

Crash Course in Statistics

ZNZ 2026

IV

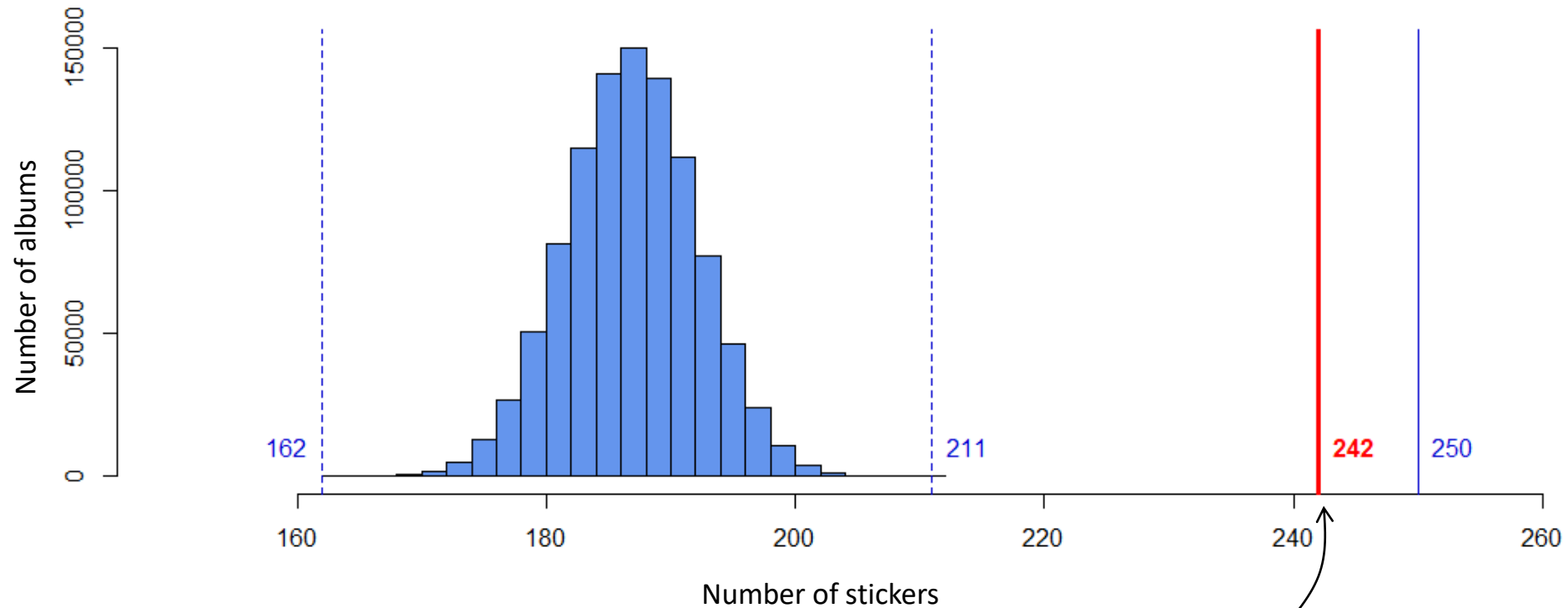
Christoph Luchsinger and Zofia Baranczuk

Based on Script by Daniel J. Stekhoven

Tests and linear models

- P-values and hypotheses
- One-sided versus two-sided testing
- Performance of tests
- Z-test, t-test
- Wilcoxon-test, sign-test
- Regression

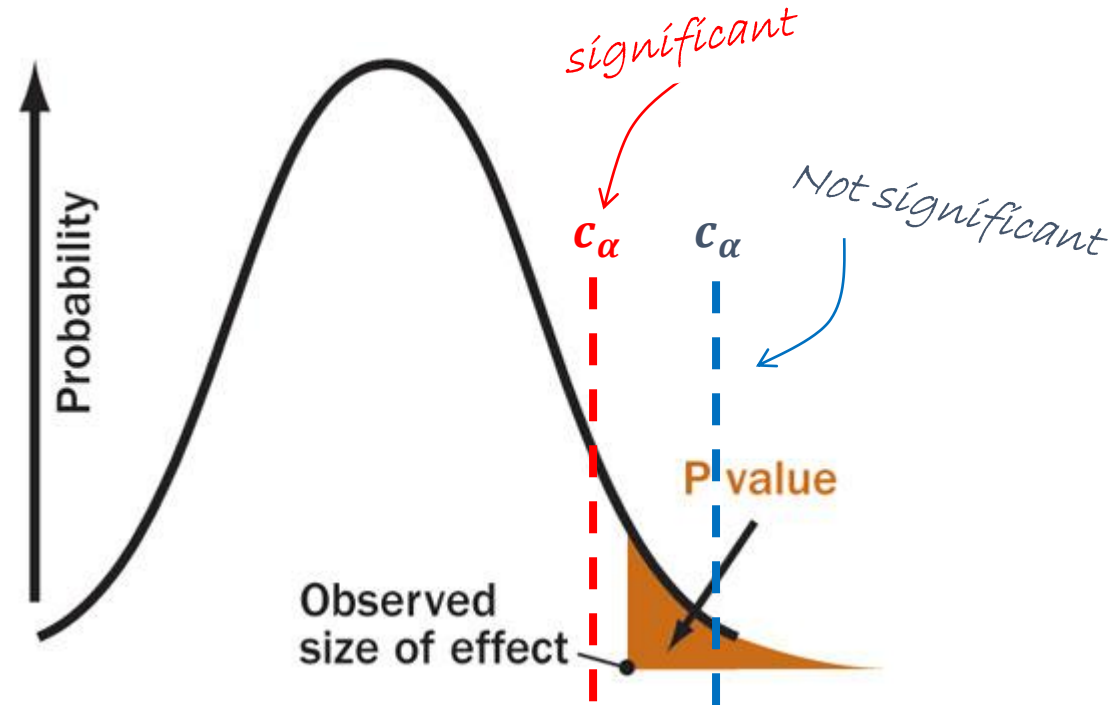
What is a p-value (now with $n=250$)?



What is the probability of such a value **or a more extreme one**, if the null hypothesis were true?

i.e. the summed probability of getting 242, 243, 244, ..., 250 stickers in one go!

What is a p-value? You need data to compute it! Not a beautiful number, not 5% or so, but $3.87e-17$




Smallest level of significance for which H_0 will just be rejected

Gut feeling and hypotheses

- Complete box → less double stickers
- Buy blisters all over town → many doubles

«Null», because no system involved



- **Null hypothesis:**
 - Stickers are filled **randomly** into the boxes
- **Alternative hypothesis:**
 - Stickers are filled **systematically** into the boxes such that there are less doubles

How can we decide between these two?

Null Hypothesis H_0 versus Alternative Hypothesis H_A

- Null hypothesis

- Nothing happened
- No difference
- No effect
- Symmetry, fair, boring

*As long as we haven't shown anything,
we are in the «null hypothesis world»*

- Alternative Hypothesis

- Something happened
- exciting
- Difference larger than something
- Effect stronger than something

*As soon as we have enough evidence,
we can enter the «alternative world»*

Popper (unquestioned in natural science):

- “All stones fall down” (simple version of gravity theory)
 - Can we *prove/verify* this? Hey: it’s (most probably) true, isn’t it?
- “All stones fall upwards”
 - Can we *prove* this? No, it is wrong; try it out *once!* But we therefore can conclude: it’s wrong!
- Conclusions:
 - We can only *confirm (bestätigen; not prove)* a (possibly true) theory through experiments/data
 - We can *falsify* a wrong theory

Null Hypothesis H_0 versus Alternative Hypothesis H_A

- Sometimes we read:

*«Since the p -value is not significant
we accept the null hypothesis that there is no difference.»*

- The null hypothesis **cannot** be accepted, we live in the null hypothesis world¹!
- The null hypothesis can only be **rejected**!
- We can accept the alternative hypothesis.

¹until enough evidence is present to prove that we have a different area code!

Null Hypothesis H_0 versus Alternative Hypothesis H_A

- Typical H_0 's:
 - The location of two samples is identical.
 - The spread of two samples is identical.
 - The distribution (basically the combination of the two above) of two samples is identical.
- Instead of two samples we can have one sample (compared to a reference) or more than two samples

Two sided versus one sided

- Typical H_A 's:
 - The location of sample A is **different** from sample B (two-sided test)
 - The location of sample A is **larger/smaller** than sample B (one-sided test)

- Why this distinction?

Two sided versus one sided



One sided

- Blind on one side
- Very sharp on the other side (statistical power)

