gambling

Do you receive heads or tails if you toss a coin?

Observations: 10 tosses, 20, 30...

The higher the number of tosses, the closer the number of heads to 0.5.

Definition of empirical probability:

P = heads/all tosses

where number of tosses tends to infinity.

P's value ranges between 0 (absolutely improbable) and 1 (absolutely certain).

Examples

- Throws with a die: given number/all possibilities $= 1/6 = 0.17$.
- Pulling an as from a set of Swiss cards: number of asses/all cards $= 4/32 = 0.124$.
- The probability to throw two heads after each other: head+head/all possible combinations $= 1/4 = 0.25$.
- The probability that a randomly chosen person is a female with an academic background: proportion of academics in that region 22.4%, proportion of females 50%: 0.224*0.5 $= 0.112$.

Confidence interval

Question: is it true that the mean of the random sample is within the range of the sample means distributed around the population mean?

Difficulty: μ is not known, only $\bar{x}$.
$\Rightarrow$ a decision with certainty is not possible, only with a probability within a defined confidence interval.

Question: can we claim with a probability of 95% that the sample mean $\bar{x}$ is within the range of the sample means that are distributed around $\mu$by the standard error?

The confidence level is set to $p = 0.95$.

Our goal: to define the critical values of a confidence interval around the mean in positive and negative direction.

Critical values of the confidence interval

Our formula:
$p(-1.96 * se + \mu < \bar{x} < \mu + 1.96 * se) = 0.95$

Problem: μ is still not known.

After simple mathematical transformations we arrive at

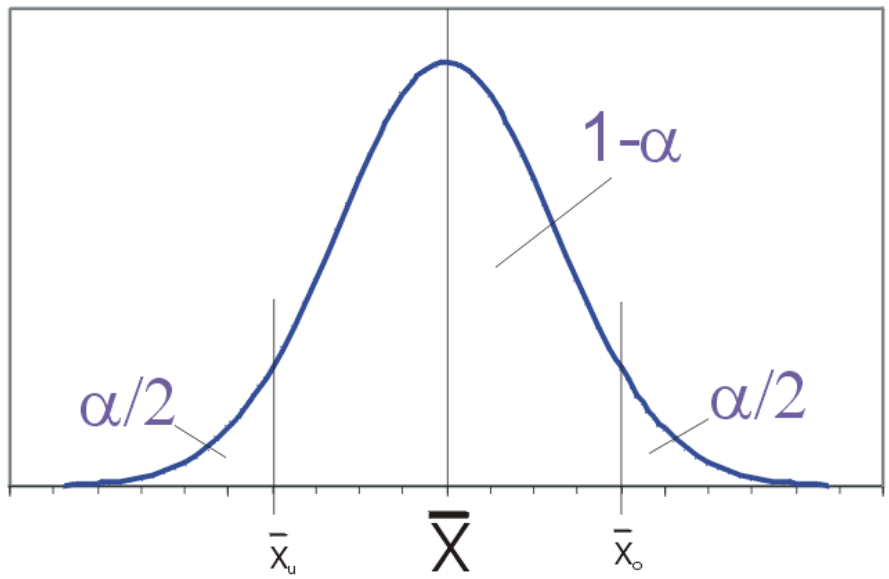
$p(-1.96 * se + \bar{x} < \mu < \bar{x} + 1.96 * se) = 0.95$

Here μ is not necessary for the limit calculation any more. Standard error and sample mean are known values.

But why 1.96? Remember: a standard deviation of 2 includes 95.45% of all cases. Here we only need 95%.

Alpha

Area outside of confidence interval: $\alpha = 1 - p$.



Remember that alpha has to be divided by 2 for the low and high end of the distribution if H_0 claims the equality of the two samples (see below).

Testing the alternate hypothesis

Remember: we test our hypothesis by falsification, i.e. we test the opposite hypothesis.

