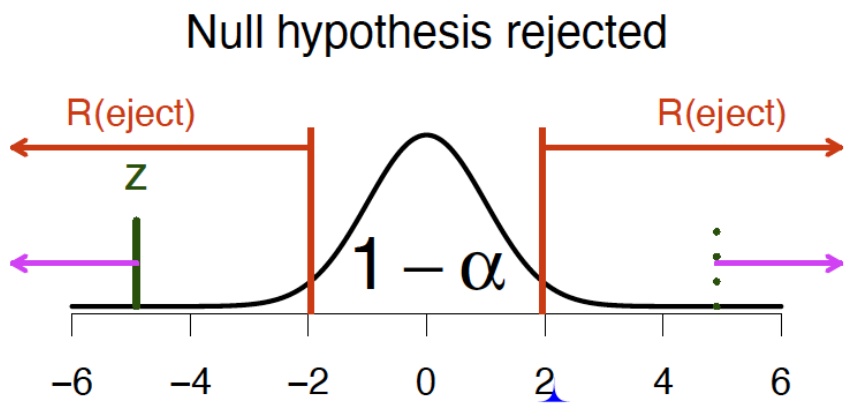
TECHNISCHE UNIVERSITÄT WIEN

Basic ideas of hypothesis testing



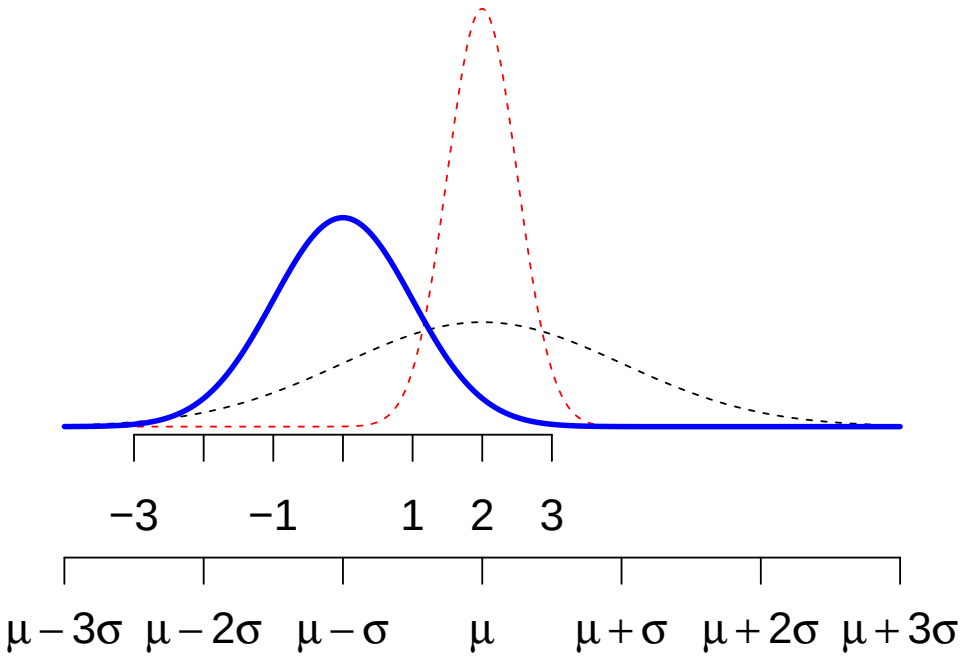
Slides were developed together with Dr. Michael Messer, former ASTAT member
Please send your comments to: T. Levajković, tijana.levajkovic@tuwien.ac.at

All examples are fictitious. All data are simulated and the graphics were created with the statistical program package R.

Sämtliche Beispiele sind frei erfunden. Alle Daten sind simuliert und die Grafiken wurden mit statistischen Programmpaket R erstellt.

Reminder

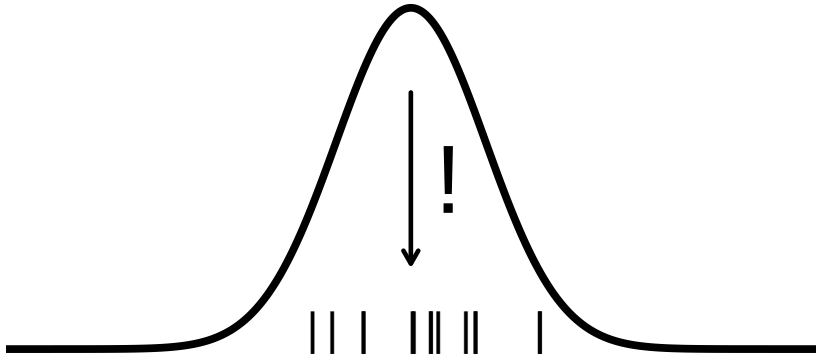
- How is the mean distributed under normal distribution?
- Let $X_1, \ldots, X_n$ be independent and identically distributed (i.i.d.) random variables and $X_1 \sim N(\mu, \sigma^2)$
- For the mean it holds $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$
  - $\bar{X}$ is also normally distributed
  - $\bar{X}$ has expectation $\mu_{\bar{X}} = \mu$ (equal to the expectation of X_i)
  - $\bar{X}$ has standard deviation $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ (decrease of factor $1/\sqrt{n}$)
    Interpretation: the typical deviation of the mean from its expectation is $\sigma/\sqrt{n}$
- Standardization: $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$



