



Empirical Software Engineering Research

Data Analysis

Norman Peitek, Annabelle Bergum, Lina Lampel, Sven Apel

Learning Goals

- Gain an overview over a typical data analysis process
- Be familiar with requirements of clean data
- Understand (dis)advantages of outlier removal

Homework:

What kind of data analysis have you done in the past? What were the steps?

Did you use a tool as support (Python, SPSS, Matlab, Excel, ...)?

“Smart” tools (Excel, ...) can be problematic

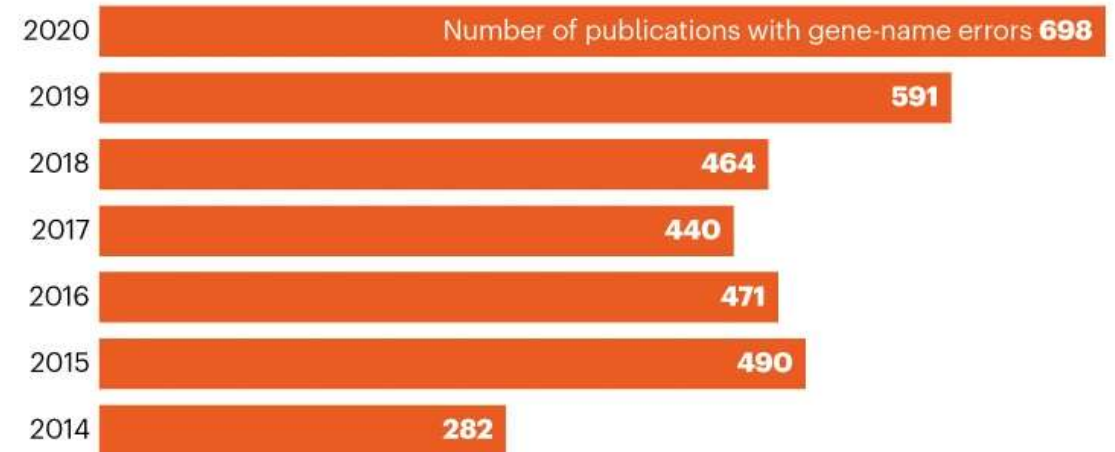
Ideally, use a tool that allows

- Automation
- Testing
- Version control

→ Popular choice nowadays: Jupyter notebooks

A GROWING PROBLEM

A 2016 analysis found that 20% of papers featuring gene names had errors created by spreadsheet autocorrect functions, but a bigger survey now finds the proportion is up to 30%. Since 2014, the number of papers with errors has increased significantly.



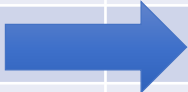
©nature

<https://www.nature.com/articles/d41586-021-02211-4>

Data cleaning is a semi-manual, supervised process. Depending on the scenario, it may cost more time than any other data-analysis step.

The first goal is to transform “dirty” data to “clean” data.

“Dirty” Data	“Clean” Data
Inaccurate	Accurate
Invalid	Valid
Incomplete	Complete
Inconsistent	Consistent
Duplicate entries	Unique



Data Cleaning: Accurate Data

- Data can be inaccurate for different reasons
 - The experiment design was not tested thoroughly
 - For example: “biweekly”
 - The measurement equipment has limitations in accuracy
- For example, eye-tracker allow a “validation” procedure before collecting data to check accuracy. If this check fails, redo the calibration procedure
 - After data collection, you may be able to correct a simple offset
- In many cases, there is no fix for inaccurate data, but you may need to remove it

```
public class RectangleDemo {  
    public static void main ( String [] args ) {  
        Rectangle box = new Rectangle ( 10 , 10 , 50 , 50 );  
        box.translate ( 10 , 10 );  
        System.out.println ( box.getX ( ) );  
        box.show ( );  
        box.draw ( );  
    }  
}
```

```
public class RectangleDemo {  
    public static void main ( String [] args ) {  
        Rectangle box = new Rectangle ( 10 , 10 , 50 , 50 );  
        box.translate ( 10 , 10 );  
        System.out.println ( box.getX ( ) );  
        box.show ( );  
        box.draw ( );  
    }  
}
```

Data Cleaning: Valid Data

- Data (especially user-submitted) can violate expectations
- For example: questionnaires asks participants for their "Age" and someone enters "1992"
- Online systems allow setting validation rules to avoid incorrect entry
- In many cases, you can manually transform invalid data to valid data
 - Otherwise, you may have to remove the specific information or the entire data set

4. In the fields below, please provide how many years you have been learning programming, programming in Java, and programming professionally. PG04

Learning programming means when you first started with programming, for example a class in high school or university. It continues as long as you are still programming in some capacity.

Programming professionally starts when you first get paid for your programming skills, for example your first job, which could also be an internship or student assistance.

Learning programming years

Programming in Java years

Programming professionally years

Your answer exceeds the maximum of 100.

Data Cleaning: Complete Data

- Data (especially user-submitted) can be incomplete
- For example: Participants may (un)intentionally skip questions in a questionnaire
- Online systems allow setting validation rules to force an entry → but could lead to bogus data
- Another example is eye-tracker missing data from participants' closing their eyes
- Typically, you cannot recover from incomplete data
 - Unless it was a simple copy & paste or download error
 - In some cases, you can replace missing values with approximated values (for example, average of the neighbors)

Please also answer this question – Your answer to this question is of importance to this study.

1. What is currently your profession?

PG02

- ☐ University student (undergraduate, bachelor)
- ☐ University student (graduate, master/PhD)
- ☐ University employee (postdoc, professor)
- ☐ Professional programmer/developer
- ☐ Other:

Data (especially user-submitted) can be inconsistent within a participant or across participants

For example: Participants checked that they are currently pursuing bachelor's degree, but already have a completed PhD

It is more likely that this was a user error rather than reality

But, in SE unusual cases exist (e.g., professional programmer without formal education)

An example from my experience: "Which semester are you currently in?"