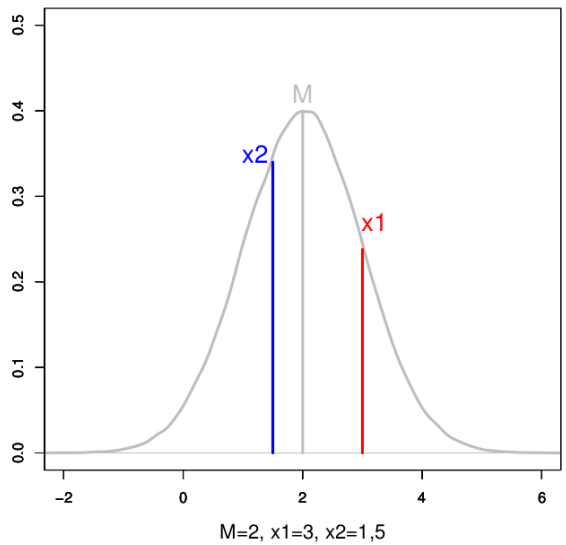
In empirical research, we usually hope to receive a mean outside of $1 - p$, i.e. within the range of α
$\Rightarrow$ the significance level is usually given by the value of α i.e. as 0.05 or 5%.

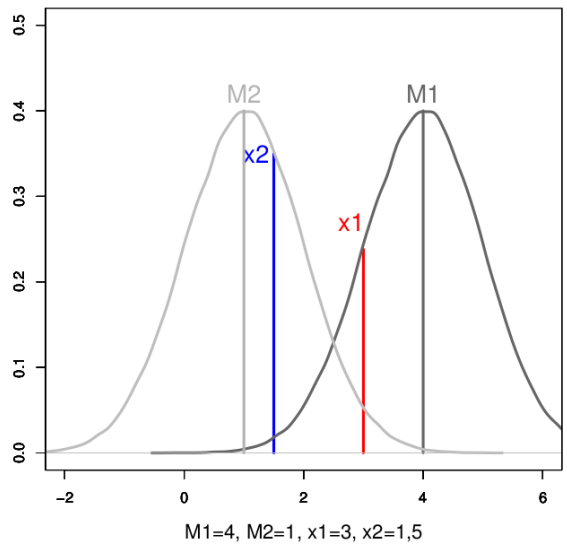
If we want to prove that a given sample does not belong into the confidence interval p, i.e. to 95% of the probable means, then the sample mean has to belong to the interval $\alpha/2$ on either the lower or the upper end.

This is bad news, because the probability that we can reject H_0 is in fact not 5%, but 2.5% in all two-sided tests where we assume equality.

Testing the null hypothesis

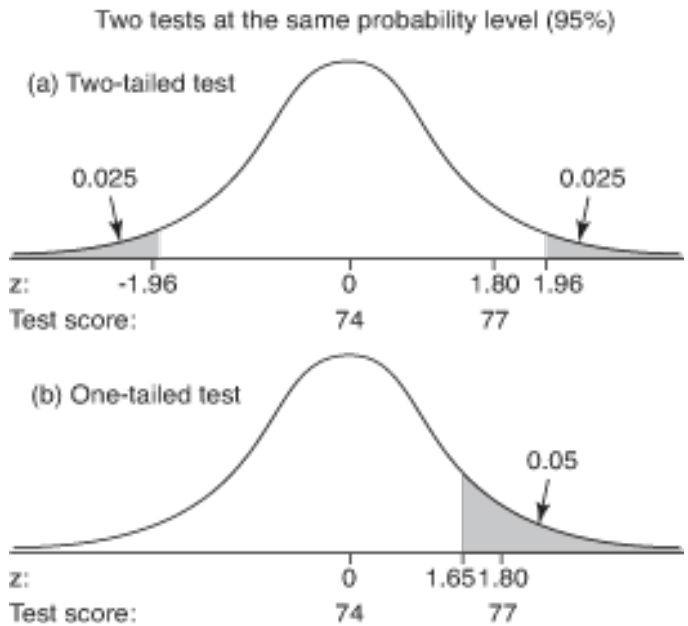
- Assumption: sample a belongs to a different population than sample b.
- Null hypothesis (H_0): samples a and b belong to the same population, i.e. their means are distributed around the same population mean μ.
- Alternate hypothesis (H_1): in can be claimed with p probability that the mean of sample b belongs to a different population than the mean of sample a.