Overview

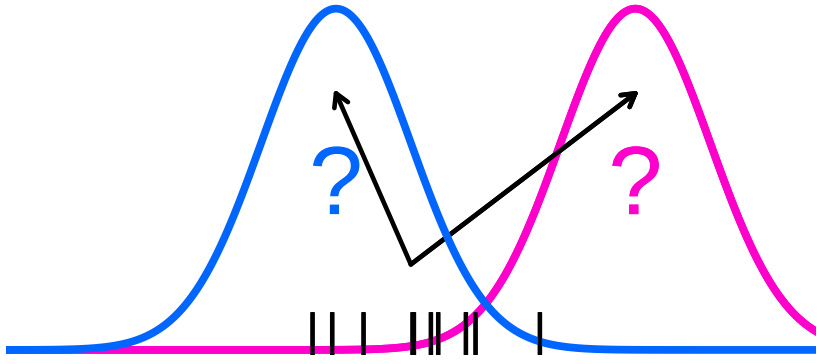
We differentiate:

Probability theory
(Stochastics)
=
Theory of randomness



and

Statistics
=
Description of data $\longrightarrow$
(using stochastic models)



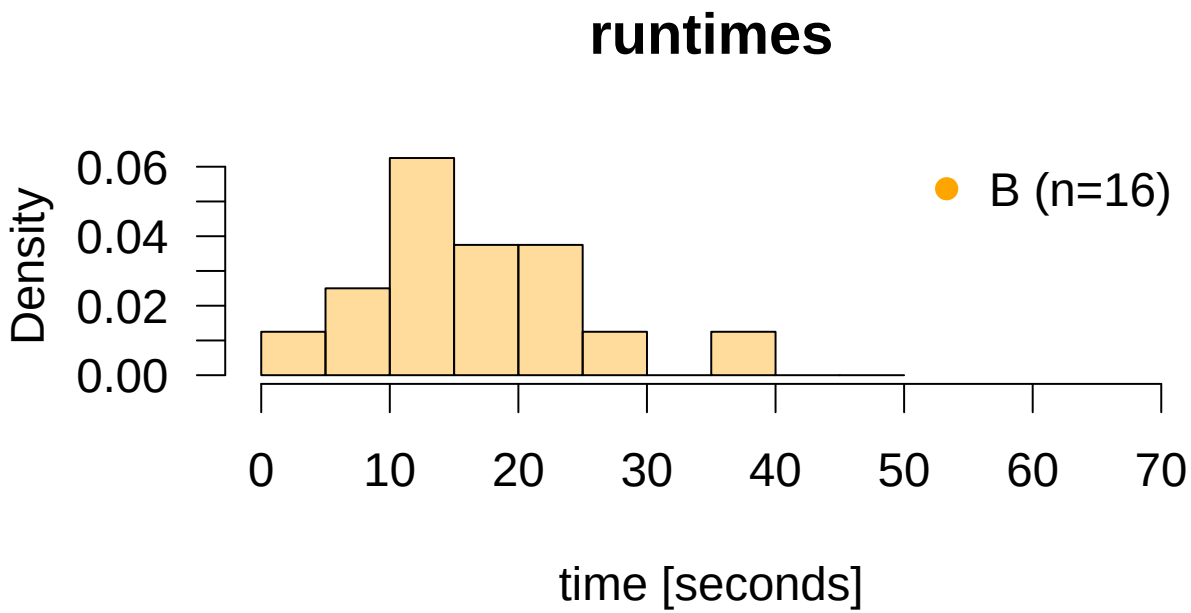
Previous lecture: Short excursion to descriptive Statistics
How do data look like? How can they be summarized?

Today: inferentiell Statistics (Modelling)
How did the data occur?

Basic problem

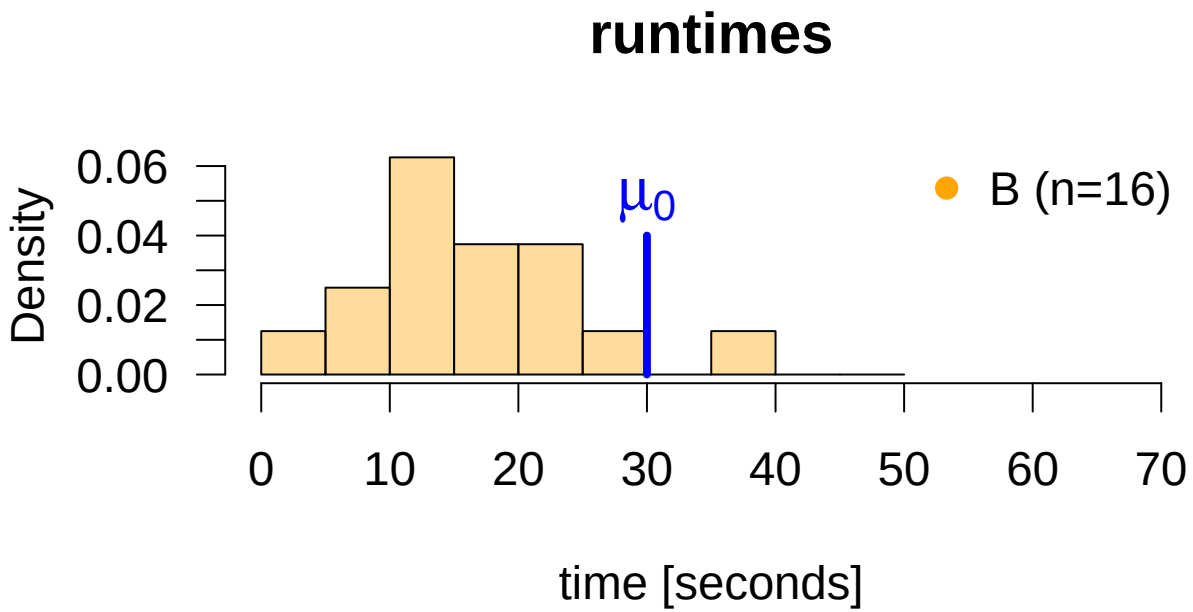
Reminder:

- Runtimes of an algorithm implemented by $n_A = 121$ students
- Additionally implemented from $n_B = 16$ students that took a certain programming course



- Distributions shifted against each other $\rightarrow$ the course seems to have a positive effect...
- In the following consider only group B (*one-sample situation*)