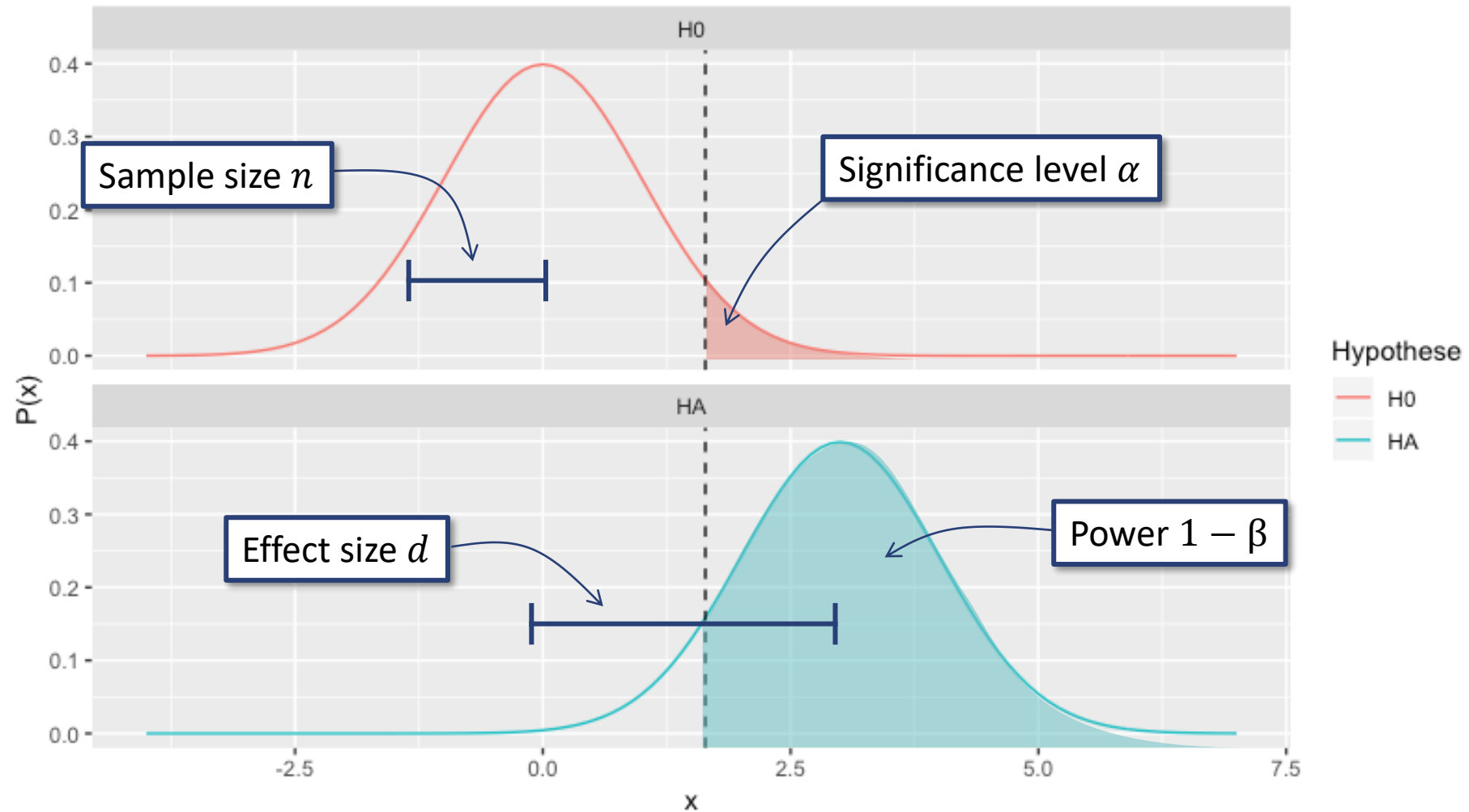


Two sided

- Sees on both sides
- Not so sharp on both sides (less power)

Statistical power: Chance to **detect** a difference (reject H_0) if there is one

Statistical power – one of 3+2 elements in a test



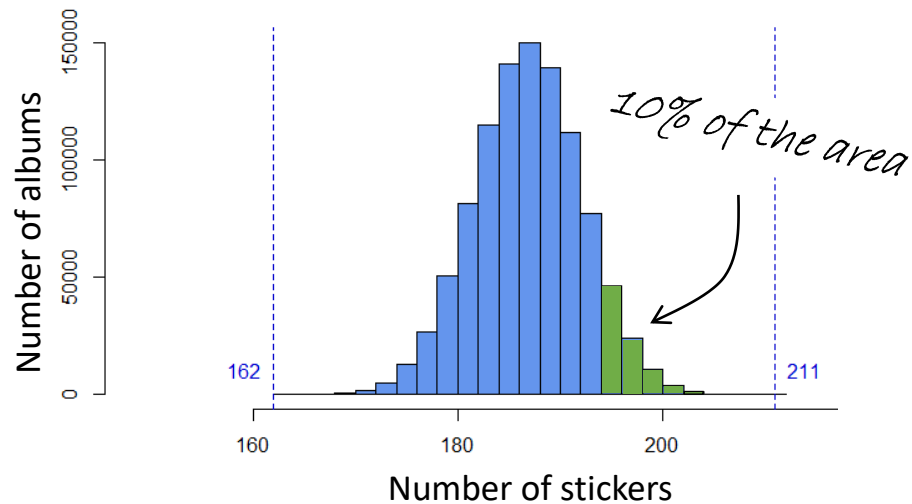
The 3+2 elements in a test...

Two sided versus one sided (data given)

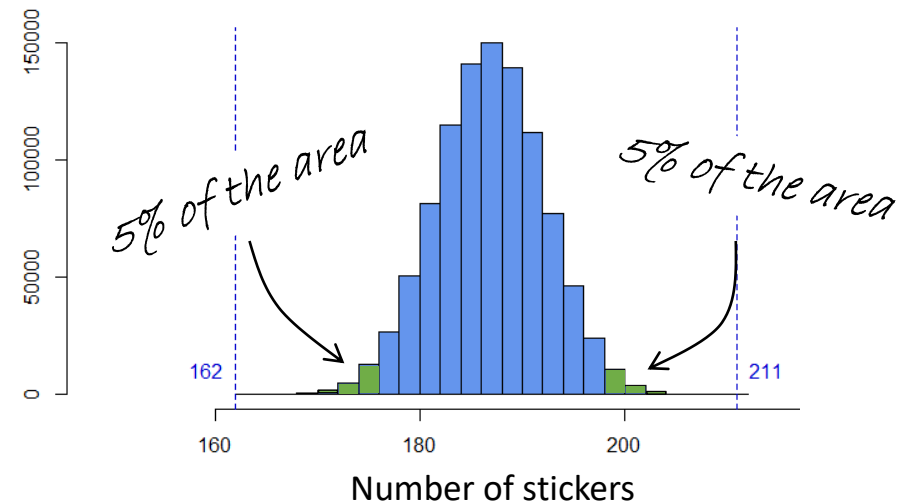
Intuition:

- If for a given sample the two-sided t-test has a p-value of y
- The p-value for this sample using a one-sided t-test is $\frac{y}{2}$

Two sided versus one sided (α given)



We require only 194 stickers for a significant result



Here, we require 198 stickers for a significant result

Paired tests

- If the two samples are labelled such that for each observation in the first sample (or control) x_i , there is a unique observation in the second sample y_i , we can use a paired test with $z_i := x_i - y_i$
- Paired tests have more statistical power
- Example:
 - ocular pressure after surgery
 - for each patient measure pressure left and right
 - test whether the pressure is higher in operated eye

Performance of tests

- Sample → Test →

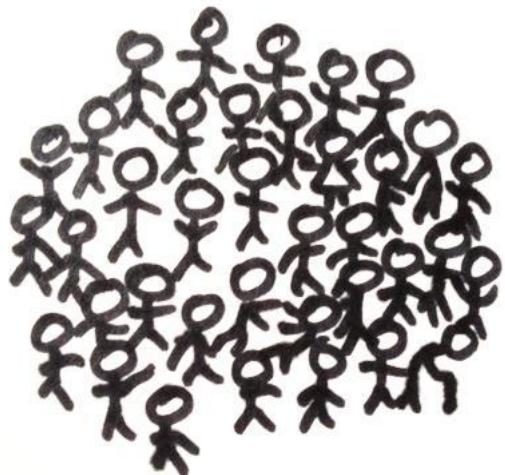


Positive (reject H_0)

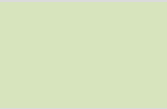


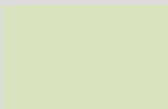
Negative (do not reject H_0)

- Population → Truth → Correct

Wrong



Performance of tests

		Truth/Reality	
		Correct (C)	Wrong (W)
Test outcome	Total (n)		
	Reject H_0 (P, positive)		
	Do not reject H_0 (N, negative)		

Performance of tests

		Truth/Reality			
		Correct (C)	Wrong (W)	Prevalence C/n	
Test outcome	Reject H_0 (P, positive)	True Positive (TP)	False Positive (FP, Type I error)	Precision TP/P	False Discovery Rate FP/P
	Do not reject H_0 (N, negative)	False Negative (FN, Type II error)	True Negative (TN)	False Omission Rate FN/N	Negative predictive value TN/N
		Sensitivity TP/C	False Positive Rate FP/W	Accuracy $(TP+TN)/n$	
		False Negative Rate FN/C	Specificity TN/W		

Which is worse? Type I (FP) or type II (FN)?

Fire alarm		Is there a fire?	
		Yes	No
Is there an alarm?	Alarm! (positive)	TP Fire, alarm goes off	FP No fire, alarm goes off
	No alarm (negative)	FN Fire, no alarm!	TN No fire, no alarm

- Type II error **much** worse!
- Reduce by making the test more lenient
(the test rejects earlier → more type I errors)



Which is worse? Type I or type II?

Spam filter		Is it spam?	
		Yes	No
Does the filter remove it?	Remove (positive)	TP Spam, is removed	FP No spam, is removed!
	Leave (negative)	FN Spam, is not removed	TN No spam, is not removed

- Type I error **much** more annoying!
- Reduce by making the test more stringent (the test rejects later → more type II errors)