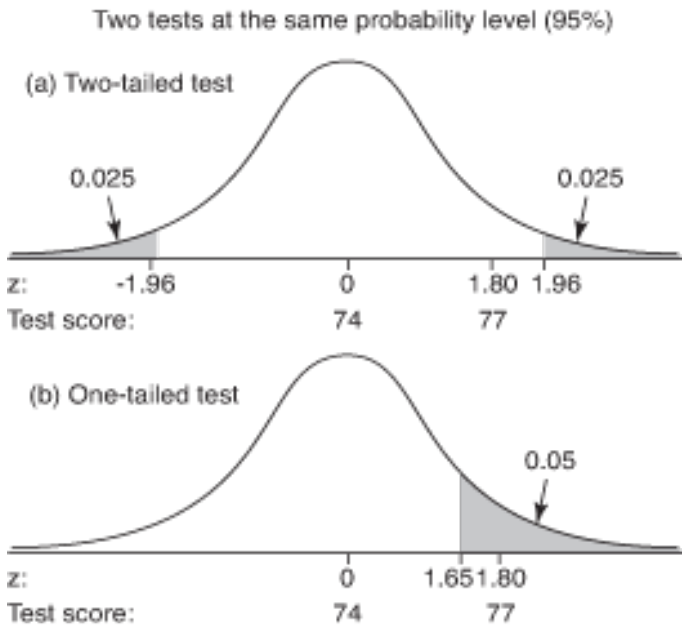
Hypothesis testing with a confidence of p = 95%

1. H_1: *a* differs from *b*-től with a high probability.
   H_0: *a* and *b* belong to the same population. Rejection: if $\bar{x}$ is within the range of $\alpha/2$ at one of the edges of the density function $\Rightarrow$ two-tailed test (upper figure).



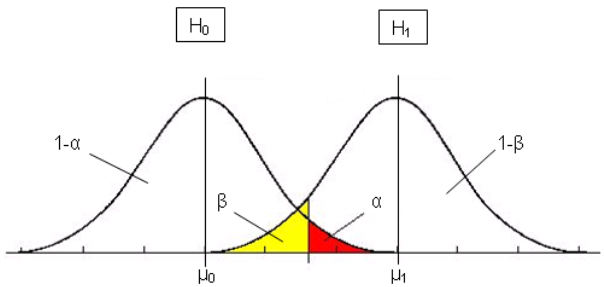
Hypothesis testing with a confidence of p = 95%

1. H_1: *a* differs from *b*-től with a high probability.
   H_0: *a* and *b* belong to the same population. Rejection: if $\bar{x}$ is within the range of $\alpha/2$ at one of the edges of the density function $\Rightarrow$ two-tailed test (upper figure).
2. H_1: *a* is larger than *b* with a high probability.
   H_0: *a* is not larger than *b*. Rejection: if $\bar{x}$ is within the range of α at the right edge of the density function $\Rightarrow$ one-sided test (lower figure).



Error types

1. α **error (Type I error):** H_0 is rejected because mean is outside of the range of the conficende interval – it belongs toα (red area).
2. β **error (Type II error):** H_0 is not rejected, although the mean belongs to a different population (yellow area).



	H_0 is not rejected	H_0 is rejected
H_0 is true	correct decision	α error (false-positive)
H_1 is true	β error (false-negative)	correct decision

Degrees of freedom

Number of elements that can be varied freely without changing a certain relevant feature of the sample.

$df = n - 1$

Why? Let's take a sample $n = 5$ with a mean $\bar{x} = 10$. How many elements can be changed without effecting the mean?

Degrees of freedom

Number of elements that can be varied freely without changing a certain relevant feature of the sample.