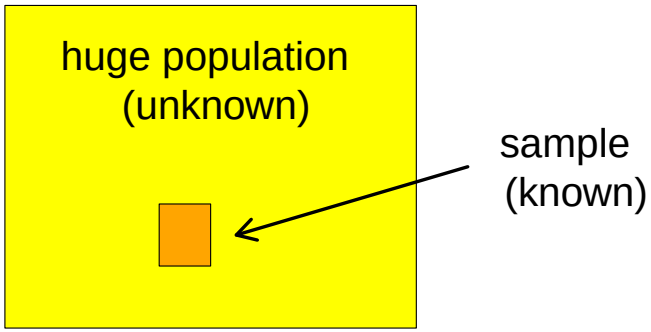
Basic problem



- Distributions shifted against each other $\rightarrow$ the course seems to have a positive effect...
- ...the lecturer is happy!
- A skeptic colleague claims: "the course is useless. The 16 students were just over average beforehand!" And further he claims:
  "The course was held by the lecturer a couple of times before. If all participants that have ever taken the course had implemented this algorithm, then the mean runtime would have been $\mu_0 = 30$."

Basic problem

Assertion:
"If all participants that have ever taken the course, had implemented this algorithm, then the mean runtime would have been $\mu_0 = 30$."

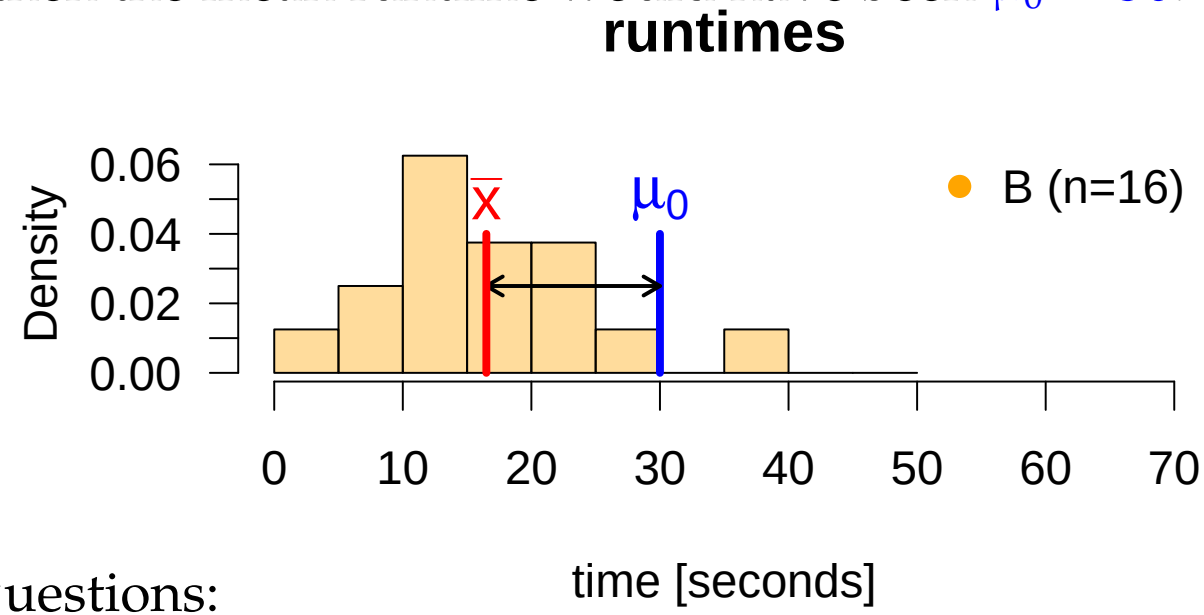


- Problem: Assertion about a huge unknown *population*
- However, a subset known: the *sample* $x_1, \ldots, x_n$
- Main questions:

  How 'compatible' are the data with the assertion?
- To do:
  Quantification of the 'discrepancy' of the data and the assertion

Basic problem

Assertion:
"If all participants that have ever taken the course, had implemented this algorithm, then the mean runtime would have been $\mu_0 = 30$."



- Main questions:

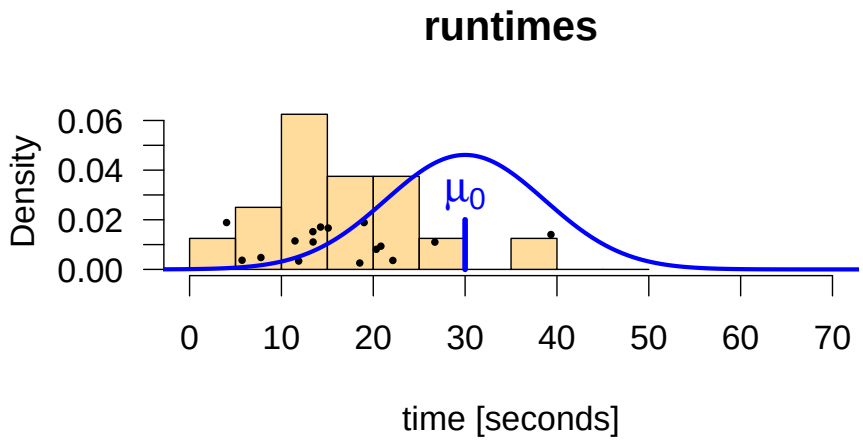
  How 'compatible' are the data with the assertion?

- To do:
  - 'Quantification of the discrepancy' between data and assertion
  - For example through the difference $d = \bar{x} - \mu_0$
  - Line of thought: $|d|$ large $\leftrightarrow$ data hardly compatible with the assertion
  - Question: What does 'large' mean? So far we do not have a notion for 'size'! $\rightarrow$ need the concept of the *statistical model*!

