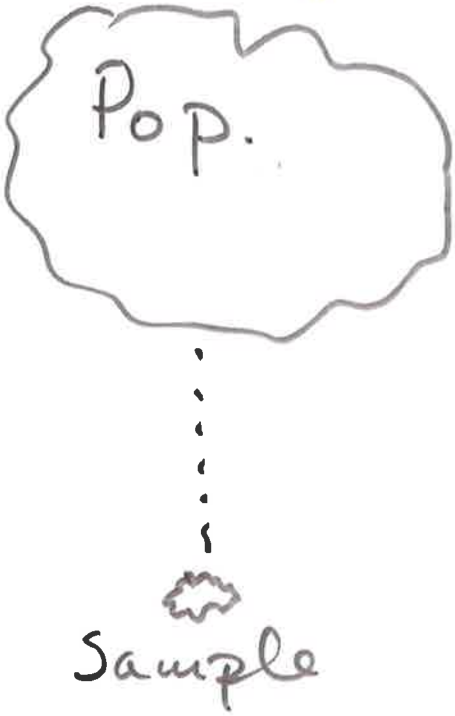
1 Key Concepts and Basic Statistics

1. Key concepts
2. Descriptive statistics
3. Frequency and probability-distributions
4. Hypothesis testing
5. P-values
6. Interval estimation
7. Suggested exercises

① Key concepts



Def. Population: All the units to be studied

Def. Sample: A subset of the population

Some common sampling methods:

* Simple random sampling: All subsets of the same size have the same probability of being drawn

↳ Usually means: All units have an equal probability of being selected

* Systematic sampling: Selecting the units at regular intervals from a "list" of all the population members
  - E.g. an alphabetically ordered list

* Stratified sampling: Classify the population into strata (i.e. categories) and then sample from each stratum (i.e. category)
* Convenience sampling: Select those that are close at hand/ easy to investigate

② Descriptive statistics

Measures of central tendency

→ Intended to indicate where the majority of the values are

→ The sample mean:

$$\bar{x} = \frac{\Sigma x}{n}$$

Example: 8, 10, 9, 7, 8

$$\bar{x} = \frac{8+10+9+7+8}{5} = \underline{\underline{8.4}}$$

→ The median: The middle value that separates the highest values from the lowest

→ The mode : The most frequent value

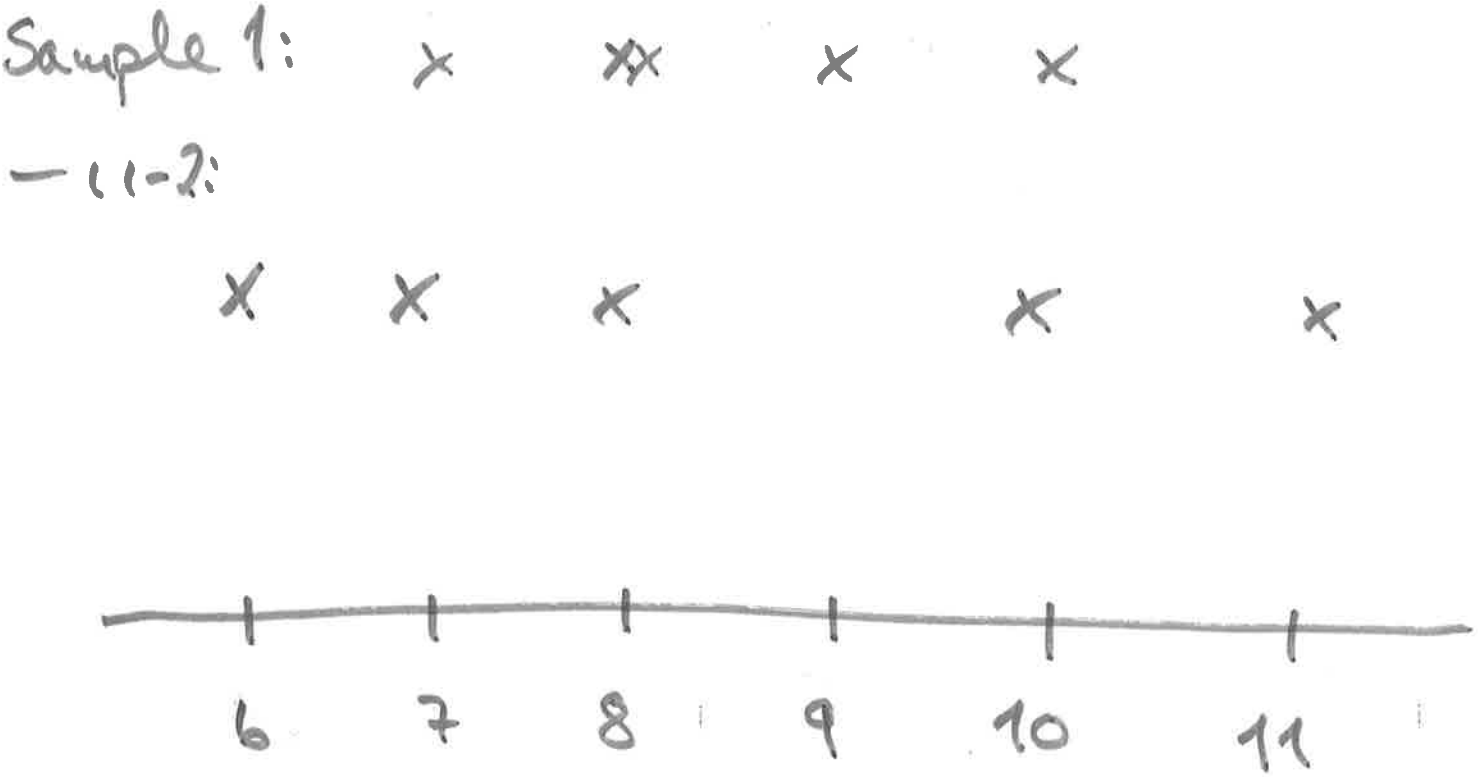
→ The weighted mean

Measures of variability:

→ Intended to indicate the degree of variability or dispersion

Example: Sample 1 = 8, 10, 9, 7, 8 $\bar{x} = 8.4$

— " — 2 = 6, 7, 11, 8, 10 $\bar{x} = 8.4$



→ The sample variance:

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$$

Example:

sample 1 | sample 2

$\bar{x}$ = 8.4 | $\bar{x}$ = 8.4

x	$(x-\bar{x})$	$(x-\bar{x})^2$	x	$(x-\bar{x})$	$(x-\bar{x})^2$
8	−0.4	0.16	6	−2.4	5.76
10	1.6	2.56	7	−1.4	1.96
9	0.6	0.36	11	2.6	6.76
7	−1.4	1.96	8	−0.4	0.16
8	−0.4	0.16	10	1.6	2.56
SUM		5.2			17.2

$$s^2 = \frac{\sum (x-\bar{x})^2}{n-1} = \frac{5.2}{4} = \underline{\underline{1.3}}$$

$$s^2 = \frac{17.2}{4} = \underline{\underline{4.3}}$$

→ The sample standard deviation:

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

Examples: Sample 1 = $s = \sqrt{1.3} = 1.14$

2 = $s = \sqrt{4.3} = 2.07$

→ The sample range: Highest value minus the lowest

Examples: Sample 1: $10 - 7 = \underline{\underline{3}}$

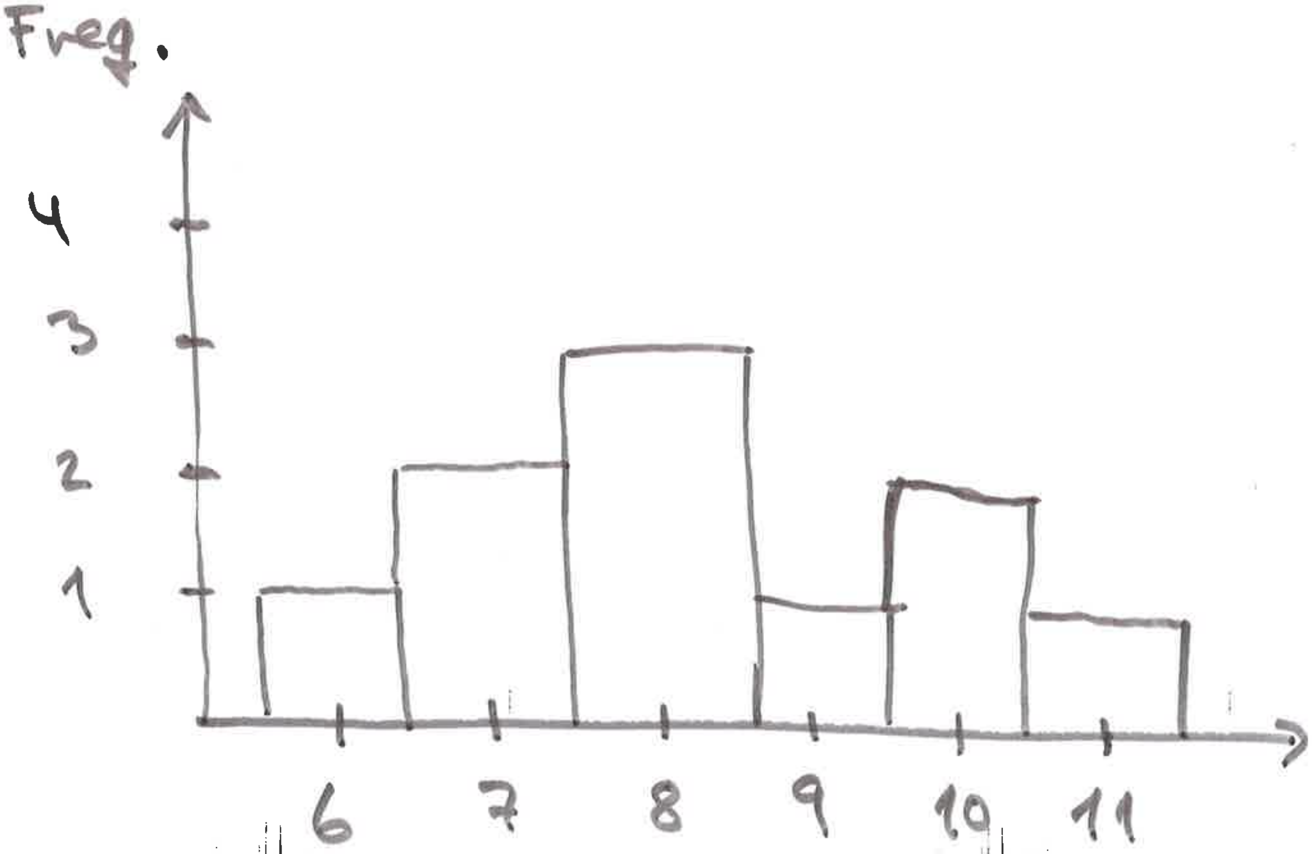
— " — 2: $11 - 6 = \underline{\underline{5}}$

3 Frequency and probability distributions

Def. Frequency distribution: A description of the number of times each value of a variable appears in the sample