Some in their master's degree responded with "2" since they were in their second semester of their masters

Others responded with "10", since they had 8 semesters of a bachelor degree beforehand

Some inconsistent data input can be avoided with well-phrased and tested questions

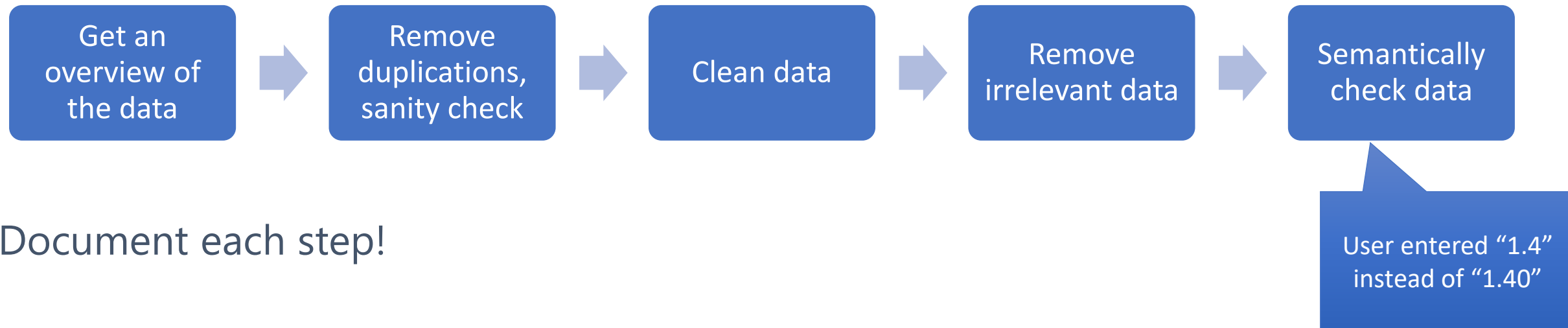
Data Cleaning: Unique Data

- In case of technical errors, data can be saved twice
- For example, a user completing an online questionnaire clicks “done” twice
- Modern systems handle these errors by default, but custom experiment scripts can be problematic
- But, typically, duplicate entries are easy to spot and remove

Data Cleaning: Prevent Issues

A lot of data issues can be avoided with tests and pilot studies!

I would recommend to continuously monitor data quality when running a study and, if necessary, make (minor) adjustments, such as adding an additional validation rule

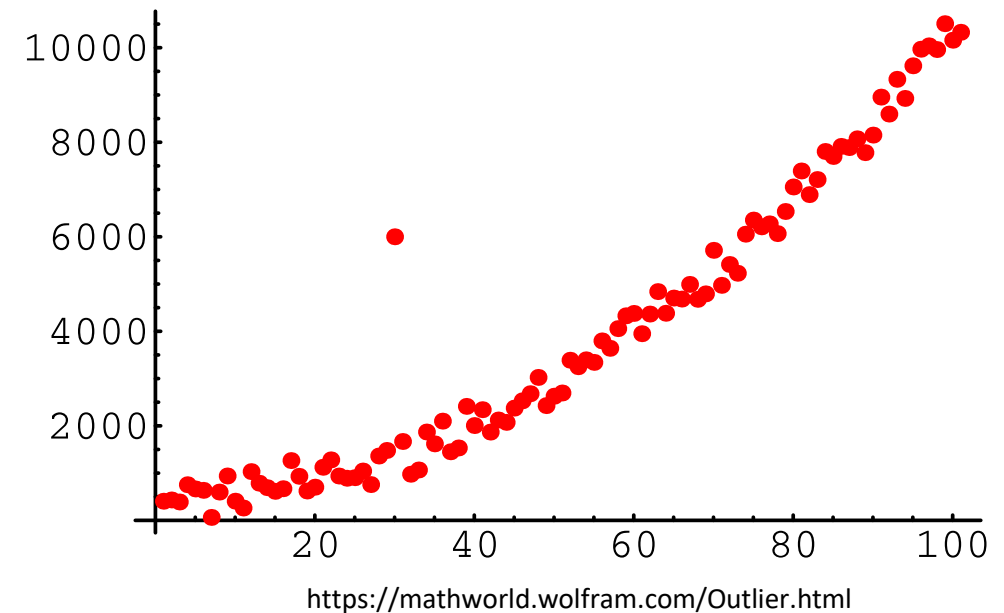


Document each step!

Outliers

Outliers are extreme values outside of most other data points. They can be true values or some form of error.

- True value: actual variation in data (can be interesting to study itself, 10xer)
- Error due to
 - Data entry error
 - Measurement error
 - Technical error (e.g., comma separation)



To understand whether a data point is an outlier, compare it to other data points (for example, from the same participant). Is it a reasonable value?

Typical approaches for detecting outliers

- Visualize data
- Top 5% (or 10%) and bottom 5% (or 10%)
- $IQR > 1.5$ (box plot), 3 SD from mean, ...

If there are many outliers, there might be some underlying confounding factor that was not considered