Statistical model

- Main questions: How 'compatible' are the data with the assertion?
- Idea of the statistical model:

> Interpret the data $x_1, \ldots, x_n$ as realizations of random variables $X_1, \ldots, X_n$

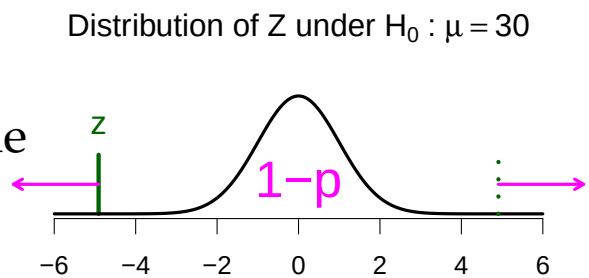


- Upside: Notion of 'size' through probability statements
  - Observe: almost all data lie in the left tail of the blue distribution
  - Interpretation: Assume that the data indeed derived from independent drawings of the blue distribution, then something unlike has happened. ($\leftrightarrow$ incompatibility of data and assertion)
  - Gain: In the context of the model, we can quantify the discrepancy via probability statements ('notion of size')
  - $\rightarrow$ *Hypothesis test*: the procedure is as follows...

Hypothesis test, exemplary: z-Tests

- *model assumption*: $X_1, \ldots, X_n$ i.i.d. RVs, with $X_1 \sim N(\mu, 11^2)$ and $\mu \in \mathbb{R}$ (here $n = 16$)
  (The data $x_1, \ldots, x_n$ are assumed to be realizations of i.i.d. normal-distributed RVs, with unknown expectation $\mu \in \mathbb{R}$, but known variance $\sigma^2 = 121$)
- *null hypothesis*: $H_0 : \mu = 30$
  (Describes the assertion: the claimed expectation is $\mu_0 = 30$)
- *test statistic:* for the evaluation of the data (measures discrepancy). Here z-statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \approx \frac{16.5 - 30}{11/4} \approx -4.9$$



- Theoretical distribution of the test statistic if H_0 is true

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$$

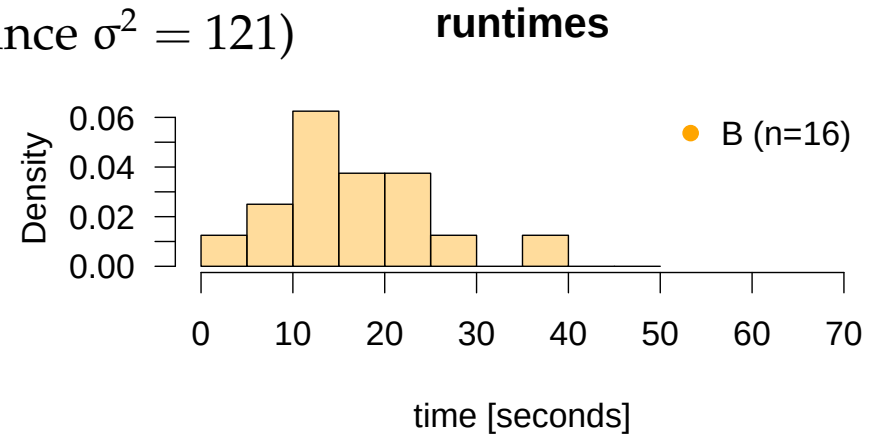
- *p-value:* quantifies the discrepancy (judge z according to the distribution of Z)

$$p = \mathbb{P}_{H_0}(|Z| \geqslant |z|) \approx 9 \cdot 10^{-7}$$

  Probability to make an observation which is *at least as extreme* as in the data, if H_0 holds true
- *Decision:* Reject the null hypothesis (reason: p is small)
  We say: the observed discrepancy was *significant* ($p < 10^{-6}$)
- *Interpretation*: the p-value is negligible. If the null hypothesis holds true, then something unlikely has happened. In that sense, the data are barely compatible with the null hypothesis.

Remark - Model assumptions

Model assumption: $X_1, \ldots, X_n$ i.i.d. RVs, with $X_1 \sim N(\mu, 11^2)$ and $\mu \in \mathbb{R}$ ($n = 16$)
(The data are realizations of i.i.d. normal-distributed RVs, with unknown expectation $\mu \in \mathbb{R}$, but known variance $\sigma^2 = 121$)



- Why are the assumptions reasonable?
  - Normal distribution: the data are distributed approximately bell-shaped
  - Knowledge of the standard deviation σ: Actually, this assumption is nonsense – why should we know something about the variance, when we do not know the expectation? Here assumed to be known for simplicity. Next lecture: replace σ by estimate $s \rightarrow$ yields t-test
  - Independence: We do not have a reason to assume that the observed individuals have too much in common.
  - In general: A model is always a simplification. The description of the 'reality' through a theoretical construct ('model') is basically always 'wrong'. A complicated model (which is possibly not appropriately understood) is often useless. Models should be chosen as 'simple' objects.
  - George Box: 'All models are wrong' (but some are useful)

Remark - Test statistic

The *test statistic* should accomplish two things:

1. It should measure the discrepancy between the data and the null hypothesis H_0

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

   Otherwise the procedure was nonsense.

2. It should be chosen such that its distribution was known under H_0

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$$

   Otherwise we can not judge the discrepancy in 1.