Example: A histogram, i.e. a bar graph with no space between the bars

Sample = 8, 10, 9, 7, 8, 6, 7, 11, 8, 10

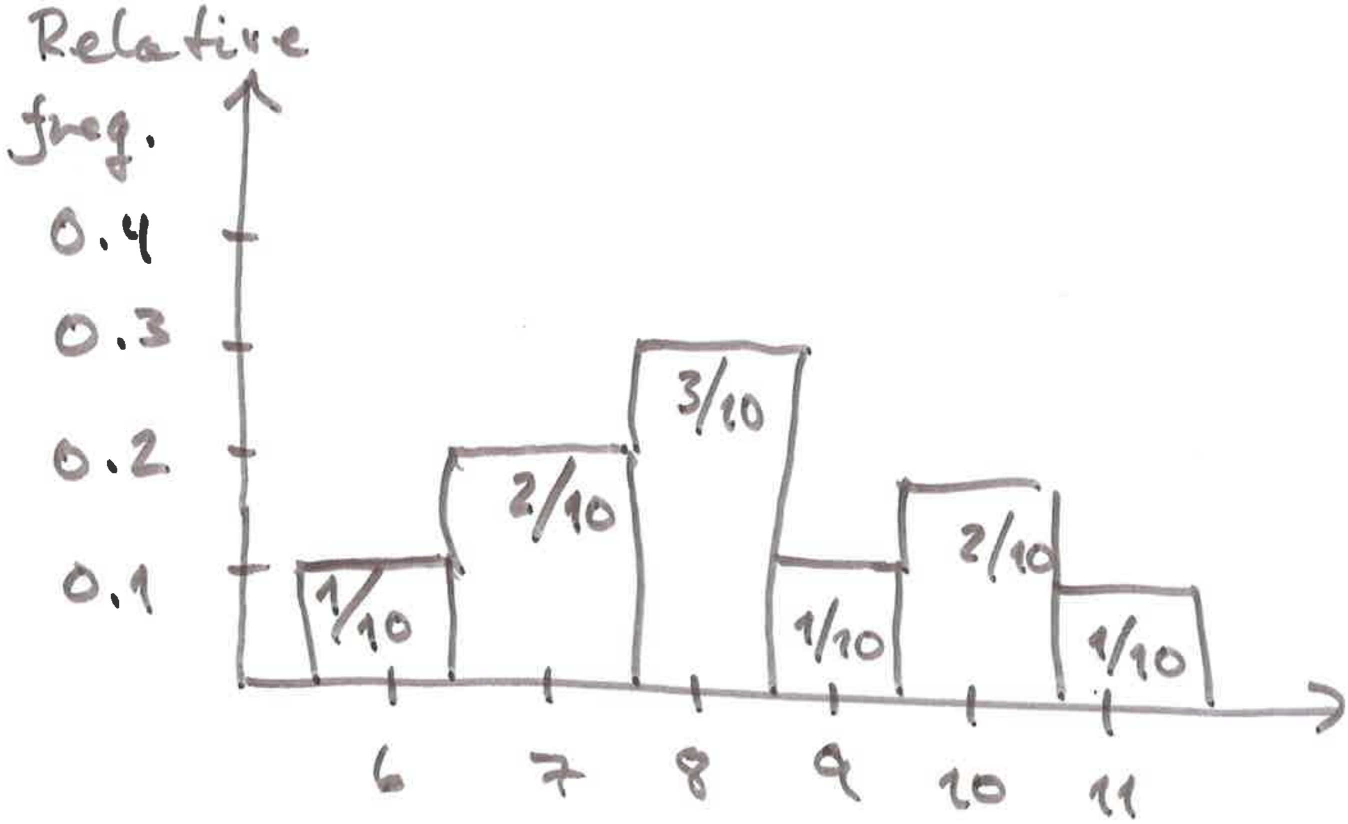


Def. Relative frequency distribution:

A description of the proportion of times each value appears in the sample

Example: A histogram

Sample = Same as previous ($n = 10$)

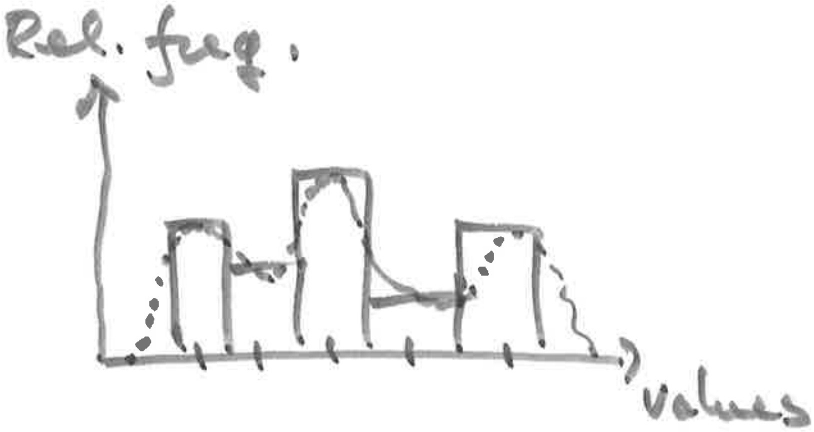


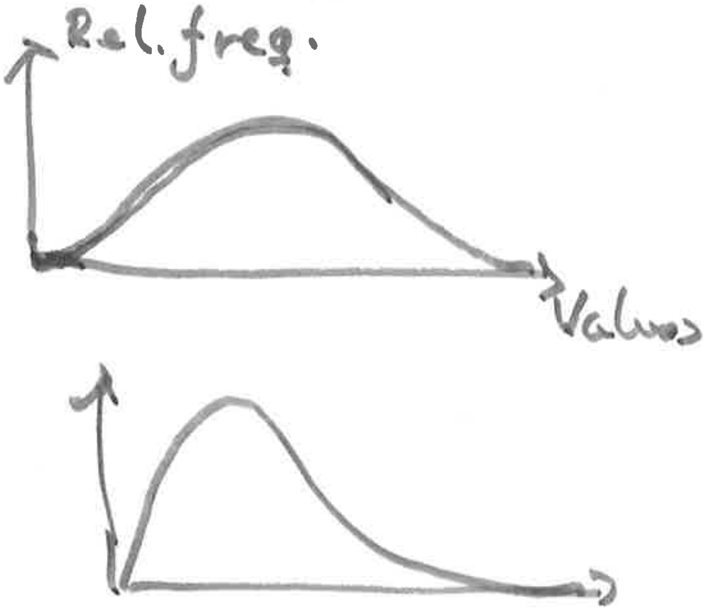
Def. Probability distribution:

A histogram of relative frequencies

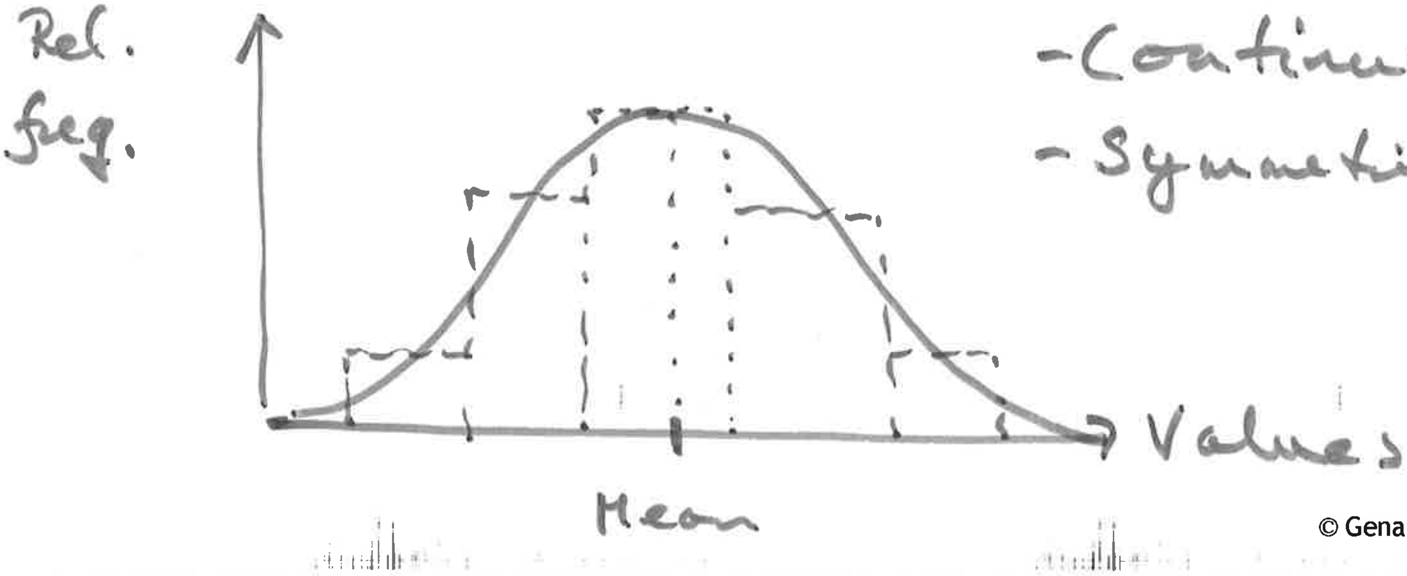
Example: Previous!

Discrete (categorical) probability distributions vs. Continuous probability distributions



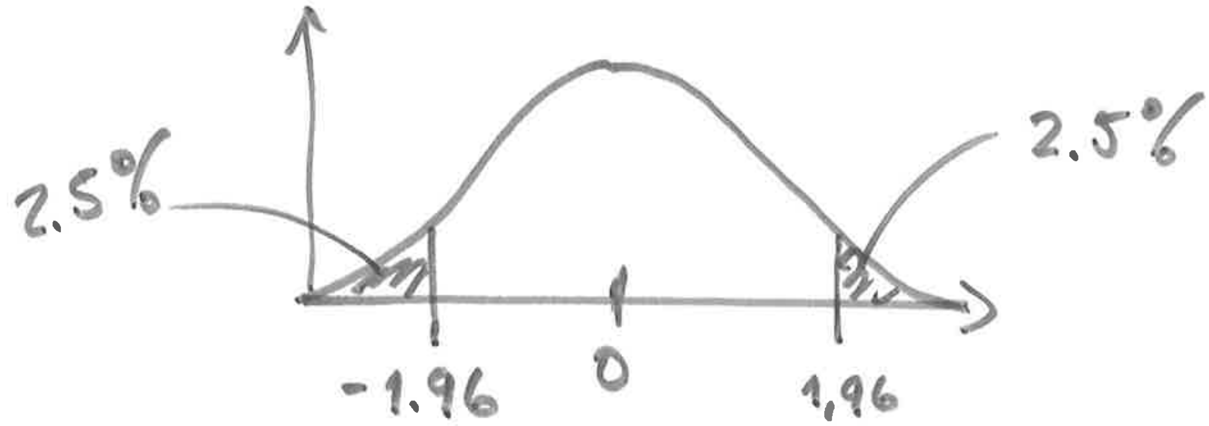


The normal distribution:



- Continuous
- Symmetric

Some properties of the standard normal distribution:



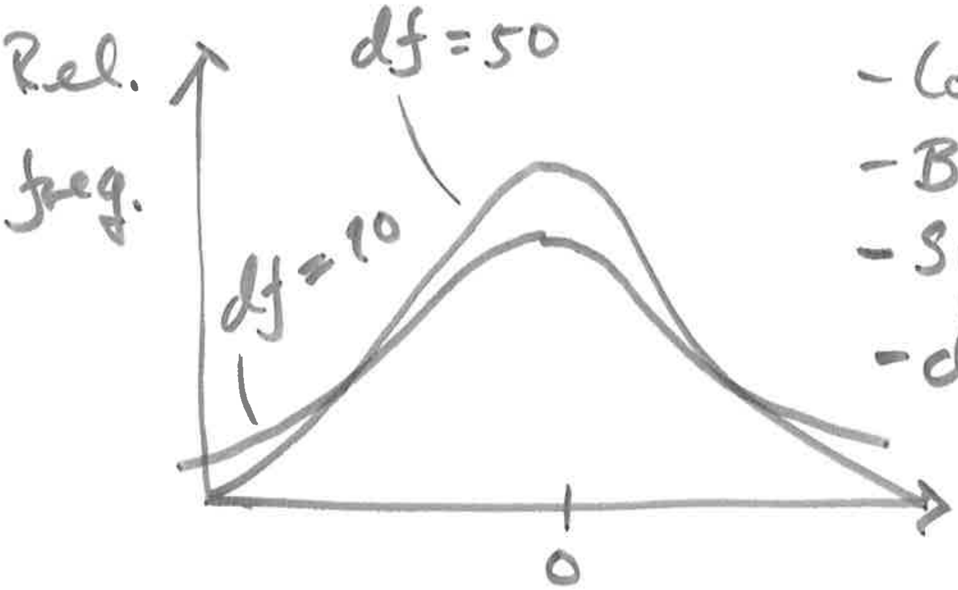
→ mean is 0

→ Area under the curved line is 1 or 100%

→ Symmetry:

- Area to the left of 0 is 0.5 or 50%
- Area to the right of 0 is 0.5
- Area to the left of -1.96 is 2.5%

The t-distribution:



- Continuous
- Bell-shaped
- Symmetric
- df: degrees of freedom

Some properties:

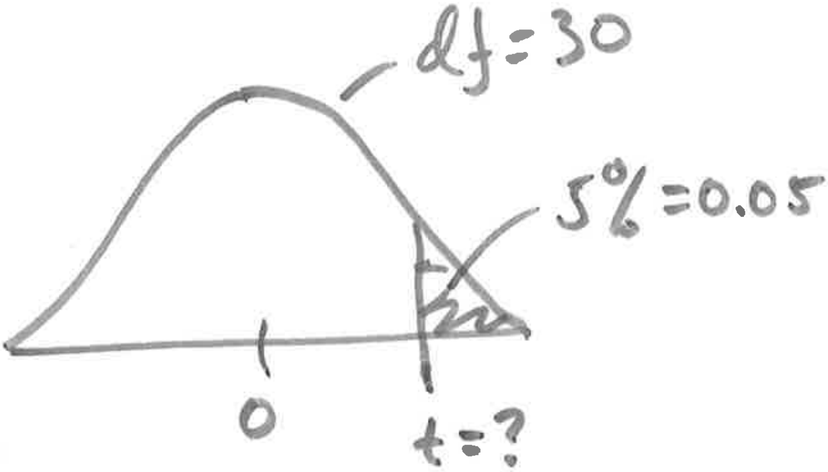
→ There is in fact an infinite number of t-distributions, one for each df

→ df is usually approximately equal to the number of observations

→ The "flatness" of the t-distribution depends on df

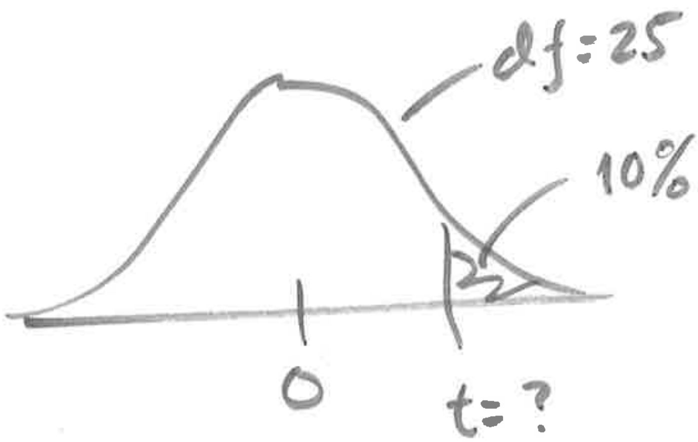
→ When $df \to \infty$, then the t-distribution becomes a standard normal distribution

Examples:



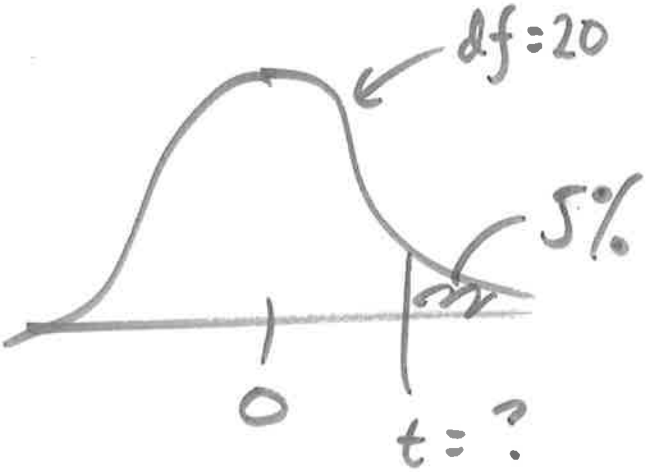
What is t?

$t_{0.05}(30) = 1.697$



What is t?

$t_{0.10}(25) = 1.316$



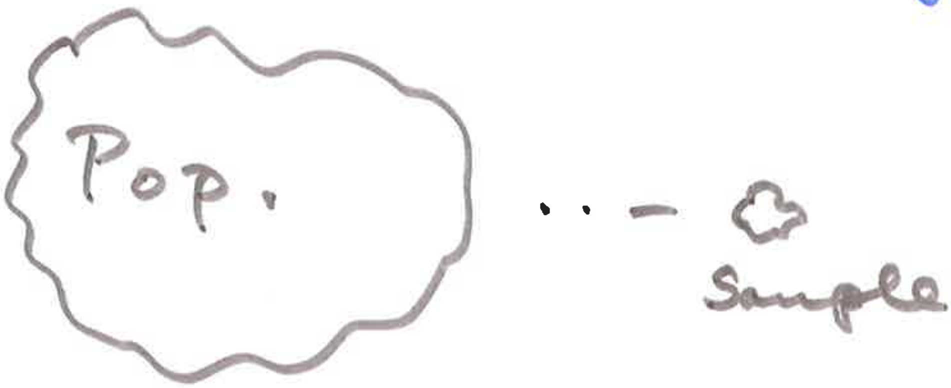
What is t?

$t_{0.05}(20) = 1.725$

Other important probability distributions:

- Chi-squared dist.
- F-distribution

④ Hypothesis testing



Def. Hypothesis: A claim about a population

Def. Null hypothesis (H_0): A claim about the population in terms of an equality

Examples:

→ $\mu = 20$: H_0 (population mean equal to 20)

→ $H_0 : \mu = 9$

→ $H_0 : \mu = 2$

Def. Alternative hypothesis (H_A):

A claim about the population of the "not H_0" type

Examples:

→ $H_A : \mu \neq 20$ $\quad H_A : \mu > 20$ $\quad H_A : \mu < 20$

→ $H_A : \mu \neq 9$ $\quad H_A : \mu > 9$ $\quad H_A : \mu < 9$

→

NOTE: The investigator chooses H_0 and H_A

Some conventions:

→ H_A is usually the claim you would like to test or investigate

→ H_0 is what previous knowledge or findings suggest

→ Moral and/or health-reasons that come into play

Def. Significance level (α): The probability of rejecting H_0 when it is true

NOTE: The investigator chooses α

Examples:

$$\left.\begin{aligned} \alpha &= 0{,}01 = 1\% \\ \alpha &= 0.05 = 5\% \\ \alpha &= 0.10 = 10\% \end{aligned}\right\} \text{most common ones}$$

Def: Test expression (statistic):

A formula whose value indicates whether to keep or reject H_0

Test expressions that are t-distributed typically have the following form:

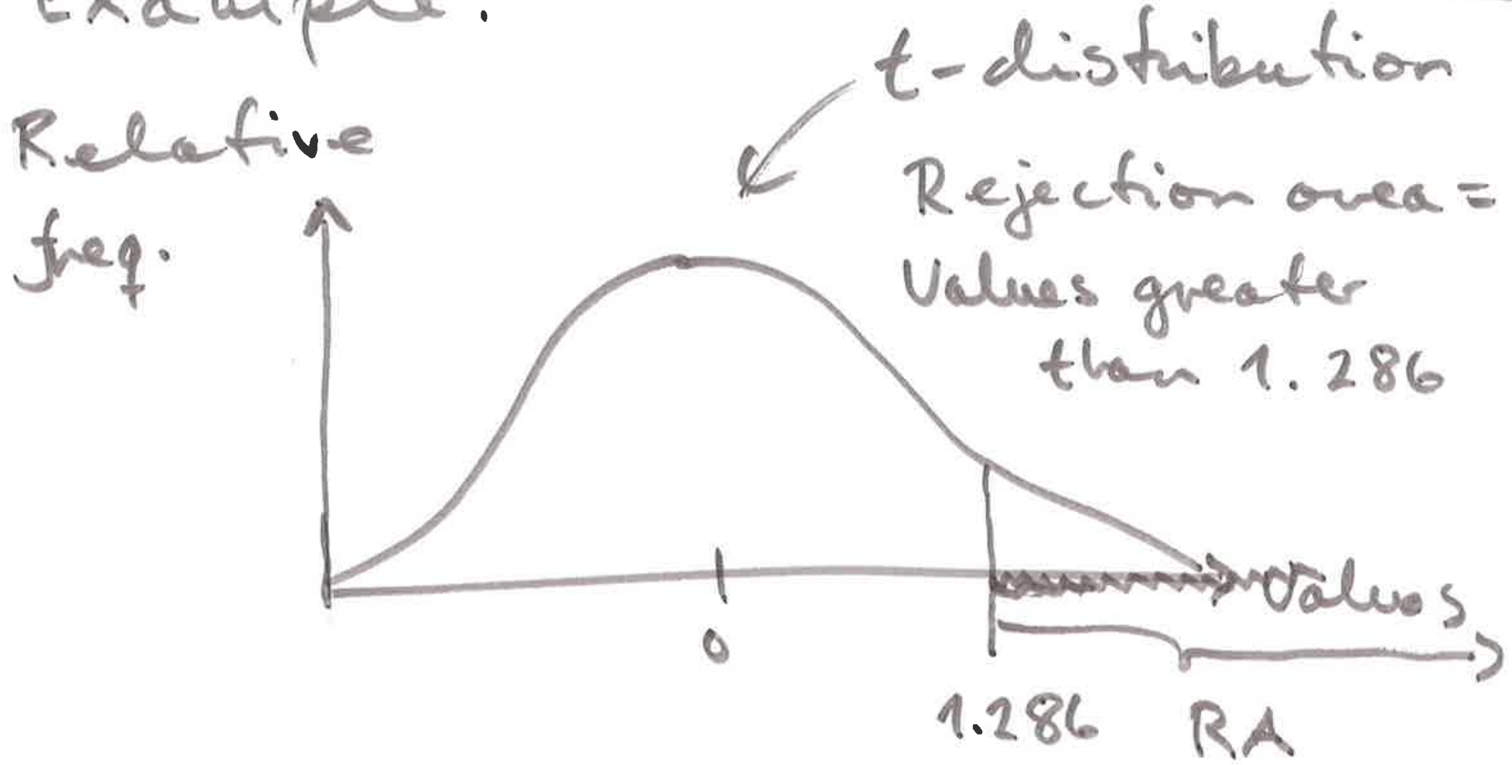
$$\frac{\bar{X} - H_0 \text{ value}}{S}$$

sample estimate

sample standard dev.

Def. Rejection area: The values of the test-expression that makes us reject H_0

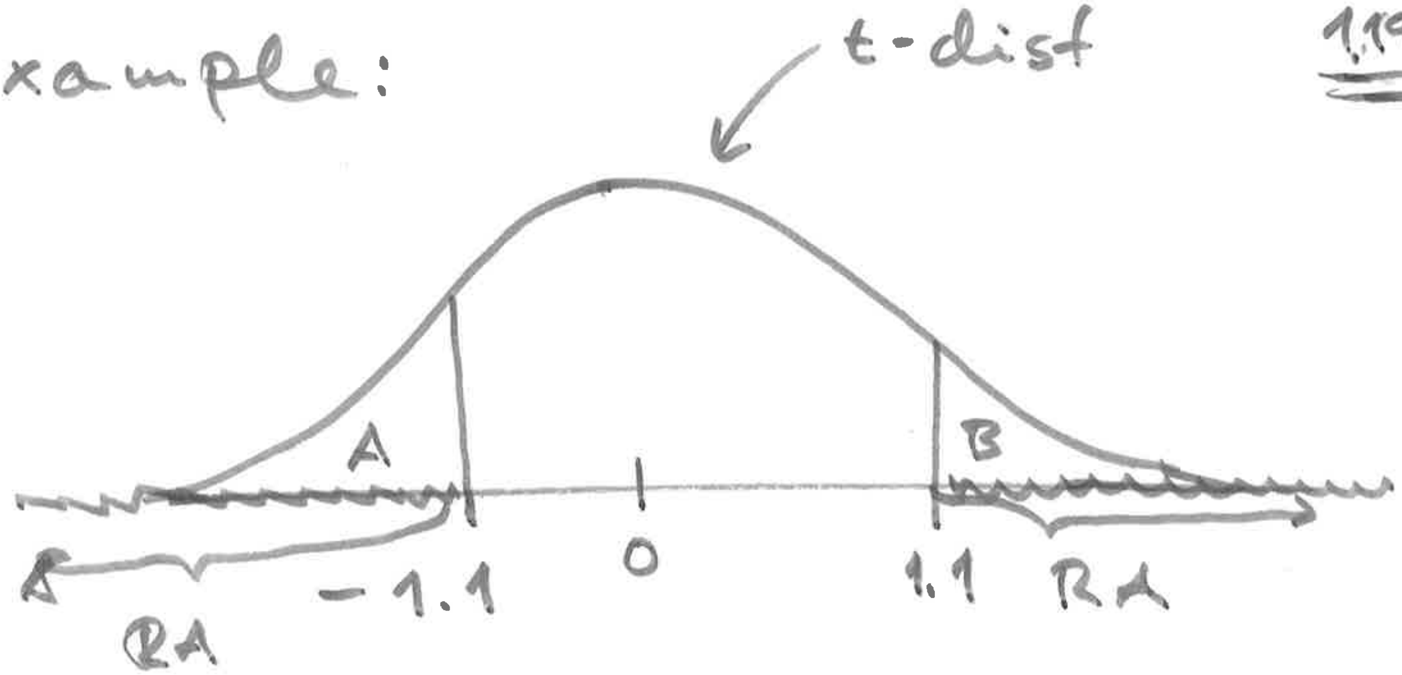
Example:



Def. Critical value(s): The boundary (or boundaries) of the rejection area

Example: Above (i.e. previous example) we have 1 critical value, and it is equal to 1.286

Example:



Rejection area = Values greater than 1.1 and values smaller than -1.1. Critical values = 1.1 and -1.1.