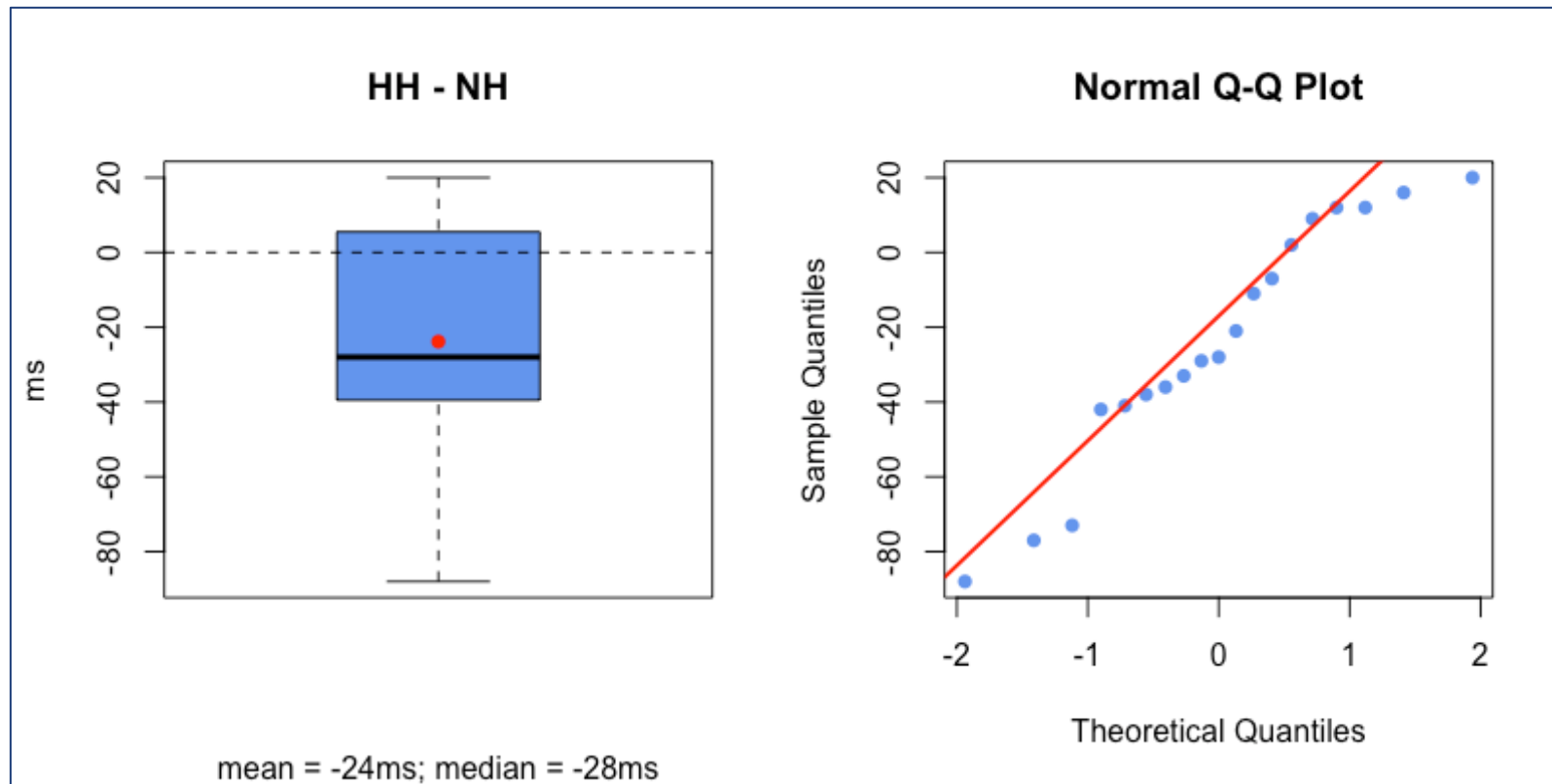
Reaction time

Are we reacting faster with our main hand (left or right) compared to our support hand (off hand/Nebenhand)?

- Experiment:
 - **Population:** all ETH students
 - **Sample:** participants of my lecture in 2017
- Measurement:
 - Reaction test on the internet
- Measure each side 3 times and average
- Randomize order of hands using birthday
- Get difference from the two hands, $z_i := x_i - y_i$

Result

- Asked 50 of the class
- Got: 19
- The main hand is on average 24ms faster as a mean, and 28ms as a median



Sample versus Population

- In that sample the main hand was 24ms faster
- Is that sufficient to say that the main hand is generally faster in the population?
- To get an answer we need one of these:
 - **z-test**, we will need: $qnorm(0.95)=1.644854$, $qnorm(0.975)=1.959964$
 - **t-test**, we will need: $qt(0.95, 18)=1.734064$, $qt(0.975,18)= 2.100922$
 - Wilcoxon-test (Mann-Whitney-U-test)
 - Sign-test
- from now on: iid=independent and identically distributed

z-Test (σ_X known)

1. **Model:** X_i continuous measurements; X_1, X_2, \dots, X_n i.i.d., $\mathcal{N}(\mu, \sigma_X^2)$, σ_X known

2. **Null hypothesis:** $\mathcal{H}_0: \mu = \mu_0$
Alternative: $\mathcal{H}_A: \mu \neq \mu_0$ (or $<$ or $>$)

3. **Test statistic:**

$$Z = \frac{\bar{X}_n - \mu_0}{\sigma_{\bar{X}_n}} = \frac{\bar{X}_n - \mu_0}{\frac{\sigma_X}{\sqrt{n}}} = \frac{\text{observed} - \text{expected}}{\text{standard error}}$$

distribution under $\mathcal{H}_0: Z \sim \mathcal{N}(0,1)$

4. **Significance level:** α

5. **Critical region** for the test statistic:

$$K = (-\infty, -\Phi^{-1}(1 - \alpha/2)] \cup [\Phi^{-1}(1 - \alpha/2), \infty)$$

$$K = (-\infty, -\Phi^{-1}(1 - \alpha)] \text{ with } \mathcal{H}_A: \mu < \mu_0$$

$$K = [\Phi^{-1}(1 - \alpha), \infty) \text{ with } \mathcal{H}_A: \mu > \mu_0$$

6. **Decision:** is the observed value of z of the test statistic in K

Problem: σ_X usually not known in real life

- Estimate the variance:

$$\hat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$$

- New test statistic:

$$T = \frac{\bar{X}_n - \mu_0}{\frac{\hat{\sigma}_X}{\sqrt{n}}} \quad Z = \frac{\bar{X}_n - \mu_0}{\frac{\sigma_X}{\sqrt{n}}}$$

- Distribution of T , under \mathcal{H}_0 :

$$T \sim t_{n-1}$$

$$Z \sim \mathcal{N}(0,1)$$

«Student's» t - distribution

- $X_1, X_2, \dots, X_n \sim \mathcal{N}(\mu, \sigma_X^2)$ and independent
- $\hat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2$ the estimated variance
- The test statistic

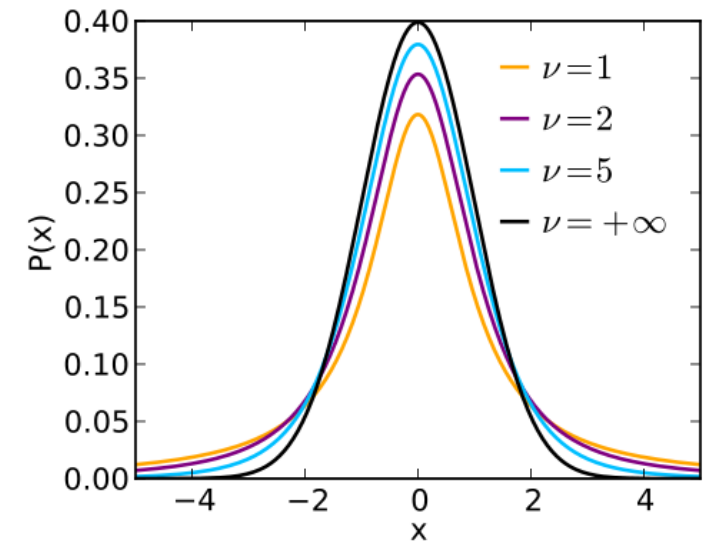
$$T = \frac{\bar{X}_n - \mu_0}{\frac{\hat{\sigma}_X}{\sqrt{n}}} \sim t_{n-1}$$

has a « t -distribution with $n-1$ degrees of freedom»

- If $n = \infty$, it holds that $t_\infty = \mathcal{N}(0,1)$



William
«Student»
Sealy
Gosset



t-test (σ_X not known)

1. **Model:** X_i continuous measurements; X_1, X_2, \dots, X_n i.i.d., $\mathcal{N}(\mu, \sigma_X^2)$, σ_X will be estimated using $\hat{\sigma}_X$

2. **Nullhypothesis:** $\mathcal{H}_0: \mu = \mu_0$
Alternative: $\mathcal{H}_A: \mu \neq \mu_0$ (or $<$ or $>$)

3. **Test statistic:**

$$T = \frac{\bar{X}_n - \mu_0}{\hat{\sigma}_{\bar{X}_n}} = \frac{\bar{X}_n - \mu_0}{\frac{\hat{\sigma}_X}{\sqrt{n}}} = \frac{\text{observed} - \text{expected}}{\text{estimated standard error}}$$

distribution under \mathcal{H}_0 : $T \sim t_{n-1}$

4. **Significance level:** α

5. **Critical region** for the test statistic:

$$K = (-\infty, -t_{n-1; 1-\frac{\alpha}{2}}] \cup [t_{n-1; 1-\frac{\alpha}{2}}, \infty)$$

$$K = (-\infty, -t_{n-1; 1-\alpha}] \text{ with } \mathcal{H}_A: \mu < \mu_0$$

$$K = [t_{n-1; 1-\alpha}, \infty) \text{ with } \mathcal{H}_A: \mu > \mu_0$$

6. **Decision:** is the observed value t of the test statistic in K

Example reaction time

1. **Model:** X_i difference in reaction time between main and support hand of student i

2. **Nullhypothesis:** $\mathcal{H}_0: \mu = 0$ ms
Alternative: $\mathcal{H}_A: \mu \neq 0$ ms

3. **Test statistic:**

$$T = \frac{\sqrt{n}(\bar{X}_n - \mu_0)}{\hat{\sigma}_X} \Rightarrow t = \frac{\sqrt{19}(-24.0 - 0)}{32.27} \approx -3.23$$

distribution under $\mathcal{H}_0: T \sim t_{18}$

4. **Significance level:** $\alpha = 0.05$

5. **Critical region:**

$$K = (-\infty, -t_{18;0.975}] \cup [t_{18;0.975}, \infty) = (-\infty, -2.10] \cup [2.10, \infty)$$

6. **Decision:** $t \in K \Rightarrow \mathcal{H}_0$ can be rejected

Sign test (= hidden binomial test)

- Everything is the same as with the t -test, except...

μ is not the mean, it is the **median**

Test statistic V : number of X_i 's where $X_i > \mu_0$

under $\mathcal{H}_0: V \sim \text{Bin}(n, p_0), p_0 = 0.5$

«Is the median of a sample equal to a specific μ_0 ?»

Sign test (= hidden binomial test)

- Assume: $\mathcal{H}_0: \mu = \mu_0 = 10, \mathcal{H}_A: \mu \neq 10$
- Observed: $x_1 = 13, x_2 = 9, x_3 = 17, x_4 = 8, x_5 = 14$
- Signs of $x_i - \mu_0$: +, -, +, -, +
- Perform binomial test

$$\mathcal{H}_0: p_0 = 0.5, \mathcal{H}_A: p_0 \neq 0.5$$
$$n = 5, v = 3$$

- Reject \mathcal{H}_0 if the binomial test rejects it. Above, we stay with \mathcal{H}_0 !