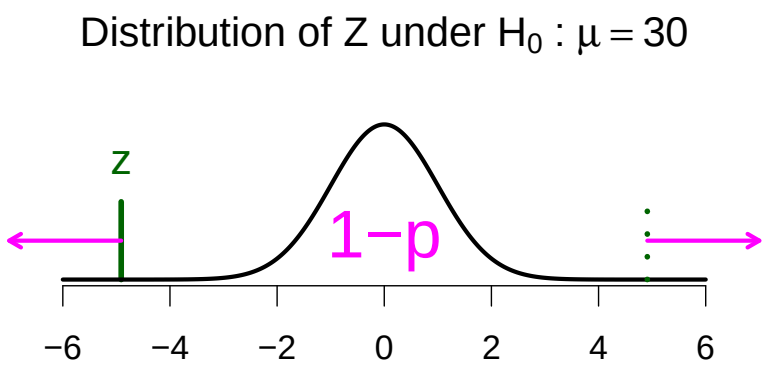
- The abbreviation $'\overset{H_0}{\sim}'$ in $Z \overset{H_0}{\sim} N(0,1)$ means that the left hand side is distributed according to the right hand side, if the null hypothesis holds true, short: 'under H_0'
- z measures the discrepancy of $\bar{x}$ from μ_0 in the units $\sigma/\sqrt{n}$, i.e., according to the variability of $\bar{X}$. Under H_0, the 'typical' deviation is one unit. In the data we observed $|z| \approx 4.9$ units. This is untypically large!
- If on the other hand H_0 does not hold true , i.e., if the X_i have an expectation μ_1, with $\mu_1 \neq \mu_0$, then Z is not distributed according to $N(0,1)$, as we did not center correctly

Remark - p-value

p-value (quantifies discrepancy)

$$p = \mathbb{P}_{H_0}(|Z| \geqslant |z|) \approx 9 \cdot 10^{-7}$$

Probability to make an observation which is *at least as extreme* as in the data, if the null hypothesis holds true.

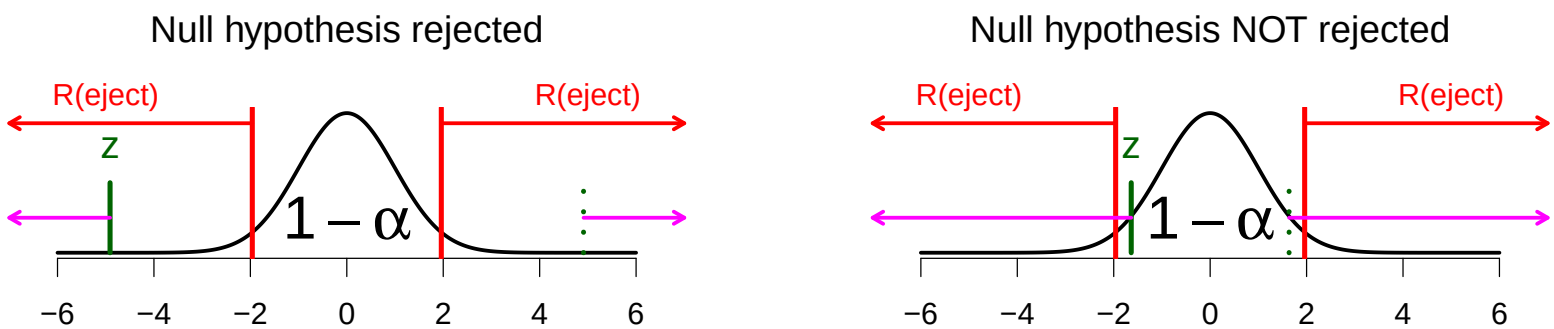


- Index H_0 again means : 'if the null hypothesis holds true'
- p-value 'small' $\leftrightarrow$ If H_0 holds true, then something unlikely has occurred $\leftrightarrow$ incompatibility of the the data and the assertion.
- The choice of 'small' / 'unlikely' is to be fixed by the statistician in advance(!). This is done by the choice of the *significance level* $\alpha \in (0, 1)$. Often $\alpha = 5\%$
- Decision rule:
  - $p \leqslant \alpha \leftrightarrow$ incompatible enough $\leftrightarrow$ reject null hypothesis
  - $p > \alpha \leftrightarrow$ not incompatible $\leftrightarrow$ do *not* reject the null hypothesis

Remark - p-value and rejection area

- Significance level $\alpha \in (0,1)$
- p-value: $p = \mathbb{P}_{H_0}(|Z| \geqslant |z|)$
- decision rule:

> - $p \leqslant \alpha \Leftrightarrow z \in R \leftrightarrow$ reject $H_0 \leftrightarrow$ say: 'the discrepancy was *significant*'
> - $p > \alpha \Leftrightarrow z \notin R \leftrightarrow$ do not reject $H_0 \leftrightarrow$ 'discrepancy was not significant'

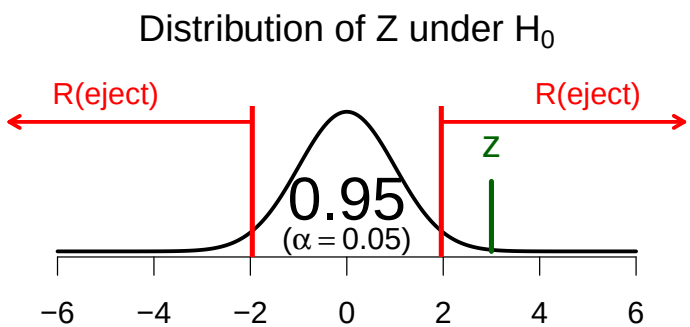


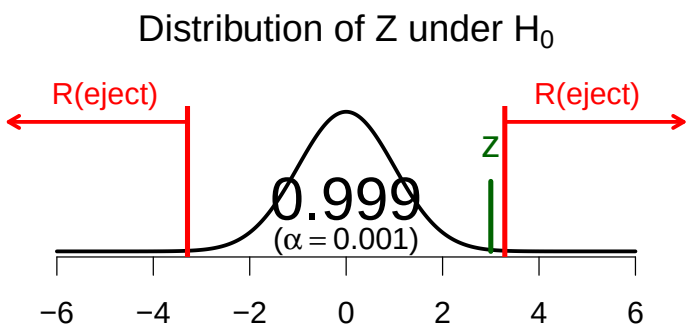
- A different point of view of the upper decision rule: the *rejection area* R
  - Reject H_0 if and only if $z \in R$
  - set R to the tails of the distribution $\leftrightarrow$ rejection at high discrepancy
  - Set R such that $\mathbb{P}_{H_0}(Z \in R) = \alpha$. Here $\alpha = 5\% \rightarrow R \approx (-\infty, -1.96] \cup [1.96, \infty)$
    Meaning: If H_0 holds true, then we falsely reject with probability α
  - Equivalent to decision rule via p-value