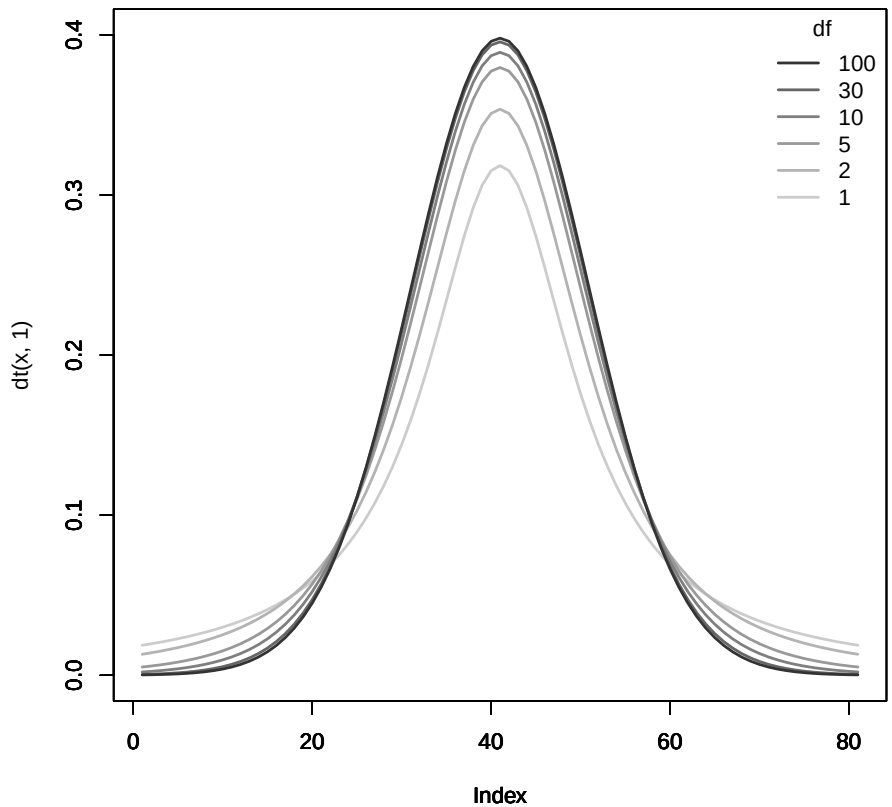
$df = n - 1$

Why? Let's take a sample $n = 5$ with a mean $\bar{x} = 10$. How many elements can be changed without effecting the mean?

Four, since 5th element has to be chosen so that sample mean still remains 10.

t-distribution and degrees of freedom

Flatness of t-distribution depends on degrees of freedom. The larger df, the closer the critical value (limits of confidence interval) to the mean.

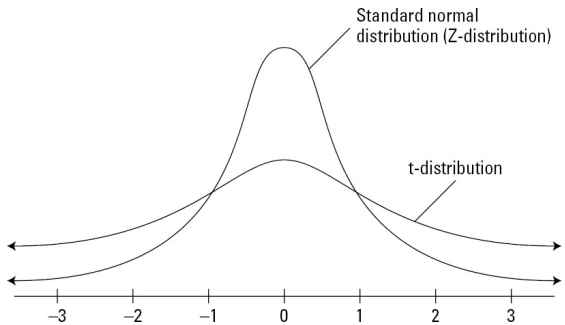


Again, the larger the sample size, the easier it is to reject the null hypothesis.

Comparing samples based on t-distribution

If the standard deviation σ of the population is known, then the z-distribution can be used. Else, Student's t-distribution is used. Same formula as for z, but contains s instead of σ.

$$z_i = \frac{\mu - \mu}{\sigma} \qquad t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$



t-distribution depends on sample size and is flatter than z-distribution $\Rightarrow$ limits of confidence interval are further away from mean. The larger n, the closer t gets to z.

The larger the sample size, the better the approximation to the population.

Student's *t*-test

Remember that applying the *t*-distribution requires that data are

- metrical,
- symmetrical,
- normally distributed.

A *t*-test cannot be run on ordinal data, on asymmetrical and not normally distributed data (unless they can be transformed to normal distribution), and on bi- or multimodal data!

Unimodality has to be tested by creating a density function, normality by the Kolgomorov-Smirnov or Shapiro-Wilk test.