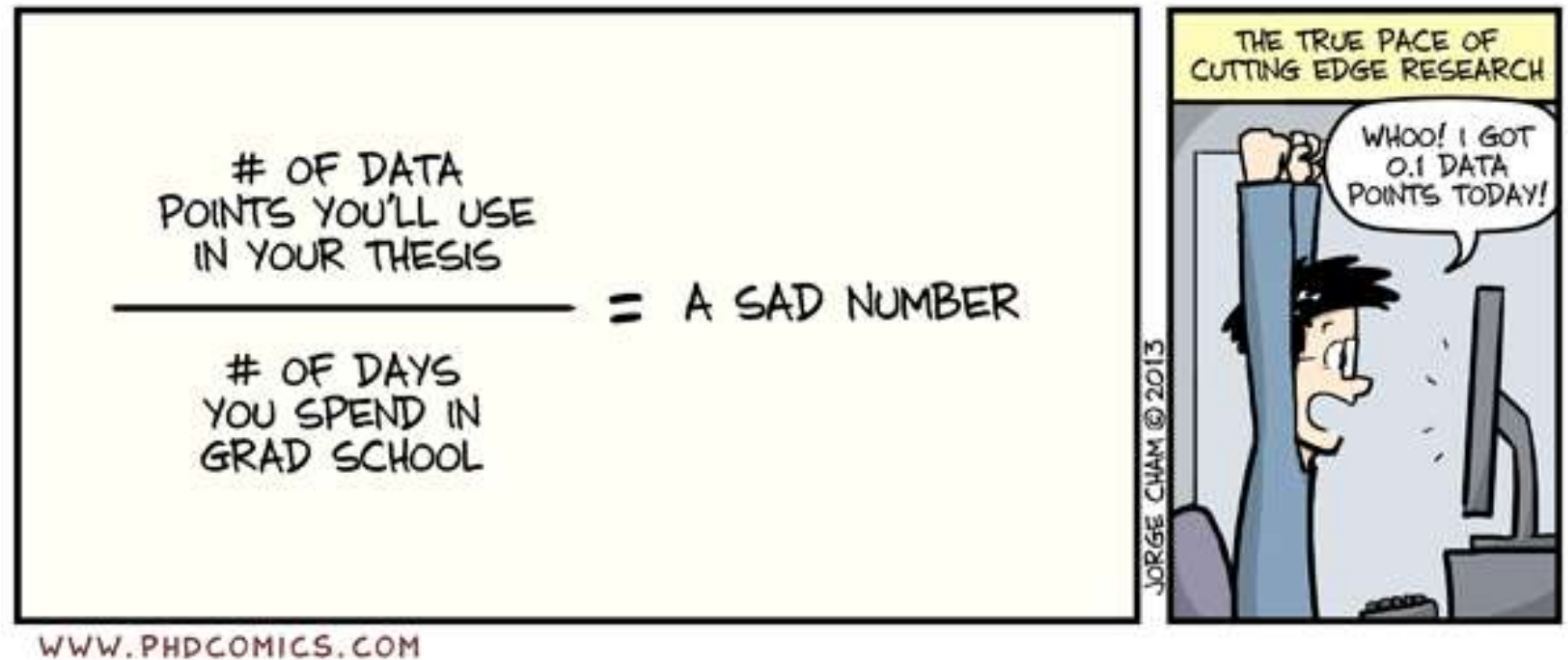
- Decision time: Keep outliers or remove them?
 - Outlier removal is controversial and there often is not a uniformly accepted approach.
 - Outlier removal is especially questionable with small data sets ($n < 25$)
- One option is to rely on reporting your experiment with outlier removal and, additionally, report without outlier removal.
 - Often the results are similar (in particular with non-parametric methods). If not, be careful.
- Ideally, decide outlier removal at the start
 - Especially if you have experience with expectations
 - Otherwise set very clear and strict rules.

TABLE III: Exclusion criteria

		Criterion	n
Language Level	German (1 - 6)	<4	2
	English (1 - 6)	<4	9
Programming Experience	C# Skill (1 - 5)	<4	24
	C# Experience (Years)	<1	8
Behavior	Encountered distractions?	Yes	17
	Worked on task conscientiously?	No	1
	Attempts to succeed	>3x	15
	Freeze, AFK (No interaction)	>1 min	4
	Too Slow (time per trial)	>10 min	14
Other	Participated in pilot study?	Yes	4
	Participated before?	Yes	8
Total (Criteria not mutually exclusive)			63

Outliers

Large n alleviates a lot of issues



Data preprocessing is an automated process that converts the data for further analysis.

- Typical are smoothing data (which can have a large effect)
- Examples: eye-tracking data, neuroimaging data
- Ideally, follow standards set in the literature
 - unless you have good reason for deviation

Questions?

Mini Test 09

Q 1: Invalid data can often be prevented by...

- a. making participants check their answers.
- b. collecting multiple data points per question.
- c. watching participants closely.
- d. setting automated validation rules.

Q 3: Missing data means that the experiment was conducted incorrectly.

- a. True
- b. False



Q 2: Outliers are...

- a. data points that do not fit our story.
- b. participants arriving late for an experiment.
- c. extreme values far away from most values.
- d. participants not showing up for an experiment.

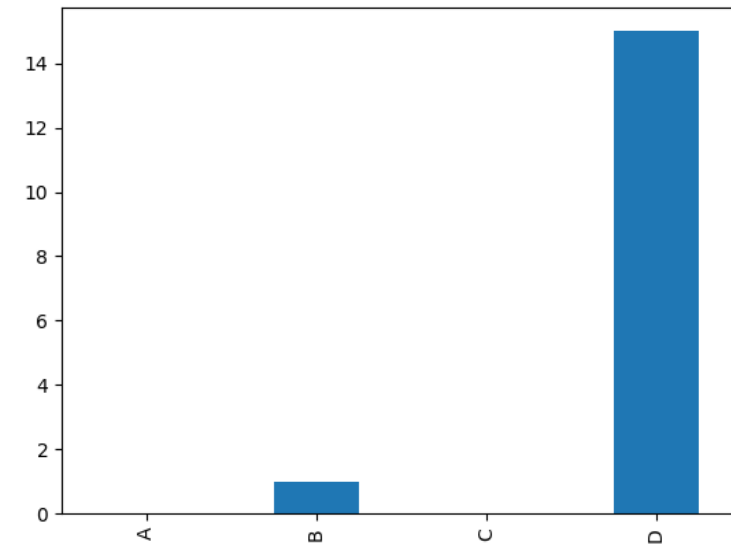
Q 4: Outliers are always measurement errors and should be deleted.

- a. True
- b. False

Mini Test 09

Q 1: Invalid data can often be prevented by...