No distributional assumptions



Less power

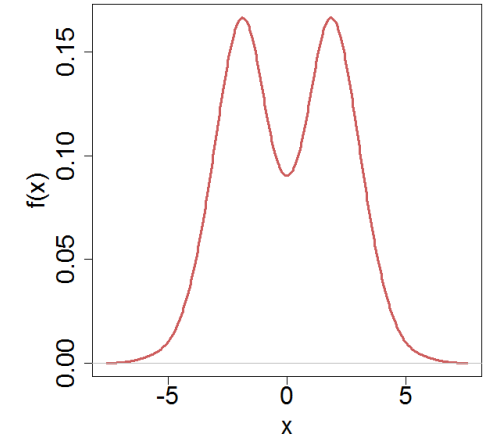
Wilcoxon-test (*Wilcoxon signed rank test*)

- Mixture of sign- and t-Test
- Assumption: $X_i \sim \mathcal{F}$ i.i.d., \mathcal{F} is symmetric
- Test median μ with $\mathcal{H}_0: \mu = \mu_0$ (one- or twosided)

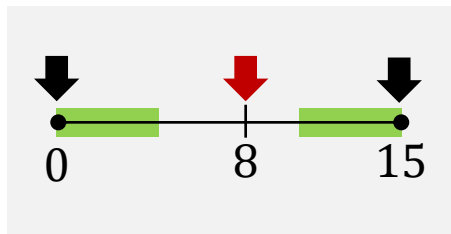
i	x_i	$sign$	abs
1	-1.9	-1	1.9
2	0.2	1	0.2
3	2.7	1	2.7
4	-4.1	-1	4.1
5	3.9	1	3.9



i	x_i	$sign$	abs	R_i	$sign R_i$
2	0.2	1	0.2	1	1
1	-1.9	-1	1.9	2	-2
3	2.7	1	2.7	3	3
5	3.9	1	3.9	4	4
4	-4.1	-1	4.1	5	-5



Rank sum:
 $W = 1 + 3 + 4 = 8$



Overview of tests

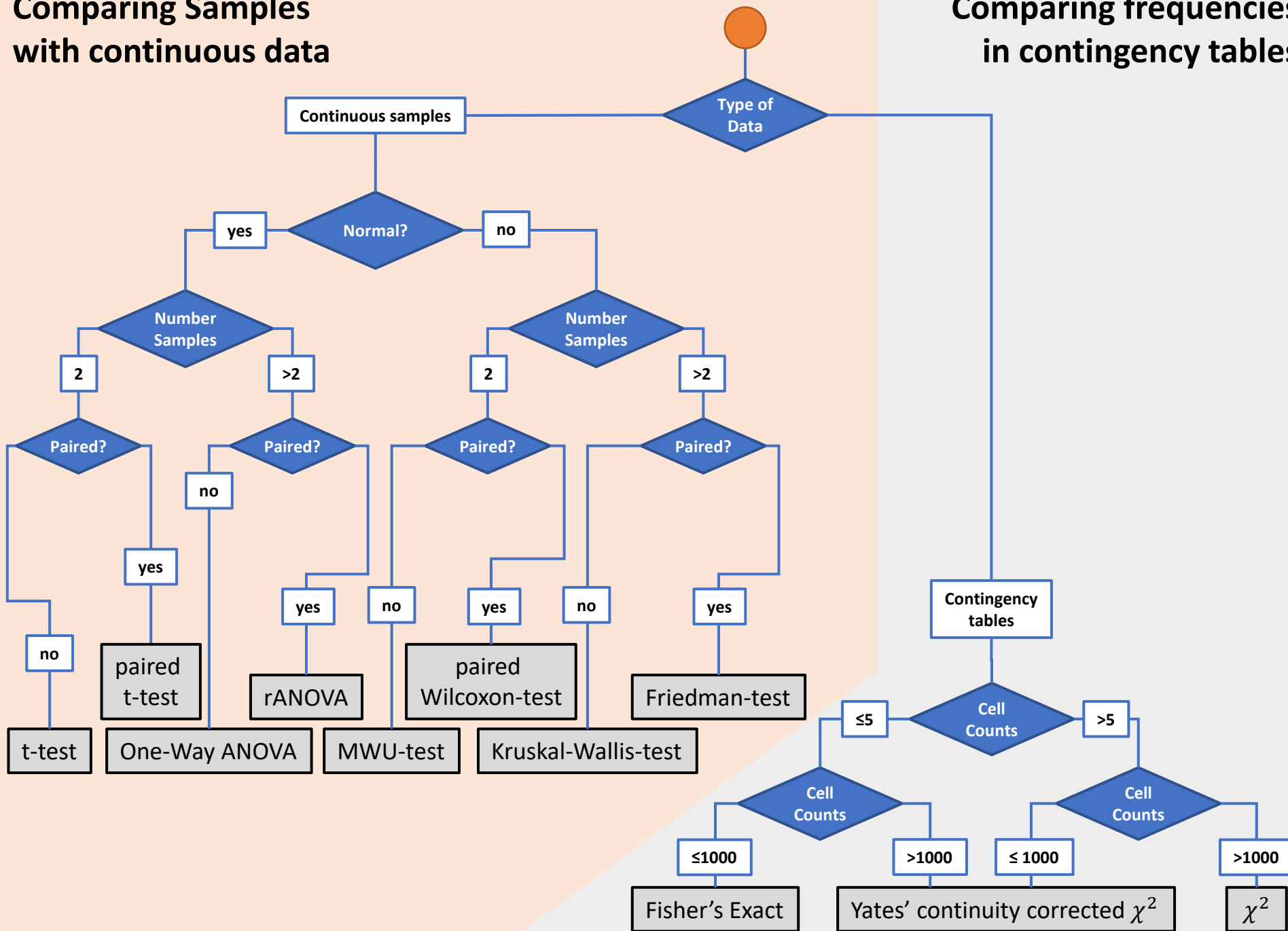
	Assumptions				n_{min} with $\alpha = 0.05$	Power for example
	σ_X known	$X_i \sim \mathcal{N}$	symmetric	i.i.d.		
<i>z</i>-test	•	•	•	•	1	89%
<i>t</i>-test		•	•	•	2	79%
Wilcoxon-test			•	•	6	79%
Sign-test				•	5	48%

Example for computation of power

- $X_i \sim \mathcal{N}(\mu, \sigma^2), n = 10$
- $\mathcal{H}_0: \mu = 0; \mathcal{H}_A: \mu \neq 0; \alpha = 0.05$
- Concrete alternative for the computation of power: $X_i \sim \mathcal{N}(1,1)$

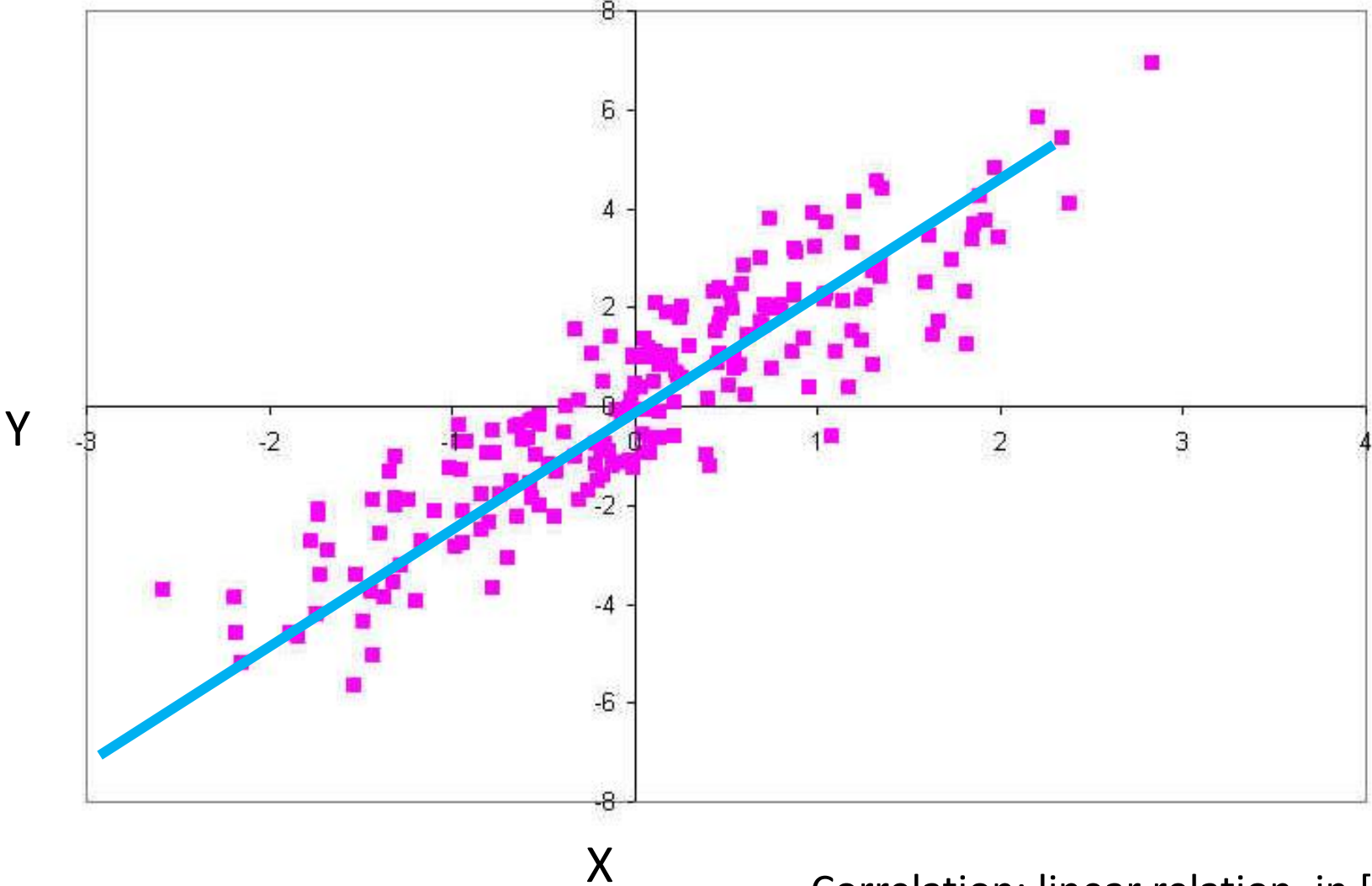
Comparing Samples with continuous data

Comparing frequencies in contingency tables



- How to hand in an exercise with tests:
 - Write down both Hypotheses
 - Write down α and n
 - Choose good test statistic
 - What is the distribution of the test statistic under \mathcal{H}_0
 - What are the critical values, the critical region
 - Calculate the value of the test statistic with the data given and decide
 - Alternatively: calculate p-value and compare with α