One-sample Student's t-test

- Requirement: normally distributed parametric variable.
- Application: if the mean of population or a large reference sample is known. E.g. IQ that is distributed around μ 100 per definition.
- Procedure: if t of the sample is larger than t that is defined for the given confidence level with the given $df \Rightarrow H_0$ is rejected.

There are 60 children in the kindergarten *Academic future*. They are said to be rather clever. Their mean IQ is 108, with a SD of 10. Are they more clever than average kids?

Example

Sample mean = 108, population mean = 100, SD = 10, number of elements = 60.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{108 - 100}{10/\sqrt{60}} = \frac{8}{1.29} = 6.2$$

t value that belongs to the defined confidence interval with 59 degrees of freedom = 1.67 (to be computed by a software or looked up in a table).

Example

Sample mean = 108, population mean = 100, SD = 10, number of elements = 60.

$$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{108-100}{10/\sqrt{60}} = \frac{8}{1.29} = 6.2$$

t value that belongs to the defined confidence interval with 59 degrees of freedom = 1.67 (to be computed by a software or looked up in a table).

$t_{sample} > t_{0.95(59)} = 6.2 > 1.67 \Rightarrow H_0$ is rejected.

Conclusion: children in the kindergarten *Academic future* are significantly more intelligent than the entire population.

Two-sample independent t-test

- Based on two samples, two unknown population means μ are compared.
- Reqiurements: normal distribution, homogeneity of sample variances.
- Samples are independent (e.g. Chinese children and Croatian pre-school children).

Here the population variance is estimated based on the sample variances.

However, sample variances are seldom homogeneous. Either we test the equality of sample variances, or we use the Welch-test.

Tests of homogeneity of variances

(At least) two samples:

- F-**test:** both samples are normally distributed, independent samples.
- **Levene-test:** approximative test, but it can be used for data without normal distribution and for more than two samples.
- **Bartlett-test:** normal distribution, also applicable to paired samples.

Welch-test

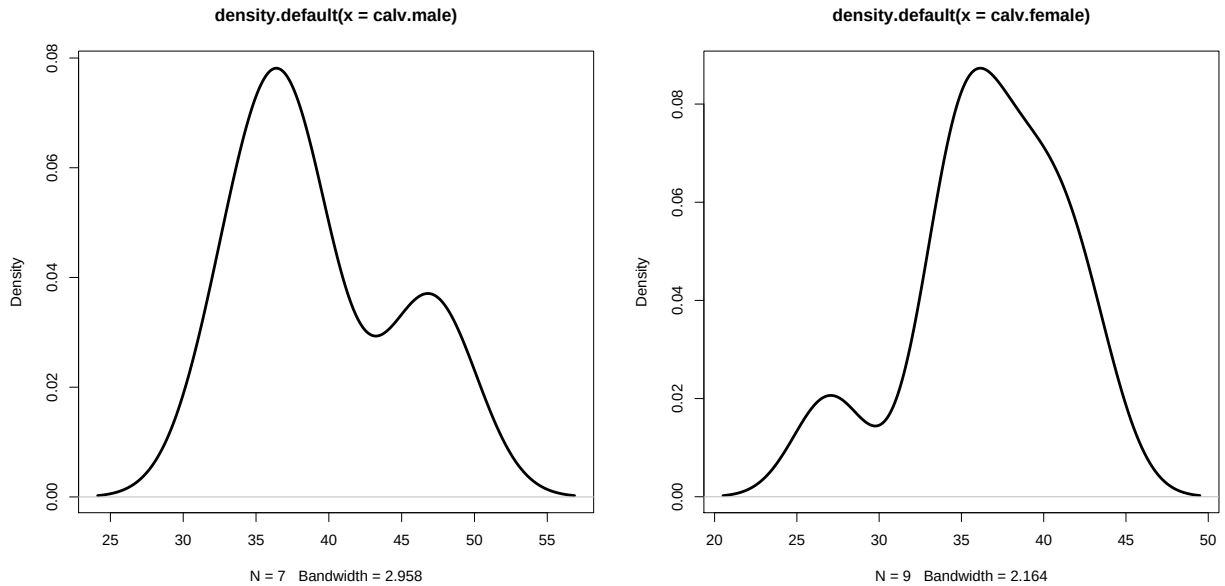
Same as two-sample independent t-test, but the equality of variances is not assumed.