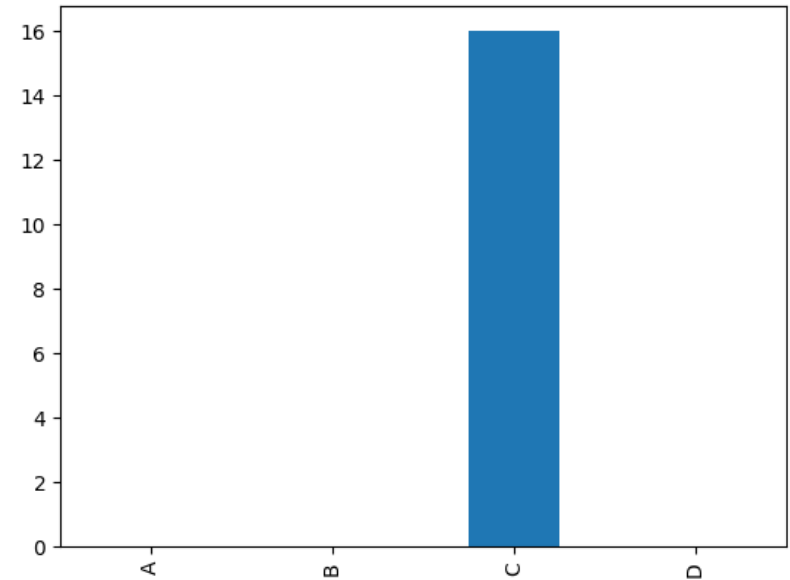
- a. making participants check their answers.
- b. collecting multiple data points per question.
- c. watching participants closely.
- d. setting automated validation rules.



Mini Test 09

Q 2: Outliers are...

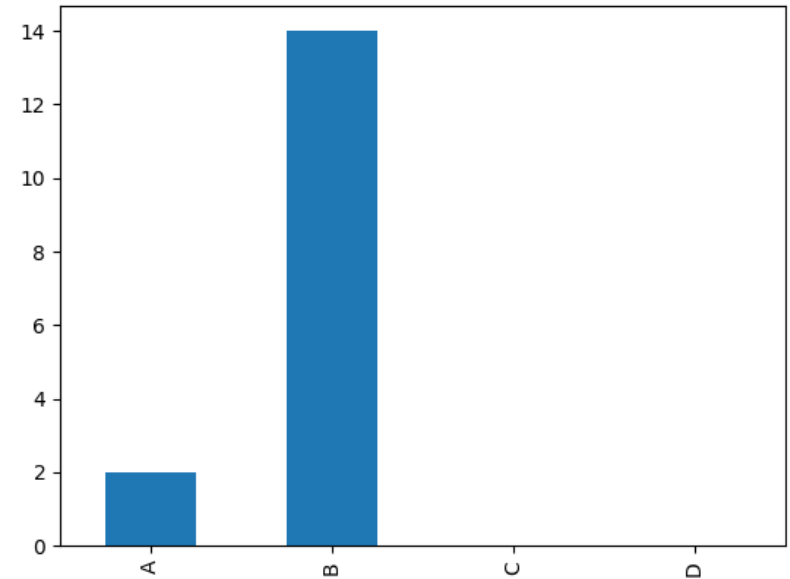
- a. data points that do not fit our story.
- b. participants arriving late for an experiment.
- c. extreme values far away from most values.
- d. participants not showing up for an experiment.



Mini Test 09

Q 3: Missing data means that the experiment was conducted incorrectly.

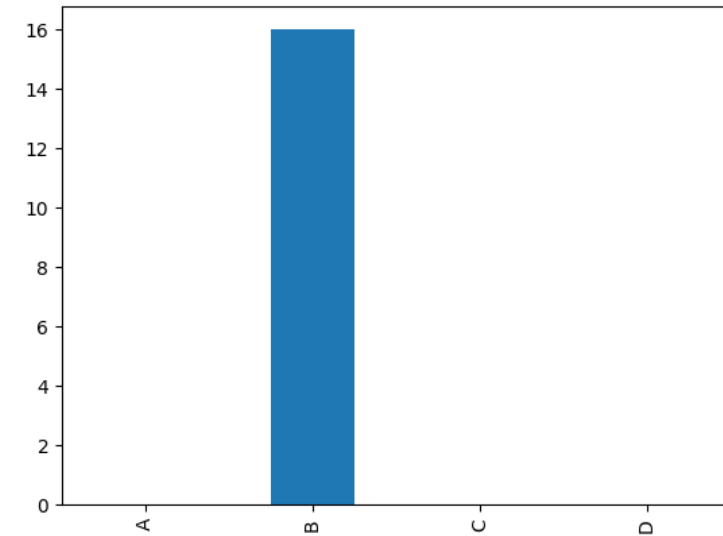
- a. True
- b. False



Mini Test 09

Q 4: Outliers are always measurement errors and should be deleted.

- a. True
- b. False



Inferential Statistics



Learning Goals

- Understand the purpose of hypothesis testing, p-values, and effect sizes
- Gain an overview over typical statistical tests*

**In this course, we treat statistical tests as black box*

- *We will not go into detail how the tests work, but you must understand when and why to use them*

- So far, we cleaned, preprocessed, and checked our data
- We computed descriptive statistics to describe our data
- Now, it is finally time to evaluate our hypothesis
 - Typically, we use inferential statistics to formally decide whether to reject or accept a hypothesis

- We already discussed hypotheses as predictive statements derived from research questions (or theory)
 - We now add that hypotheses are statistically testable: “hypothesis testing”
 - Hypothesis testing is set up with the **goal to refute the negation of the theory**
- H0 (null hypothesis): theory does not apply
 - For example: there is no effect of programming language on implementation speed
- H1 (alternative hypothesis): the theory predicts...
 - For example: there is an effect of programming language on implementation speed
- By default, we assume H0 is true, unless we find sufficient evidence to reject it

- A hypothesis test typically takes all collected data and simplifies them into a single number ("test statistic")