Relation between two variables

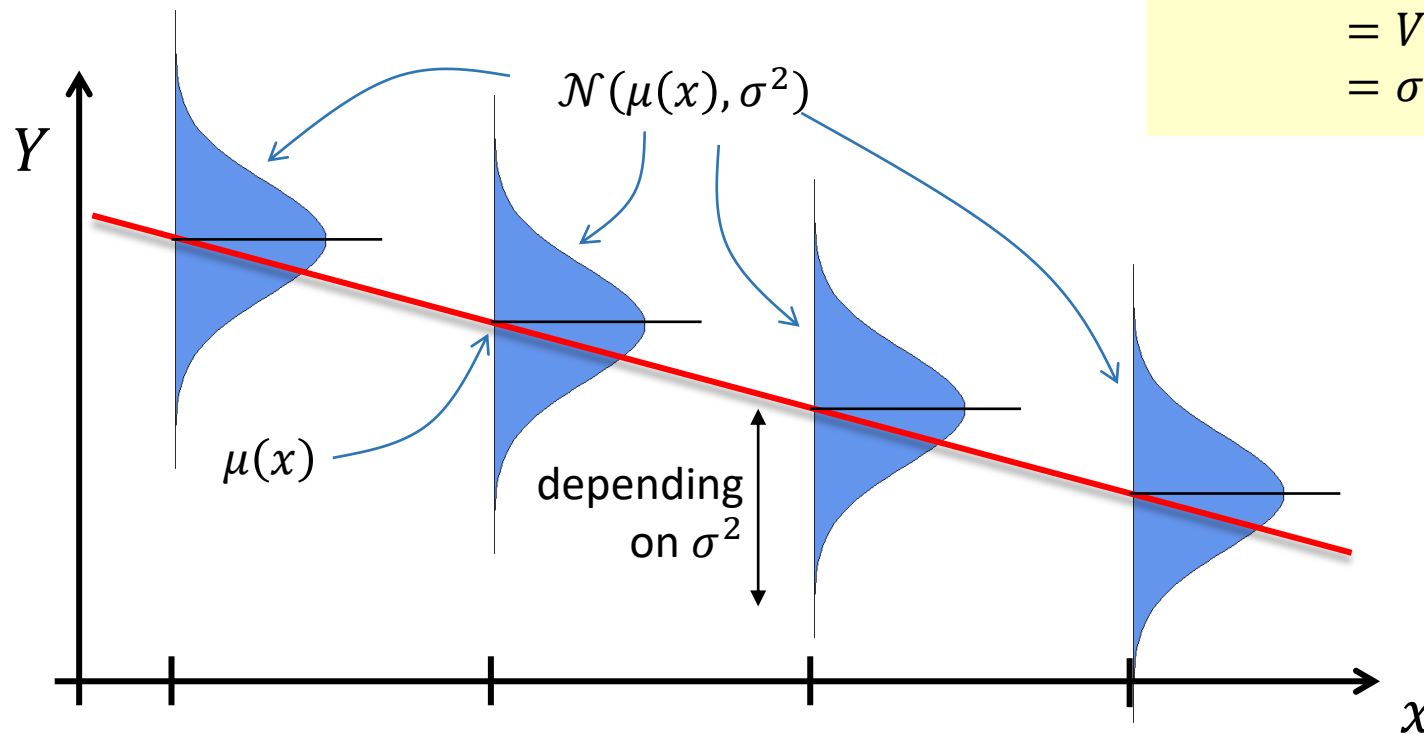


Correlation: linear relation, in $[-1, 1]$

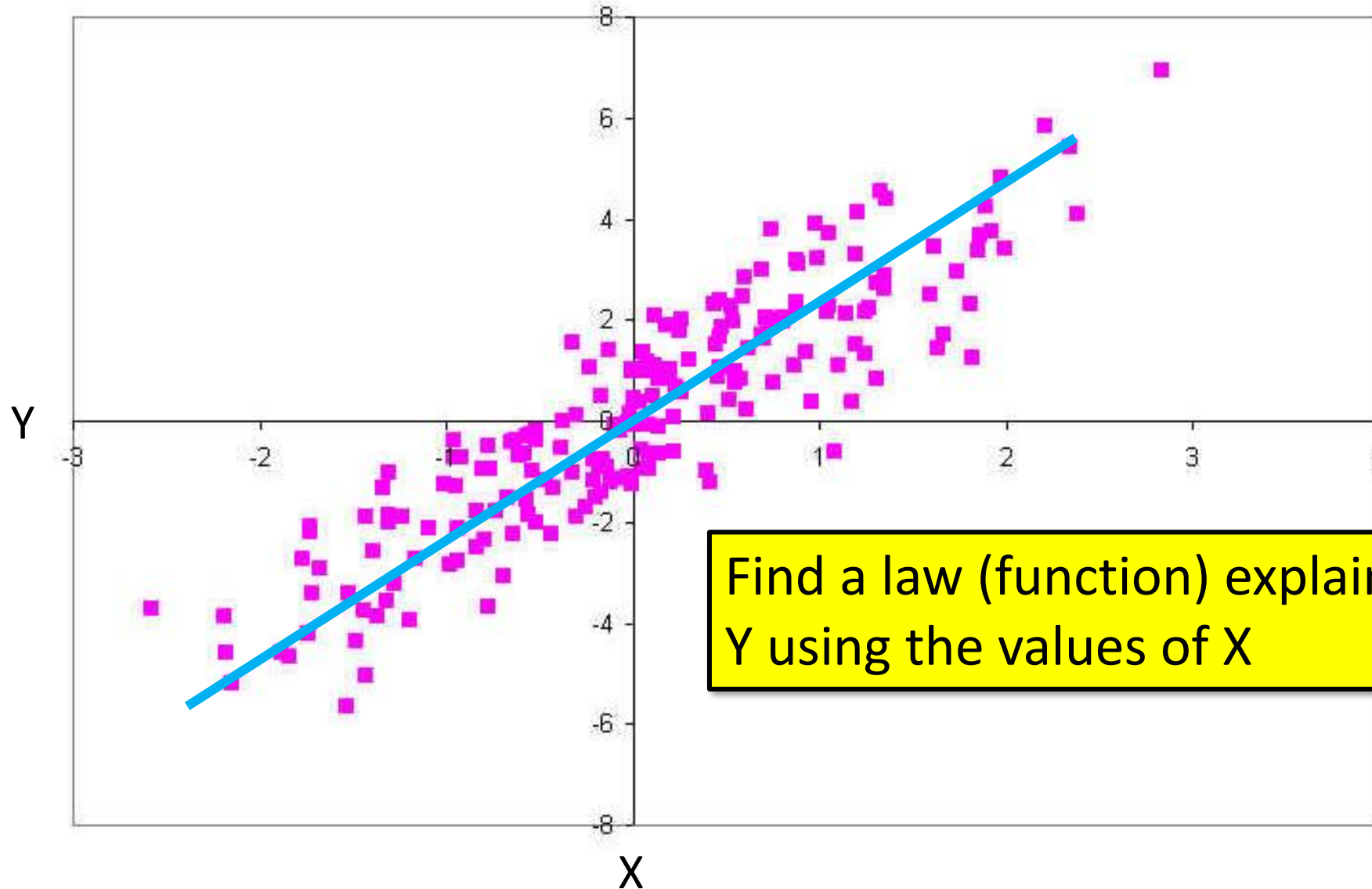
Linear regression: two definitions

1. $Y \sim \mathcal{N}(\mu(x), \sigma^2)$
 - $\mu(x) = \beta_0 + \beta_1 x$
2. $Y = \beta_0 + \beta_1 x + \varepsilon$
 - $\varepsilon \sim N(0, \sigma^2)$

$$\begin{aligned} E(Y) &= E(\beta_0 + \beta_1 x + \varepsilon) \\ &= \beta_0 + \beta_1 x + E(\varepsilon) \\ &= \beta_0 + \beta_1 x \\ \text{Var}(Y) &= \text{Var}(\beta_0 + \beta_1 x + \varepsilon) \\ &= \text{Var}(\varepsilon) \\ &= \sigma^2 \end{aligned}$$