Testing in 4 steps:

Step 1: Choose α, H_0 and H_A

2: Find the critical value(s) and identify the rejection area

3: Compute the value of the test-expression

4: Conclude: Reject H_0 if the test-value lies in the rejection area, otherwise keep H_0

Type 1 and 2 errors:

	H_0 True	H_A true
Reject H_0	Type 1 error	Correct!
Keep H_0	Correct!	Type 2 error

Testing the value of a population mean

Step 1: Choose α, H_0 and H_A

2: Identify the rejection area using a t-distribution with n - 1 df

3: Compute the test-expression

$$\frac{\bar{x} - H_0 \text{ value}}{s/\sqrt{n}}$$

4: Conclude: Reject H_0 if the test-value lies in the rejection area, otherwise keep H_0

Example: Travel time (in minutes)

Claim: Average travel time is greater than 20 minutes

Sample: 20, 15, 17, 55, 20, 10, 20, 90, 15
18, 40

$n = 11$ $\quad \bar{x} = 29.09$ $\quad s = 24.0394$

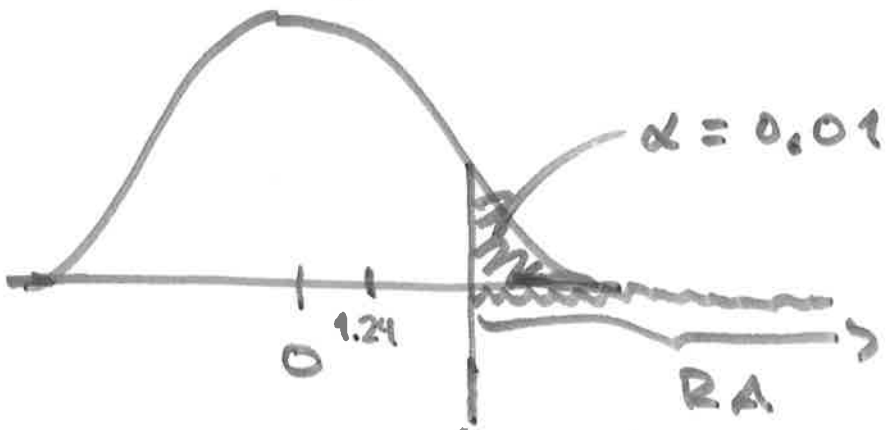
Step 1: $\alpha = 1\%$ $\quad H_0: \mu = 20$

$H_A: \mu > 20$

2: Identify the rejection area:

$df = n - 1 = 11 - 1 = 10$

$\alpha = 0.01 = 1\%$



$t_{0.01}(10) = 2.764$

3: Compute the test-value:

$$\frac{\bar{x} - H_0 \text{ value}}{s/\sqrt{n}} = 1.2417$$

29.09 → $\bar{x}$; 20 → H_0 value; 24.0394 → s; 11 → n

4: Conclusion: We keep H_0. That is,

We have found support in favour of the claim that the average travel time is 20 minutes

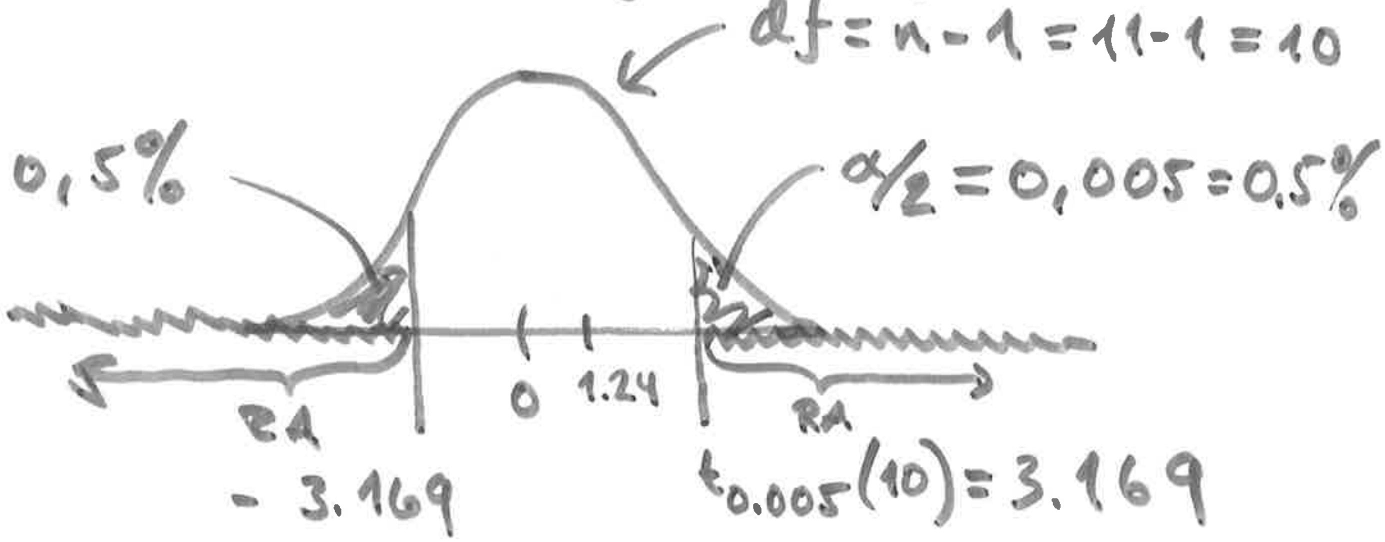
Example:

- Claim: Average travel time is different from 20 minutes
- Sample: Same as previous

Step 1: $\alpha = 0.01 = 1\%$ $\qquad H_0: \mu = 20$

$H_A: \mu \neq 20$

2: Identify the rejection area:



RA: Values higher than 3.169 and values lower than -3.169

3: Value of test-expression:

$$\frac{\bar{X} - H_0 \text{ value}}{s/\sqrt{n}} = 1.2417$$

29.09

20

24.0394

11

4: Conclusion: We keep H_0

(5) P-values

Intuition: A number between 0 and 1 that summarises the information from a hypothesis test