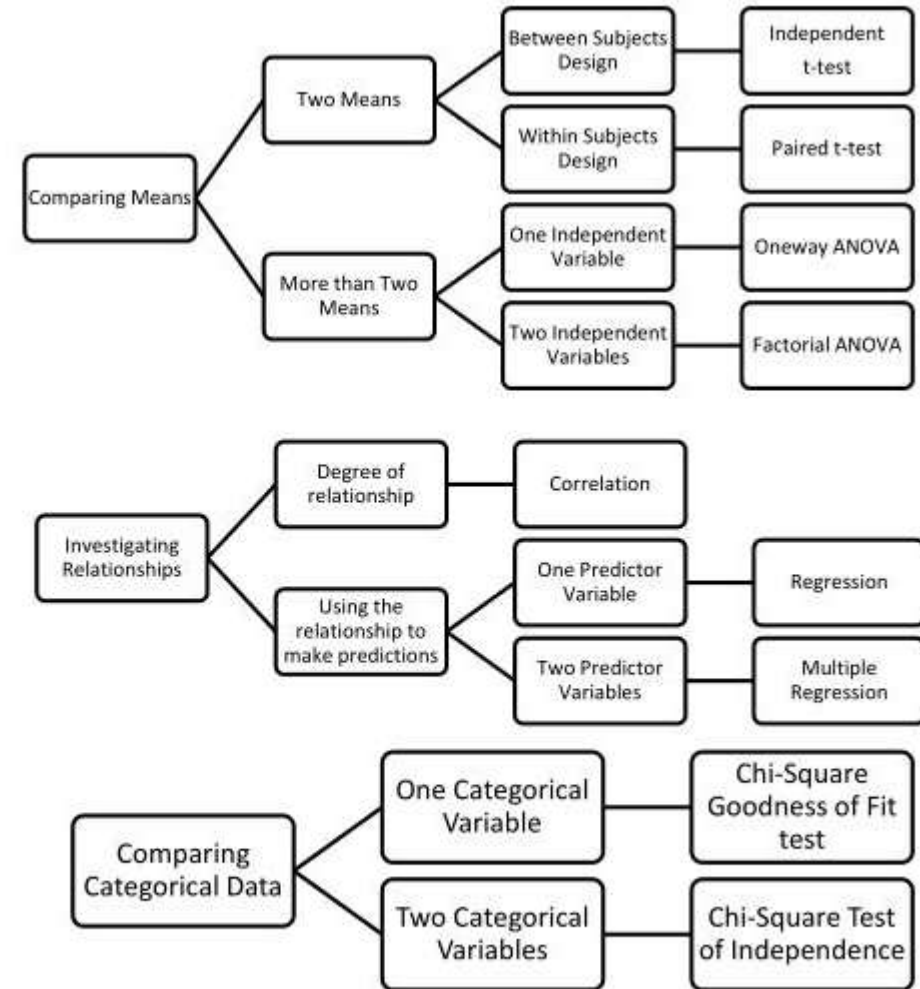


- The test statistic allows us to reject the null hypothesis (or not)
- The test statistic depends on the statistical test, but all of them allow us to compute a p-value

Overview

- Inferential statistics allow us to draw conclusions based on collected data
 - Note that they do not claim truth, they only state probabilities (depending on the chosen p-value)
- Be careful with selecting the right statistical test for your data
 - Guidelines and decision charts exist

Statistical Decision Tree



<https://www.slideserve.com/zitomira-pascha/statistical-decision-tree>

p-Value

- Most statistical tests will result in a p-value, indicating the probability that the observed data occurred by chance
 - A smaller p-value is „better“ as it allows us to reject the null hypothesis (= it did not occur by chance)
- Researchers (arbitrarily) set the significance level (α) meaning at what level we can we reject the null hypothesis
 - Typical thresholds for p-values are $p < 0.01$ and $p < 0.05$
 - Confidence level corresponds to $1 - \alpha$, so for example, 0.99 or .095

<u>P-VALUE</u>	<u>INTERPRETATION</u>
0.001	HIGHLY SIGNIFICANT
0.01	
0.02	
0.03	
0.04	SIGNIFICANT
0.049	
0.050	OH CRAP. REDO CALCULATIONS.
0.051	ON THE EDGE OF SIGNIFICANCE
0.06	
0.07	HIGHLY SUGGESTIVE, SIGNIFICANT AT THE $P < 0.10$ LEVEL
0.08	
0.09	
0.099	HEY, LOOK AT THIS INTERESTING SUBGROUP ANALYSIS
≥ 0.1	

“Good”

“Bad”