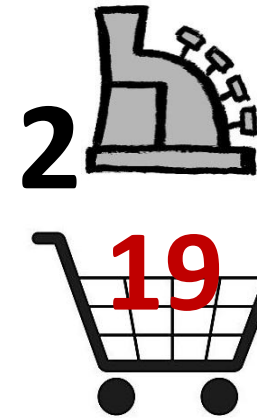


Linear regression: goal

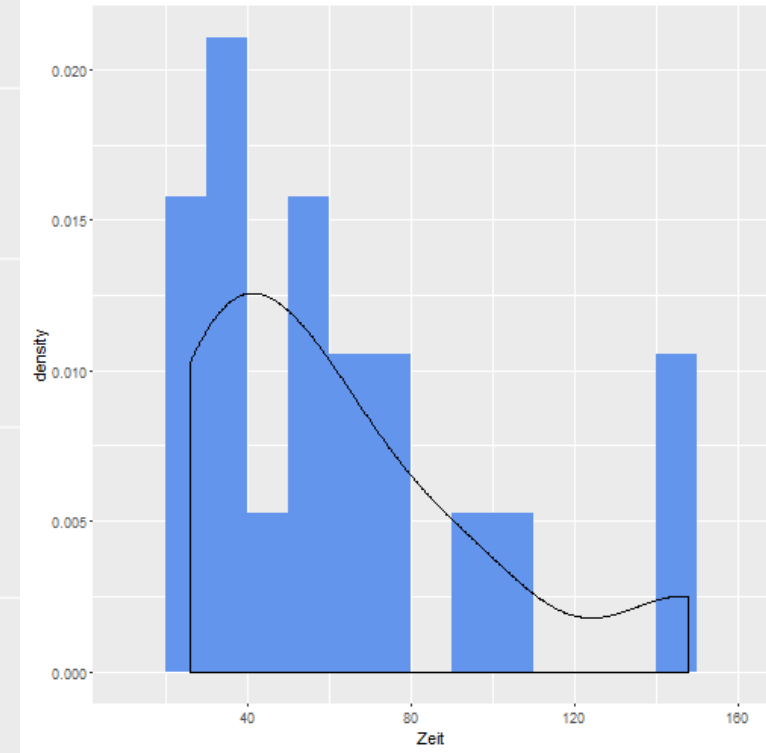
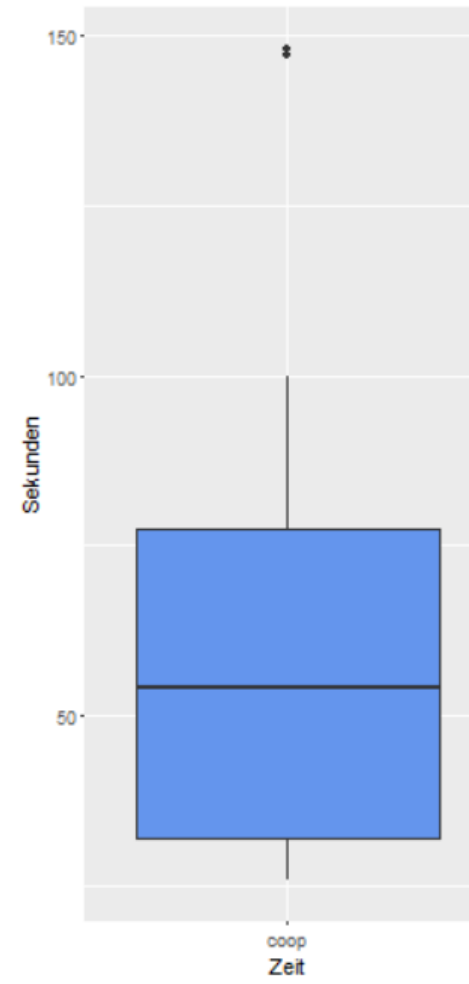
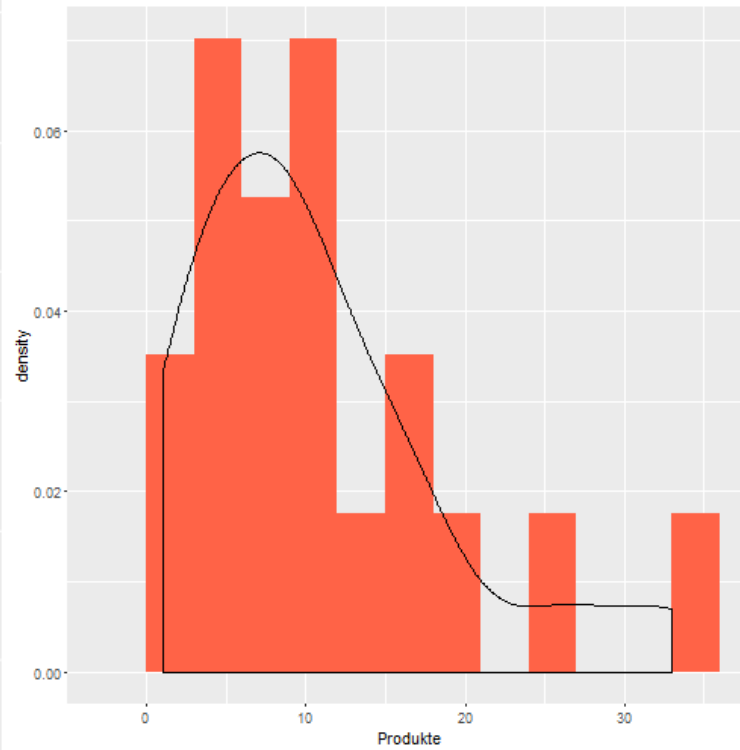
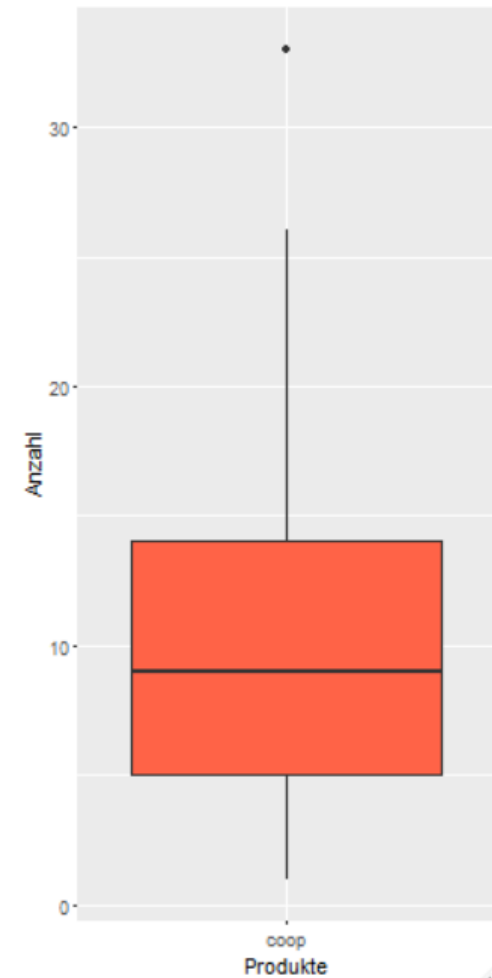


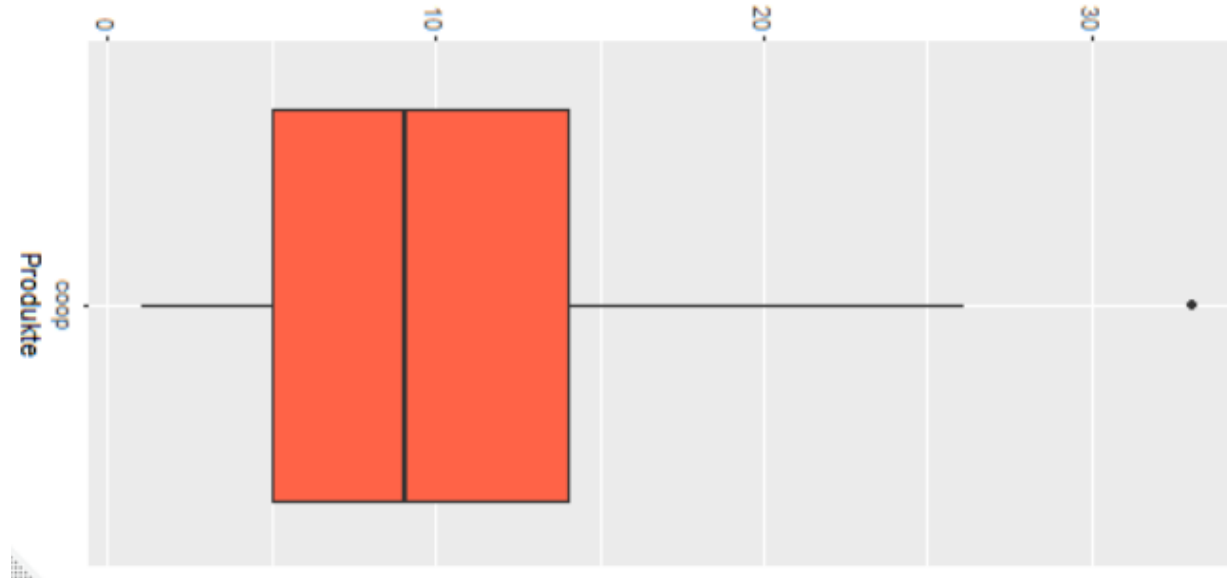
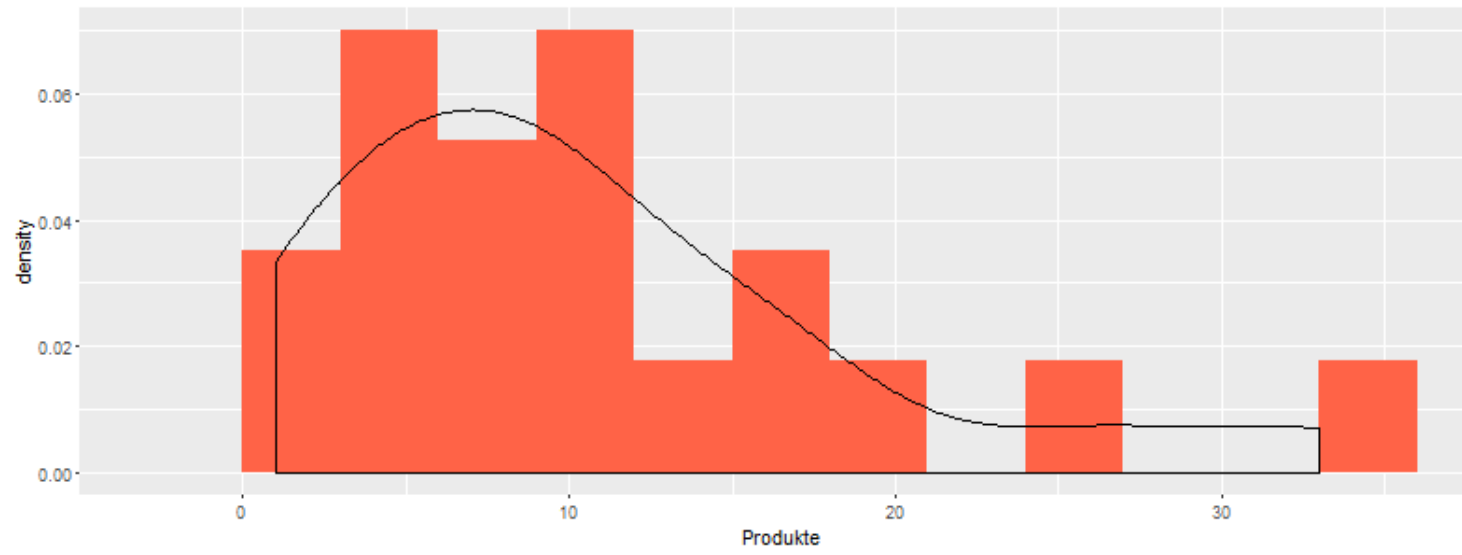


Where to queue?