Remark - significance level

- Significance level $\alpha \in (0,1)$
- p-Wert: $P = \mathbb{P}_{H_0}(|Z| \geqslant |z|)$
- Decision rule:

> - $p \leqslant \alpha \Leftrightarrow z \in R \leftrightarrow$ reject $H_0 \leftrightarrow$ say: 'the discrepancy was *significant*'
> - $p > \alpha \Leftrightarrow z \notin R \leftrightarrow$ do not reject $H_0 \leftrightarrow$ 'discrepancy was not significant'

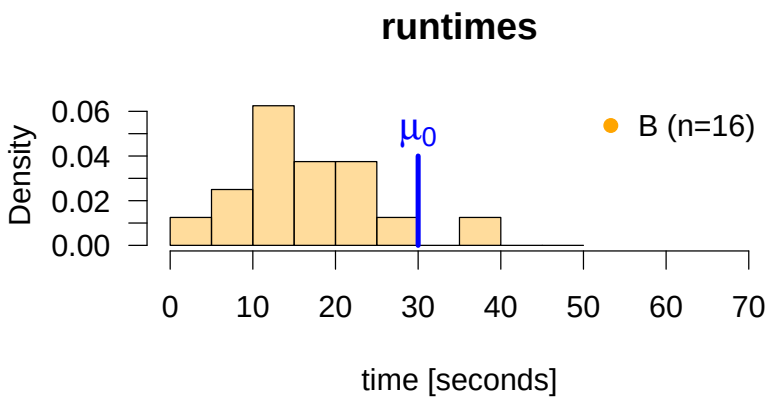


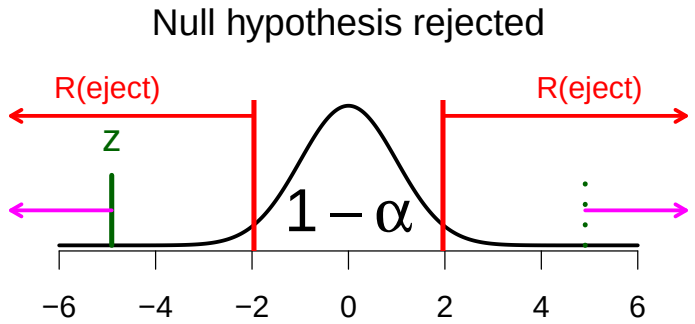


- Decrease of $\alpha \leftrightarrow$ smaller rejection area R, i.e., more 'strict' with rejecting
- popular choices: $\alpha = 5\%, 1\%, 0.1\%$.
- Here: Reject H_0 on the 5% level, but not on the 0.1% level
- Important: The level α has to be chosen in advance! It is self-delusive to increase α in order to reject!

Remark - Interpretation

A skeptic colleague claims: ”the course is useless. The 16 students were just over average beforehand!” And further he claims:
”The course was held by the lecturer a couple of times before. If all participants that have ever taken the course had implemented this algorithm, then the mean runtime would have been $\mu_0 = 30$.”



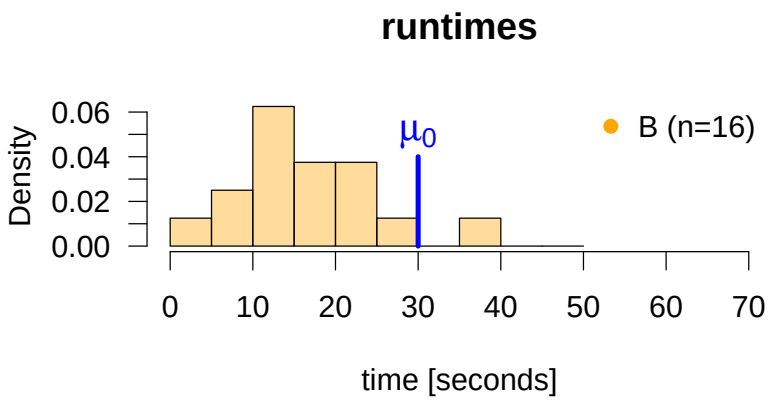


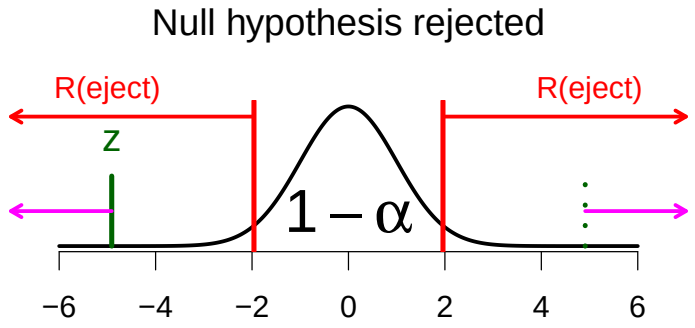
$$p = \mathbb{P}_{H_0}(|Z| \geqslant |z|) \approx 9 \cdot 10^{-7}$$

What is our opinion on that? (Apart from the fact, that we might find the delicate statement inappropriate from an interpersonal level)

- The data are not at all compatible with the assertion
- If the colleague is right, then a discrepancy which is at least as extreme as in the observed data, will occur in less than one of 1 million cases (as $p \approx 9 \cdot 10^{-7} < 10^{-6}$)
- $\rightarrow$ The data give us good reason to doubt the assertion!