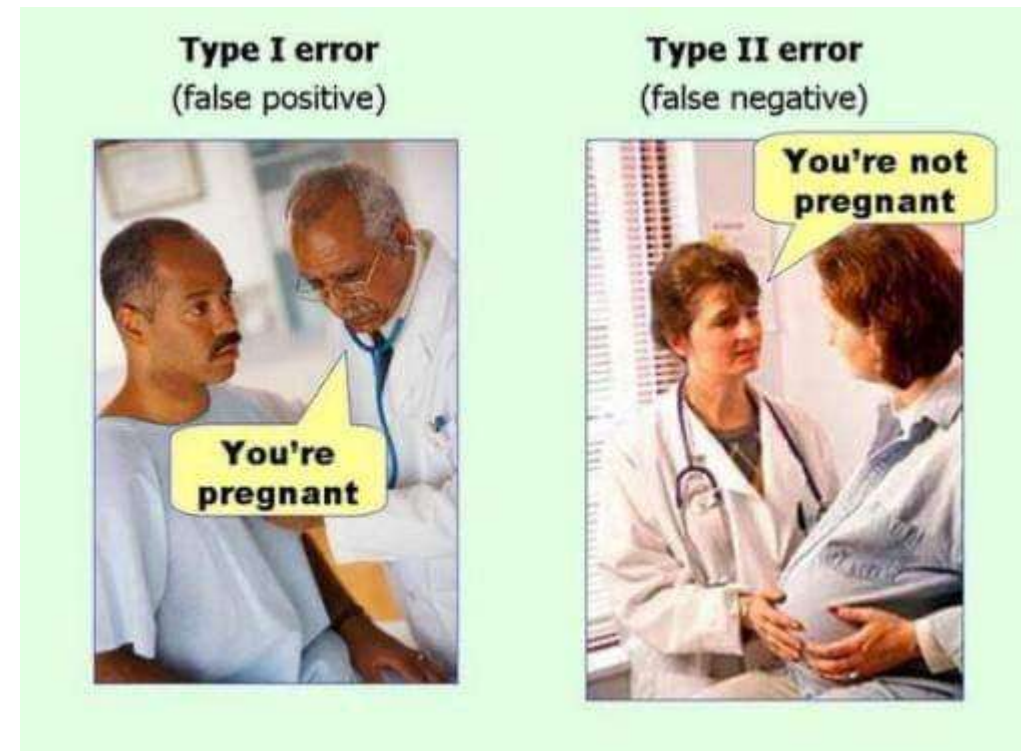
<https://xkcd.com/1478/>

Error Types

Statistical tests only provide a probability: they can be incorrect!

We typically focus on reducing Type 1 errors.

		Reality/Valid	
		H0	H1
Test Result	H0	✓	β error; Type-2 error (False negative)
	H1	α error; Type-1 error (False positive)	✓



p-Value

- In SE, many publications only report on the p-value. But that alone says very little about the observed effect
- Assume we have two offices of programmers. We observed them every day for a year and check how many lines of code they have written. The office in Saarbrücken always wrote exactly 1 LOC more than the one in Berlin.

	Saarbrücken	Berlin
March 5 th	1001	1000
March 6 th	765	764
March 7 th	1292	1291
...	... [197 more]	...

p-value < 0.00001 (extremely significant)

Effect Sizes

- Effect sizes describe how *large* the observed effect is

	Saarbrücken	Berlin
March 5 th	1001	1000
March 6 th	765	764
March 7 th	1292	1291
...	... [197 more]	...
	Saarbrücken	Berlin
March 5 th	2000	1000
March 6 th	1765	765
March 7 th	2292	1292
...	... [197 more]	...

p-value < 0.00001 (extremely significant)

Effect size of 0.005 (negligible)

p-value < 0.00001 (extremely significant)

Effect size of 0.8 (large)

- Effect sizes are typically only reported when the effect is statistically significant
- Depending on data normality and data type
 - Cohen's d (normally distributed continuous data)
 - Cliff's delta (non-normal ordinal data)
 - ...

Pairwise Different Baselines	Mann-Whitney U	Cliff's delta
PROGCOMPR>CROSSFIX vs. PROGCOMPR>MATH	0.027	−0.163 (small)
PROGCOMPR>CROSSFIX vs. PROGCOMPR>READ	0.002	0.231 (small)
PROGCOMPR>CROSSFIX vs. PROGCOMPR>PROBSOLV	0.000	−0.323 (small)
PROGCOMPR>MATH vs. PROGCOMPR>READ	0.300	—
PROGCOMPR>MATH vs. PROGCOMPR>PROBSOLV	0.019	0.177 (small)
PROGCOMPR>READ vs. PROGCOMPR>PROBSOLV	0.103	—

- How „large“ an effect is depends on the field
- Study in the medical field investigated whether aspirin helps to prevent heart issues across 22'000 participants, effect size of 0.001
 - But, in medical care, this helps a lot → unethical to continue the study and everyone started to take aspirin
- Effect sizes can be calculated retroactively, for example, if a paper did not report them
 - For that, you need: n , mean +SD
 - → now you know why that is often reported

Independent T-Test

- Comparing two measurements from independent samples
- For example, comparing response times of professional versus novice programmers
- `scipy.stats.ttest_ind(a, b)`

Null hypothesis (H_0)	Alternative hypothesis (H_1)
Statistical hypotheses	
Measurements do not differ	Measurements differ significantly
Formal: $H_0 : \bar{x}_1 = \bar{x}_2$	Formal: $H_1 : \bar{x}_1 \neq \bar{x}_2$

Independent T-Test: Results

Determines probability of observed result, under the assumption that the null hypothesis is valid → conditional probability

If probability is smaller than...

0.001:	Highly significant
0.01:	Very significant
0.05:	Typical significant
0.10:	Exploratory/initial studies

... the null hypothesis must be wrong

Significance level must be defined in advance!

Independent T-Test: Results

What does significant result mean?

- Is null hypothesis incorrect? → No
- Is alternative hypothesis correct? → No
- There is no evidence that the null hypothesis is valid
- Writing a report
 - Reject/could not reject null hypothesis
 - Never: Confirmation of null or alternative hypothesis

Welch's T-Test

The independent t-test assumes

- the measurements are normally distributed, and
- the variances between the two measurement groups are equal

Welch's t-test is a version that does not assume equal variances and is more robust. Its power is close to an independent t-test when there are equal variances → some researchers always use Welch's t-test.

```
scipy.stats.ttest_ind(a, b, equal_var=False)
```

Dependent (Paired) T-Test

The independent t-test assumes that the two samples are independent of each other.

In a within-subject design, this is often not the case.

For example, comparing a programmer's response time before and after a training.

The two values are dependent of each other ("paired").

→ It requires equal number of samples (missing data is problematic)

```
scipy.stats.ttest_rel(a, b)
```

All T-Tests: One-Tailed & Two-Tailed

- Two-tailed:
 - No assumption about direction of effect (e.g., which of two UIs is more usable, productivity of two programming languages)
 - Compute half of significance level
- One-tailed:
 - There is an assumption about the direction (e.g., darkmode UI is more usable, experts are more productive than novices)
 - No need to cut significance level in half

Live Demo

T-Tests: Assumptions/Requirements

The t-tests have some assumptions for the data:

- Data in an interval/ratio scale type
- Data is normally distributed (e.g., test with Shapiro-Wilk)
 - Or: sample size $> x \in \{30, 50\}$

Mann-Whitney-U Test

- A non-parametric test to compare two means (similar to independent t-test)
- Assumes at least ordinal data

```
scipy.stats.mannwhitneyu(a, b)
```