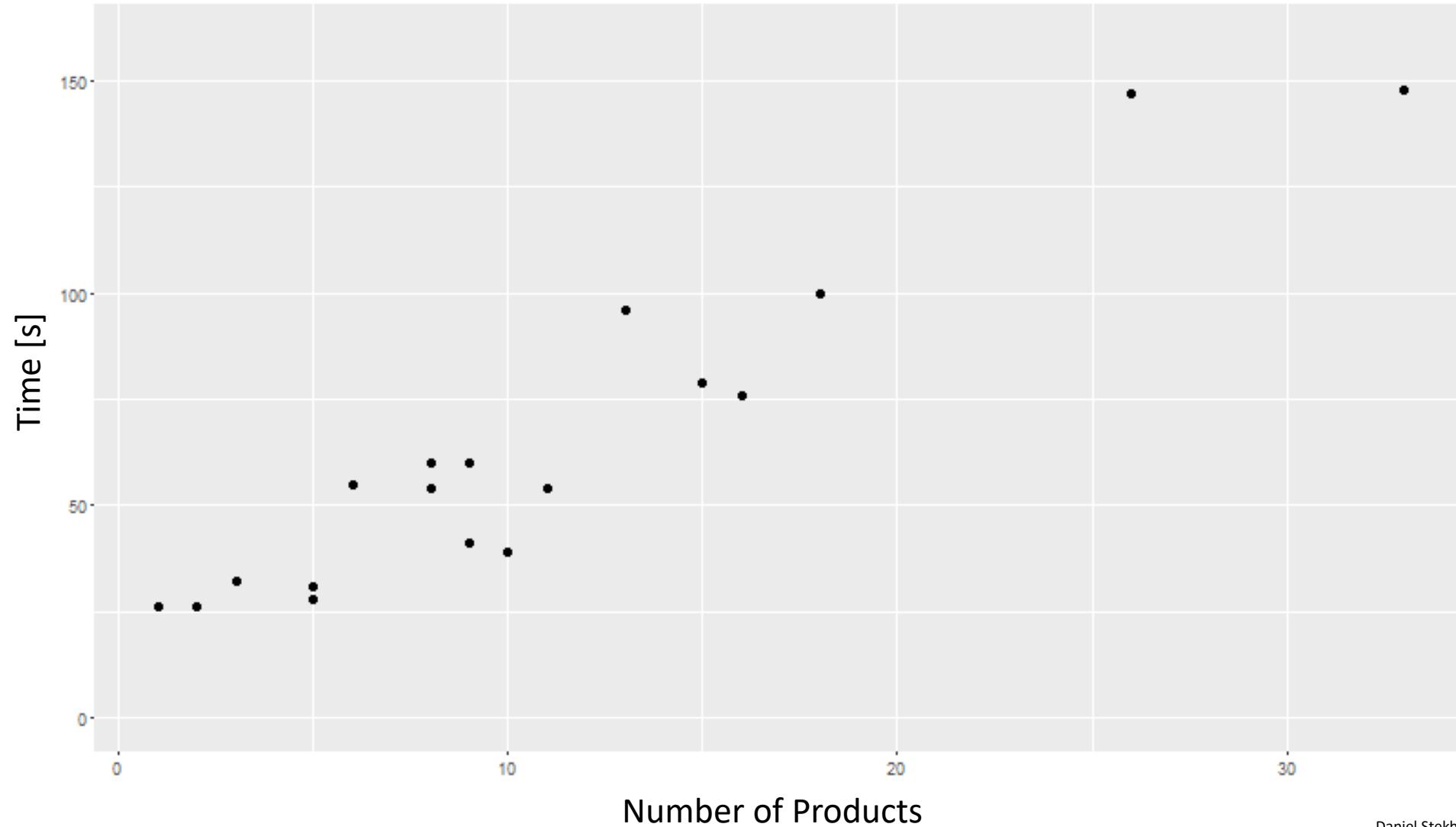


Migros Zurich main station one cashier, 17:40 – 18:00

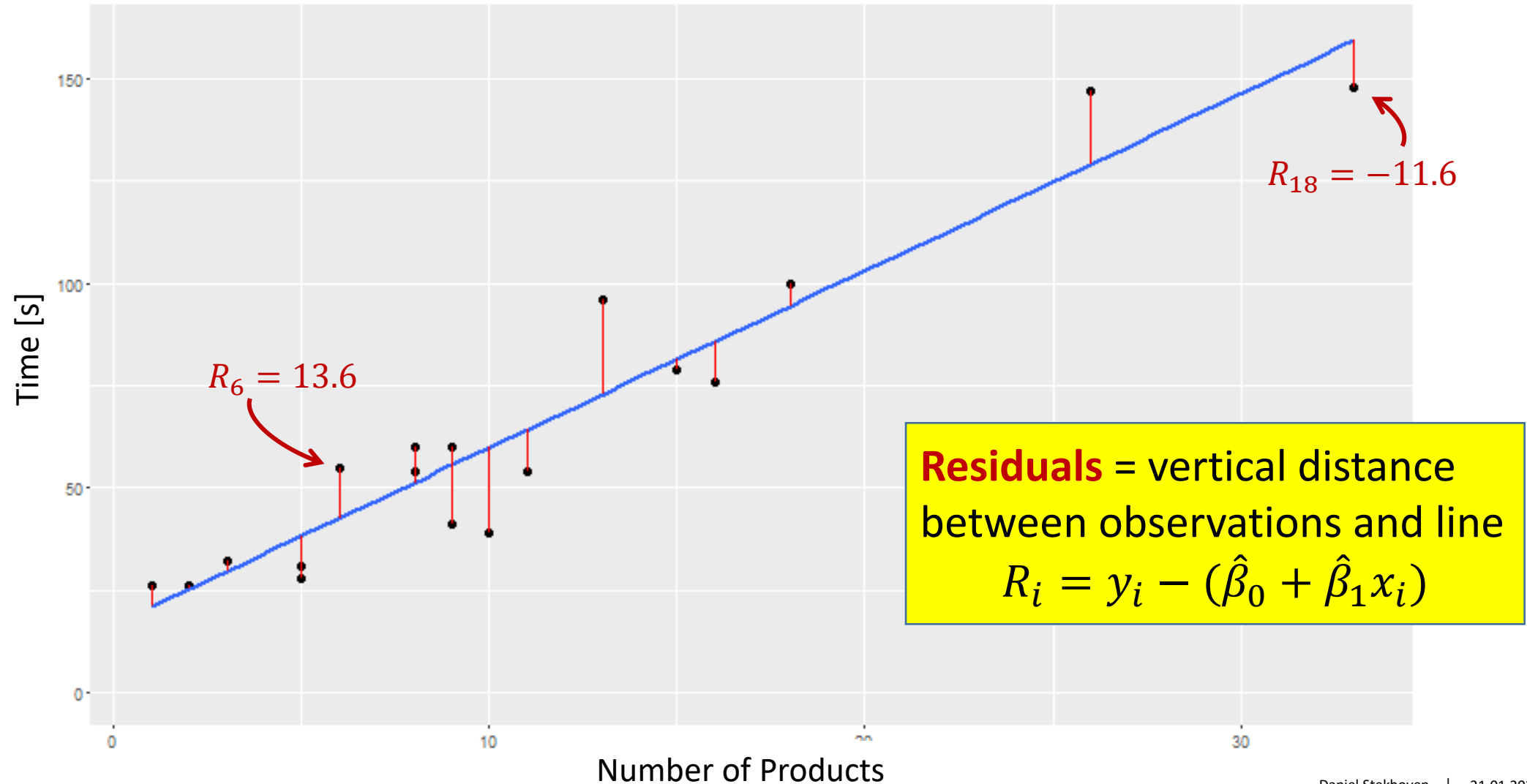




Scatter plot

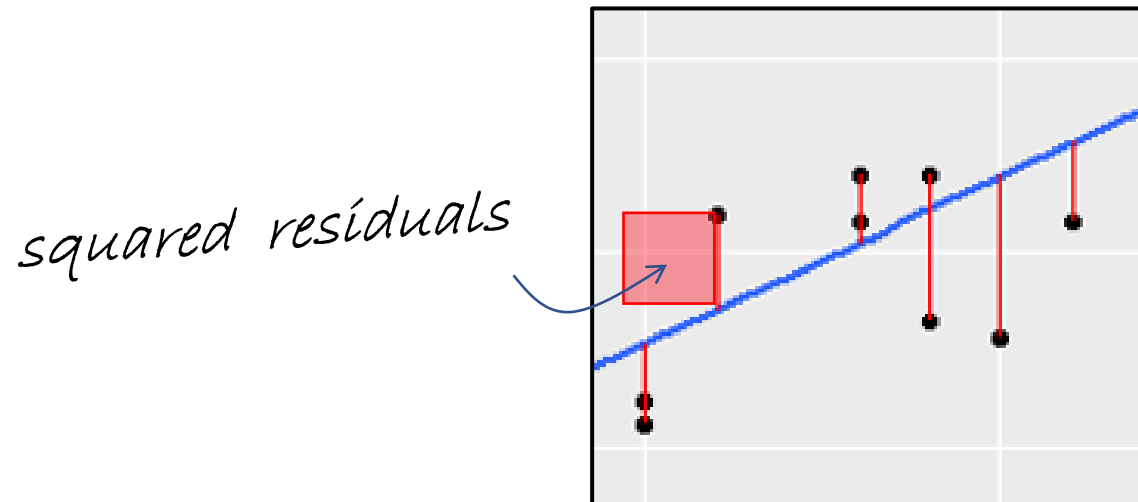


Residuals

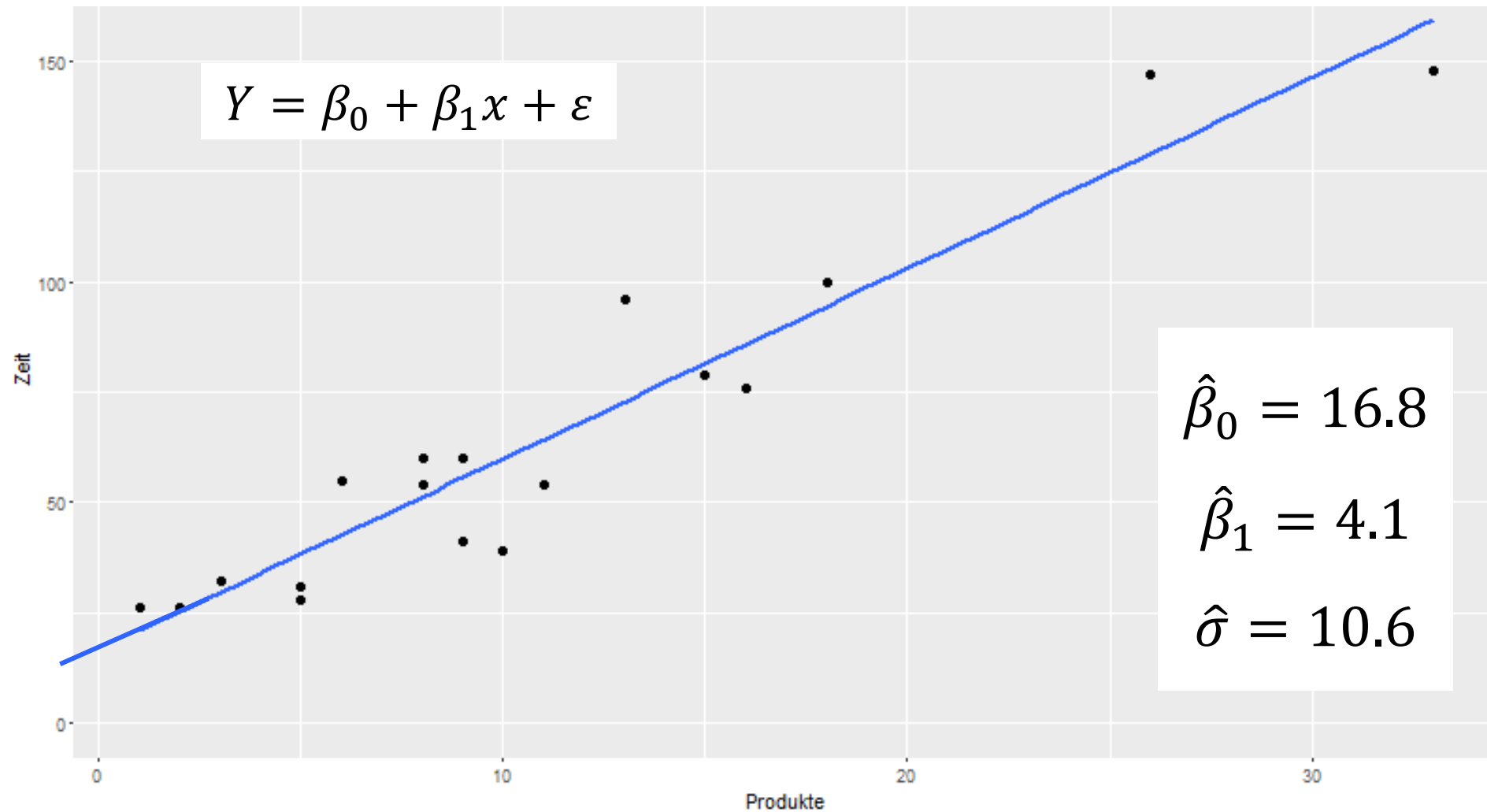


How are slope and intercept estimated?

- Which line fits best to the points? Several methods; one is:
- Choose intercept and slope such that the sum of the squared residuals is minimal

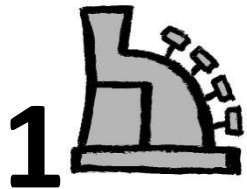


Regression line



Where to queue?

$\hat{\beta}_0 = 16.8$: intercept
 $\hat{\beta}_1 = 4.1$: slope
 $\hat{\sigma} = 10.6$: SD of error



$$16.8 + 3 \cdot 4.1 = 29.1 \text{ s}$$

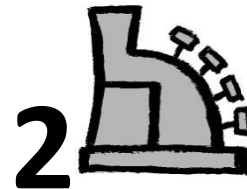


$$16.8 + 2 \cdot 4.1 = 25 \text{ s}$$



$$16.8 + 3 \cdot 4.1 = 29.1 \text{ s}$$

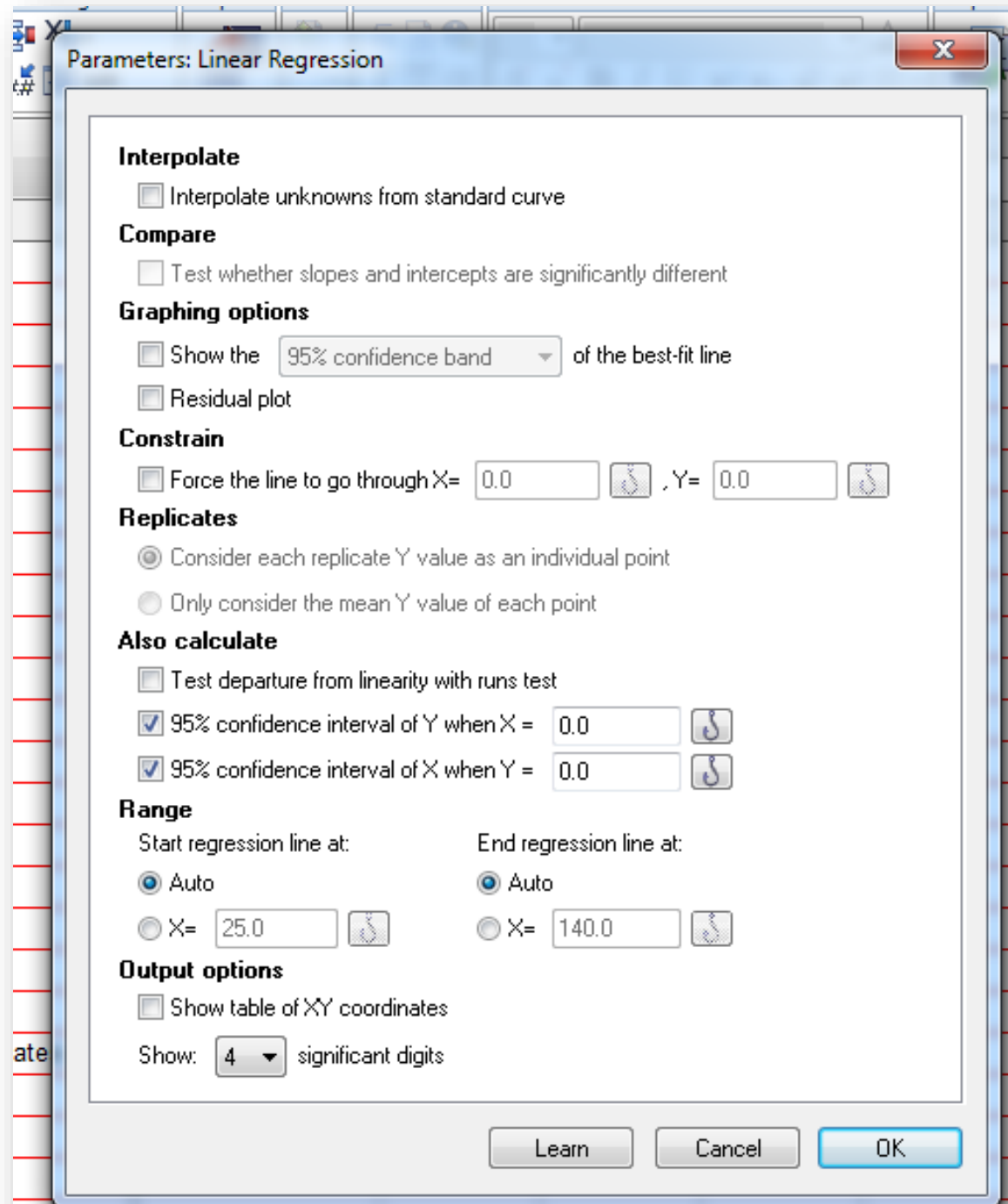
83.2



$$16.8 + 19 \cdot 4.1 = 94.7 \text{ s}$$

94.7

Linear Regression



Can we trust these parameters?

(test for β_1)

1	Best-fit values	
2	Slope	4.084 ± 0.3097
3	Y-intercept when X=0.0	16.85 ± 4.225
4	X-intercept when Y=0.0	-4.125
5	1/slope	0.2448
6	95% Confidence Intervals	
7	Slope	3.428 to 4.741
8	Y-intercept when X=0.0	7.891 to 25.81
9	X-intercept when Y=0.0	-7.347 to -1.705
10	Goodness of Fit	
11	R square	0.9158
12	Sy.x	10.60
13	Is slope significantly non-zero?	
14	F	173.9
15	DFn, DFd	1.000, 16.00
16	P value	< 0.0001
17	Deviation from zero?	Significant
18	Data	
19	Number of X values	18
20	Maximum number of Y replicates	1
21	Total number of values	18
22	Number of missing values	0
23		

Size of effect
for 1 product more it takes
4 additional seconds

True value lies within
this confidence interval

How good does the model
fit the data (1 is best)

Is the effect significant?

Summary

- P-values and hypotheses
- One-sided versus two-sided testing
- Performance of tests with fire alarms and spam filters
- Z-test, t-test
- Wilcoxon-test, sign-test
- Regression