Remark - Erroneous interpretation





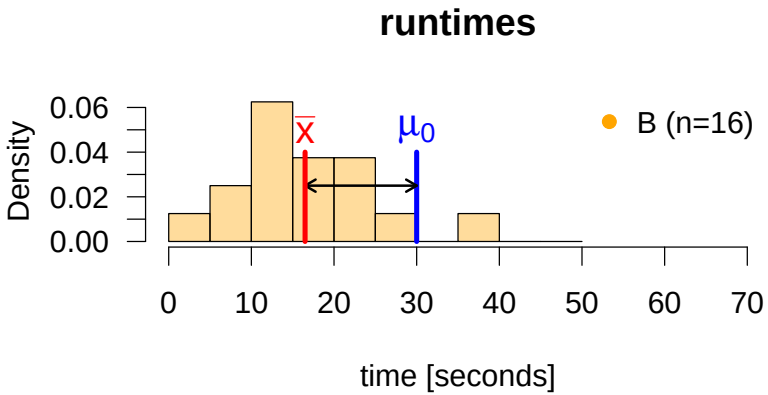
$$p = \mathbb{P}_{H_0}(|Z| \geqslant |z|) \approx 9 \cdot 10^{-7}$$

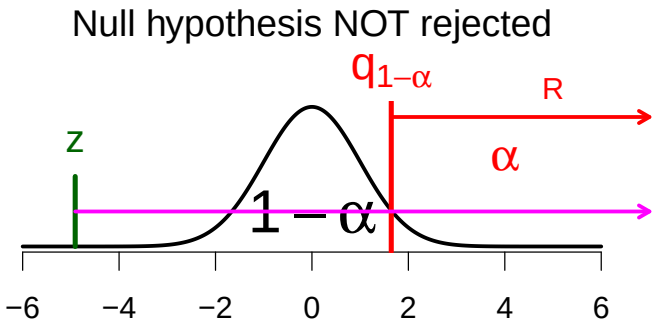
Popular misinterpretations:

- ~~the colleague is right~~, respectively ~~the null hypothesis holds true~~
- ~~the colleague is wrong~~, resp. ~~the null hypothesis does not hold true~~
- ~~the colleague is probably right / wrong~~, resp. ~~the null hypothesis is probably true / false~~
  We cannot say that, because we just make probability statements about the data, in fact under the theoretical assumption that the null hypothesis holds true.
  We cannot say anything about the huge unknown population / the null hypothesis!
  Rejecting the null hypothesis does not mean that its wrong or probably wrong!

Throughout the course we will stay in the realm of the so-called *frequentists approach* of statistics, where the unknown parameters (like μ) are treated as *non-random*. In contrast, the *Bayesian* statisticians will want to make probability statements about the parameter as in their world parameters are modeled as random variables.

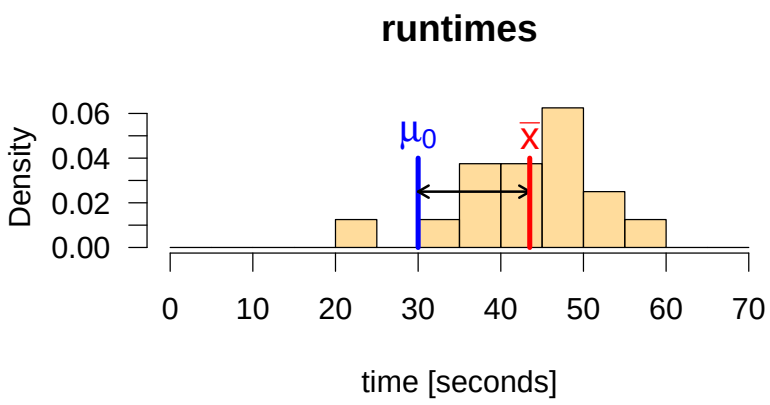
Two-sided and one-sided testing

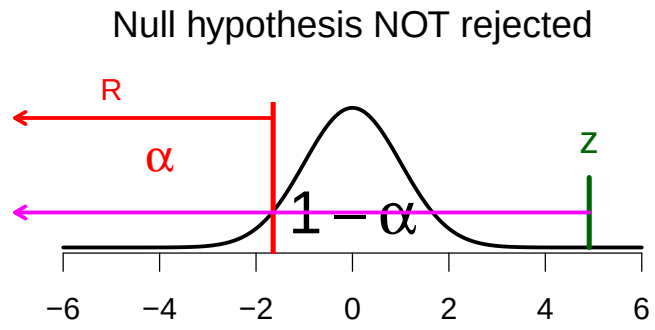




- H_0 rejected $\leftrightarrow z \in R$
- Let q_α be the α-quantile of $N(0,1)$, i.e., $\mathbb{P}(Y < q_\alpha) = \alpha$ for $Y \sim N(0,1)$
- Differentiation between *two-sided* and *one-sided* tests
  - two-sided: null hypothesis $H_0 : \mu = \mu_0$ ($\rightarrow$ alternative $H_A : \mu \neq \mu_0$)
    $R = (-\infty, q_{\alpha/2}] \cup [q_{1-\alpha/2}, \infty)$
    $p = \mathbb{P}_{H_0}(|Z| \geqslant |z|)$
    Extreme values if z speak against the null hypothesis
  - left-sided: null hypothese $H_0 : \mu \geqslant \mu_0$ ($\rightarrow$ alternative $H_A : \mu < \mu_0$)
    $R = (-\infty, q_\alpha]$
    $p = \mathbb{P}_{H_0}(Z \leqslant z)$
    Small values of z speak against the null hypothesis
  - right sided: null hypothesis $H_0 : \mu \leqslant \mu_0$ ($\rightarrow$ alternative $H_A : \mu > \mu_0$)
    $R = [q_{1-\alpha}, \infty)$
    $p = \mathbb{P}_{H_0}(Z \geqslant z)$
    Large values of z speak against the null hypothesis