Wilcoxon Signed-Rank Test

- A non-parametric test to compare two means that are dependent of each other (similar to dependent t-test)
- Assumes at least interval data

`scipy.stats.wilcoxon(a, b)`

Analysis of Variances (ANOVA)

ANOVA evaluates how much the variance in dependent variable can be explained by variance in the independent variable(s)

- Multiple levels of independent variable(s)
- Controls for multiple-testing issue

H_0 : Mean of all groups is the same

H_1 : At least two means differ (but not *which* one are different)

Comparing Categorical Data



Chi-Squared (χ^2)-Test

- Compares frequencies of categorical data
- Can answer two questions:
 - Do observed frequencies deviate from expected frequencies?
 - Do observed frequencies deviate from each other?

```
cross = pd.crosstab(data['class'], data['alive'])
stats.chi2_contingency(cross)
```

- Assumptions
 - Comparison of frequencies
 - Expected frequencies > 5 (Fisher's exact test, otherwise)
 - Nominal scale type

```
cross = pd.crosstab(titanic_data['class'], titanic_data['alive'])
print(cross)
stats.chi2_contingency(cross)
```

[44] ✓ 0.4s

	alive	no	yes
class			
First	64	120	
Second	90	83	
Third	270	85	

(91.08074548791019,
1.6675060315554636e-20,
2,
array([[109.57303371, 74.42696629],
[103.02247191, 69.97752809],
[211.40449438, 143.59550562]]))

Fisher's Exact Test

Also compares frequencies of categorical data, but with small n

```
cross = pd.crosstab(data['sex'], data['alive'])  
scipy.stats.fisher_exact(cross)
```

Assumptions:

- 2x2 contingency table
 - → does not work the previous example
- Nominal scale type

```
cross2 = pd.crosstab(titanic_data['sex'], titanic_data['alive'])  
print(cross2)  
stats.fisher_exact(cross2)
```

✓ 0.4s

alive	no	yes
sex		
female	64	195
male	360	93

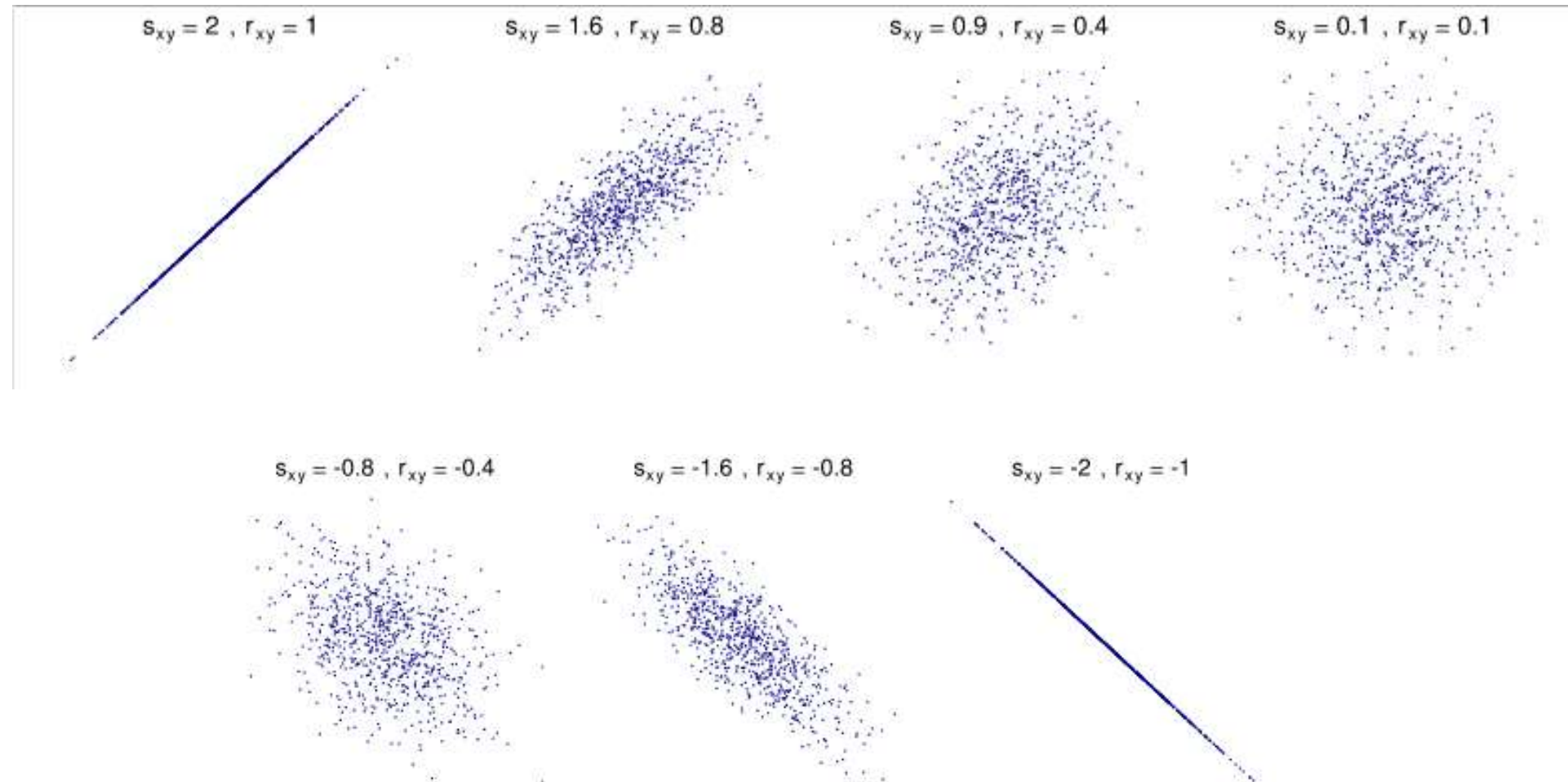
(0.0847863247863248, 1.8885888329631993e-47)