The weight of new-born calves is given below. Is it true that male calves weigh more than female calves?

male calves (kg)	46	37	39	37	33	48	35		
female calves (kg)	27	37	35	41	35	34	43	38	40

Example: test of normality

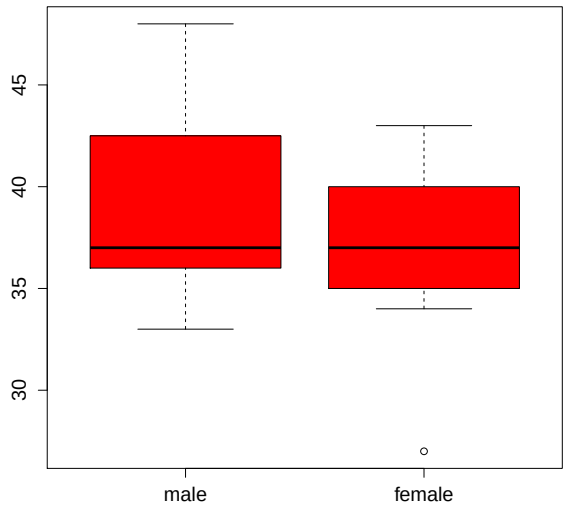
Plotting densities will give a tentantive idea about normality:



Test of normality based on Shapiro-Wilk-test: $p = 0.2592$ for males, $p = 0.6278$ for females $\Rightarrow$ both samples are normally distributed. (A value less than 0.05 would indicate non-normal distribution.)

Example: means and t-test

It is always useful to have a look at the boxplots first. If boxes overlap, the samples are expected not to differ significantly.



F-test: 0.63 $\Rightarrow$ equality of variances

Two-sample independent t-test with equal variances.
$p = 0.3276 \Rightarrow H_0$ is not rejected.
What if we used Welch-test instead?
$p = 0.3416 \rightarrow$ if you want to skip performing the F-test, use Welch-test wich is the default in R.

Two-sample paired t-test

If we have two kinds of observations on the same sample, the observations are not independent, but paired.

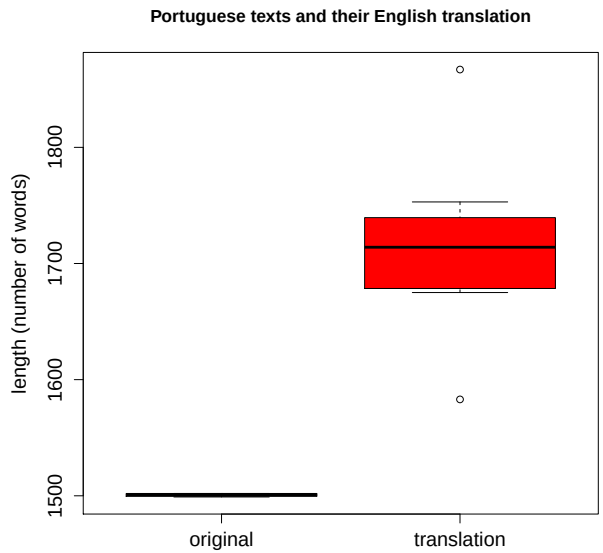
Requirement: differences between two observations for the same element are normally distributed. If $n \geq 30$, normality assumption can be ignored.

There is a data set with texts in English and Portuguese that have been translated into the other language. Does the length of the English translations of Portuguese texts significantly differ from their original length?

Example: normality of differences

Since our sample size is only 8, the length differences for the original Portuguese and English texts have to be tested for normality.

Sample differences are normally distributed $\Rightarrow$ paired t-test can be applied.



$p = 0.0001309$, thus English translations are significantly longer than Portuguese originals.