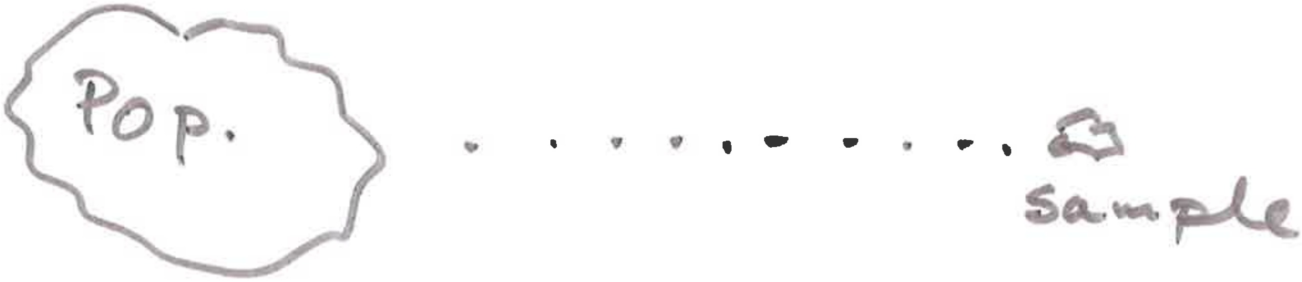
Def. P-value: The smallest significance level at which the null hypothesis can be rejected

In practice: Reject H_0 if the p-value is smaller than α

The p-value game & dance!

NOTE: In order to compute a p-value "by hand", H_0 and H_A must already have been defined, and the test-value must already have been computed

(b) Interval estimation



$$\left.\begin{matrix}\mu \\ \sigma^2 \\ \sigma\end{matrix}\right\} \text{ Sample counterparts: } \begin{matrix}\bar{x} \\ s^2 \\ s\end{matrix}$$

<u>Def. $(1-\alpha)\cdot 100\%$ confidence interval</u>: An interval that contains the population value of interest (e.g. μ, σ^2, σ, etc.) with $(1-\alpha)\cdot 100\%$ degree of certainty

Example: 90% confidence interval for the mean time it takes from you wake up until you leave the house (Q2)

Sample = 110, 45, 75, 90, 70, 90, 40, 30, 45, 60

$n = 10 \quad \bar{x} = 65.5 \quad s_x^2 = 674.72 \quad s_x = 25.98$

$$L = \underbrace{\bar{x}}_{65.5} - \underbrace{t_{\underbrace{\alpha/2}_{0.05}}(\underbrace{df}_{10-1=9})}_{1.833} \cdot s_x/\sqrt{\underbrace{n}_{10}} = \underline{50.43}$$

$$U = \bar{x} + t_{\alpha/2}(df) \cdot s_x/\sqrt{n} = \underline{80.57}$$

Interpretation: The interval [50.43, 80,57] contains the population mean with 90% degree of certainty

Example: Suppose [6, 11] is a 95% confidence interval for μ, then μ lies between 6 and 11 with 95% degree of certainty

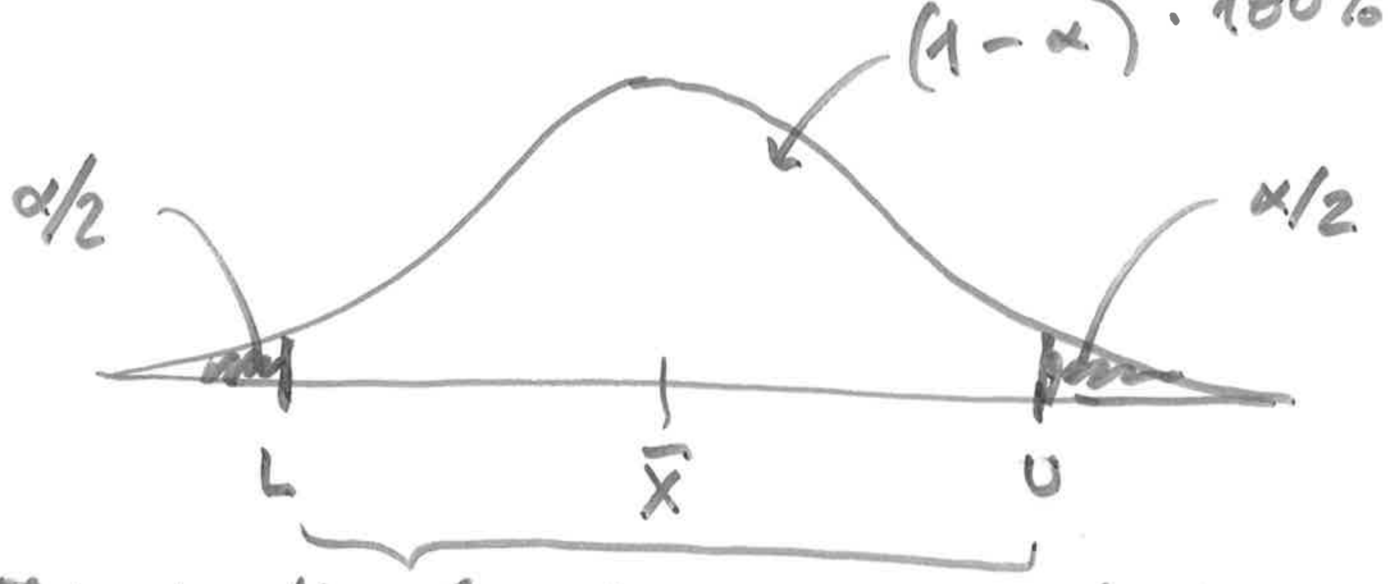
Computation of confidence interval for μ

Lower bound $= \bar{x} - t_{\alpha/2}(df) \cdot s_x/\sqrt{n}$

$n-1$

Upper bound $= \bar{x} + t_{\alpha/2}(df) \cdot s_x/\sqrt{n}$

Intuition (probabilistic):



This is the $(1-\alpha)\cdot 100\%$ confidence interval

7 Suggested exercises

<u>Exercise set 1:</u> 1(a)–g), 4(a)iv), 4b)iv), 4c)i), 5a)i), 5a)iii), 5b)i), 5b)iii), 5c)i), 7