Two-sided and one-sided testing



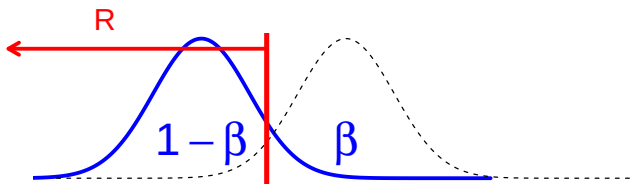


- In our example we could also have tested left-sided, because we knew in advance that the mean of the observed runtimes was smaller than the assertion, and we wanted to detect this fact of being 'faster'.
- However, we preferably formulate hypothesis before we observe the data.
- And it might have happened that the mean observed runtime was slower (larger) than the assertion. This would then have suggested a 'negative effect', which would not been detected with a left-sided test.
- Thus, depending on the context we need to decide of whether to perform a one- or two-sided test. (Rule of thumb: two-sided)

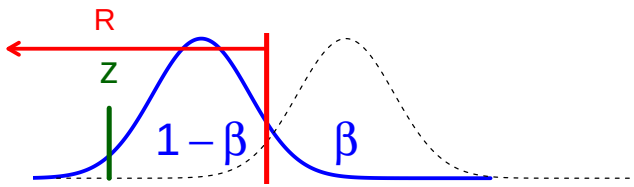
Errors and test power

Null hypothesis	rejected (with prob)	not rejected (with prob)
holds true	α-error $(= \alpha)$	$(= 1 - \alpha)$
does not hold true	(test power $= 1 - \beta$)	β-error $(= \beta)$

If H_0 holds NOT true



If H_0 holds NOT true



H_0 rightly rejected, test power 1−β

- α-error: H_0 is rejected although H_0 holds true
- The probability to commit the α-error is given through the choice of the significance level α, as by construction $\mathbb{P}_{H_0}(Z \in R) = \alpha$
- β-error: H_0 is not rejected although H_0 does not hold true
- The probability β to commit the β-error depends on the concrete alternative
- Test power $1 - \beta$ is the probability to reject H_0, if H_0 does not hold true

The question of whether we commit these errors can never be answered in practice, because hypotheses are theoretical assumptions

Multiple-choice questions

(1) For the p-value of a statistical test of significance level α it always holds true that

a. $p \leqslant \alpha/2$, if the null hypothesis was rejected
b. $p \geqslant 0$, if the null hypothesis was rejected
c. $p > 2\alpha$, if the null hypothesis was rejected
d. $p \leqslant \alpha$, if the null hypothesis was not rejected

(2) Which of the following is a **true** statement?

a. A well-planned hypothesis test should result in a statement either that the null hypothesis is true or that it is false.
b. Hypothesis tests are designed to measure the strength of evidence against the null hypothesis.
c. The alternative hypothesis is stated in terms of a sample statistic.
d. When the null hypothesis is rejected, it is because it is not true.

Multiple-choice questions

(3) Which of the following is a **true** statement?

a. The p-value of a test is the probability of obtaining a result as extreme (or more extreme) as the one obtained assuming the null hypothesis is false.
b. If the p-value for a test is 0.015, the probability that the null hypothesis is true is 0.015.
c. The alternative hypothesis is one-sided if there is interest in deviations from the null hypothesis in only one direction.
d. The larger the p-value, the more evidence there is against the null hypothesis.

(4) Suppose $H_0 : p = 0.4$, and the power of the test for the alternative hypothesis $p = 0.35$ is 0.75. Which of the following is a valid conclusion?

a. The probability of committing a Type I error is 0.05.
b. The probability of committing a Type II error is 0.65.
c. If the alternative $p = 0.35$ is true, the probability of failing to reject H_0 is 0.25.
d. If the null hypothesis is false, the probability of failing to reject it is 0.65.

Multiple-choice questions

(5) A company manufactures a synthetic rubber (jumping) bungee cord with a braided covering of natural rubber and a minimum breaking strength of 450 kg. If the mean breaking strength of a sample drops below a specified level, the production process is halted and the machinery inspected. Which of the following would result from a Type I error?

a. Halting the production process when too many cords break.
b. Halting the production process when the breaking strength is below the specified level.
c. Halting the production process when the breaking strength is within specifications.
d. None of the options given.

(6) In a hypothesis test:

a. the null hypothesis is what we are trying to prove
b. the alternate hypothesis is always assumed to be true
c. the alternate hypothesis is accepted unless there is sufficient evidence to say otherwise
d. the null hypothesis is not rejected unless there is sufficient evidence to reject it.

Multiple-choice questions

(7) Let $X_1, \ldots, X_{16}$ be i.i.d. random variables with $X_1 \sim \mathcal{N}(0, 4)$. From a data set we computed the test statistics to be 4. In the context of a right-sided test, let $H_0 : \mu = 2$. If the rejection area is $R = [3, +\infty)$, which one of the following statements is correct?

a. We will commit a Type I error
b. We will commit a Type II error
c. We will not commit a Type II error.
d. If we increase the significance level of the test, then we obtain a lower test power.

(8) For a statistical test of significance level α it holds

a. the rejection area does not depend α
b the rejection area depends on the distribution of the test statistic under the null hypothesis
c. the rejection area shrinks when α is increased
d. rejection at level α implies rejection at level $\alpha/2$

Multiple-choice questions

(9) Suppose we do six independent right-sided tests for testing $H_0 : \mu = 38$, each at the $\alpha = 0.02$ significance level. What is the probability of committing a Type I error and incorrectly rejecting a true null hypothesis in at least two of the six tests?

a. 0.114
b. 0.02
c. 0.994
d. 0.006

(10) Choosing a smaller level of significance results in

a. a lower risk of Type II error and lower power.
b. a lower risk of Type II error and higher power.
c. a higher risk of Type II error and lower power.
d. a higher risk of Type II error and higher power.

Thank you for your attention!