Relating Two/Multiple Variables



- Value for relationship in data
 - Remember, they do not imply causality
- Values range from: $-1 \leq r \leq +1$
 - $r : 0.0-0.1$: no relationship
 - $r : 0.1-0.3$: weak relationship
 - $r : 0.3-0.5$: medium relationship
 - $r : >0.5$: strong relationship
- Sign indicates whether relationship is positive or negative

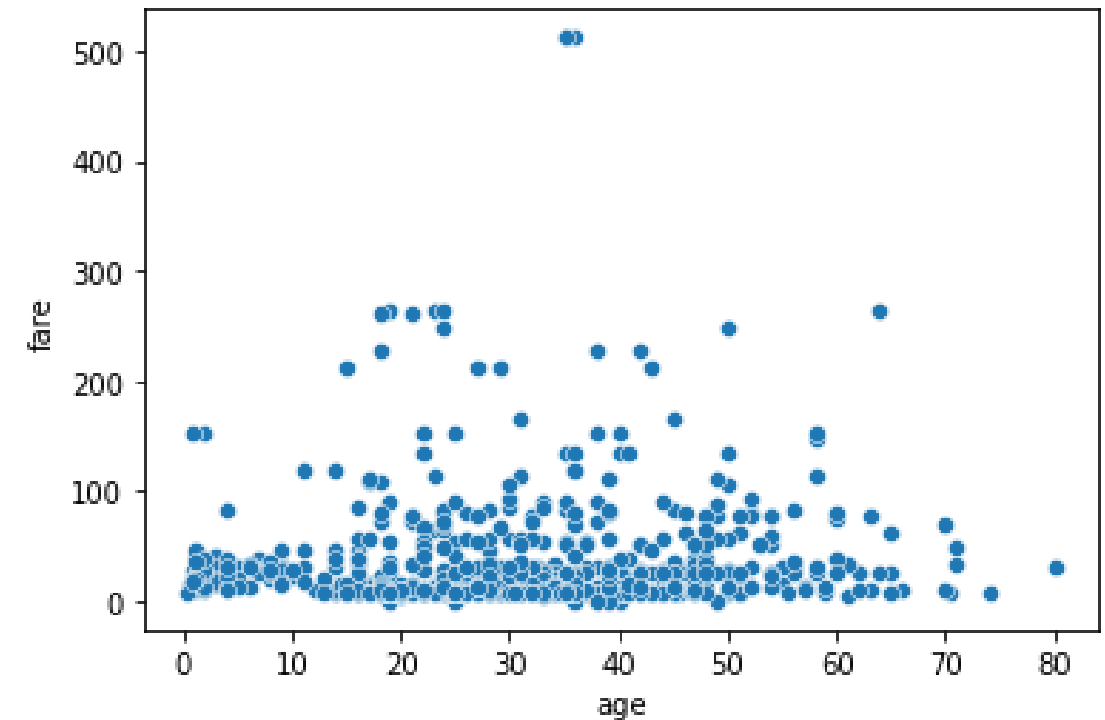
Correlations: Examples



Correlations: Pearson's r

Assumptions:

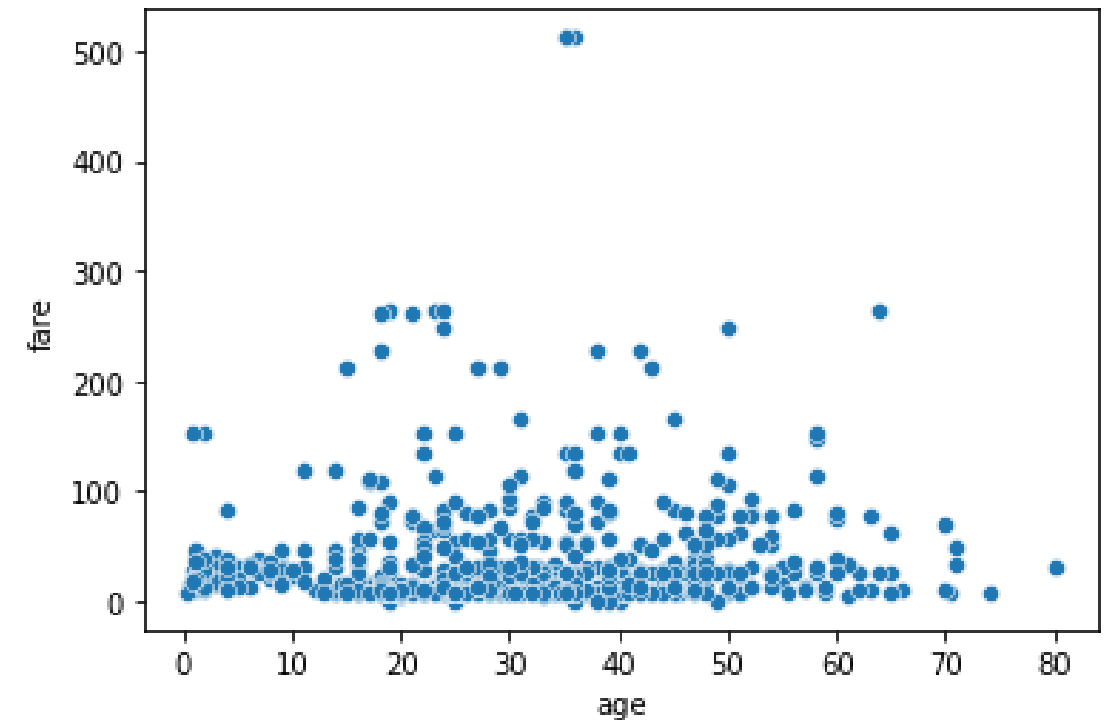
- Data on an interval/ratio scale
- Normally distributed data
- `scipy.stats.pearsonr(a, b)`



$r=0.09$

Correlations: Spearman-Rank ρ

- Accepts non-normally distributed data
- Assumes at least ordinal data
 - Ordinal-ordinal
 - Ordinal-interval/ratio
- Deals well with outliers (unlike Pearson's r)



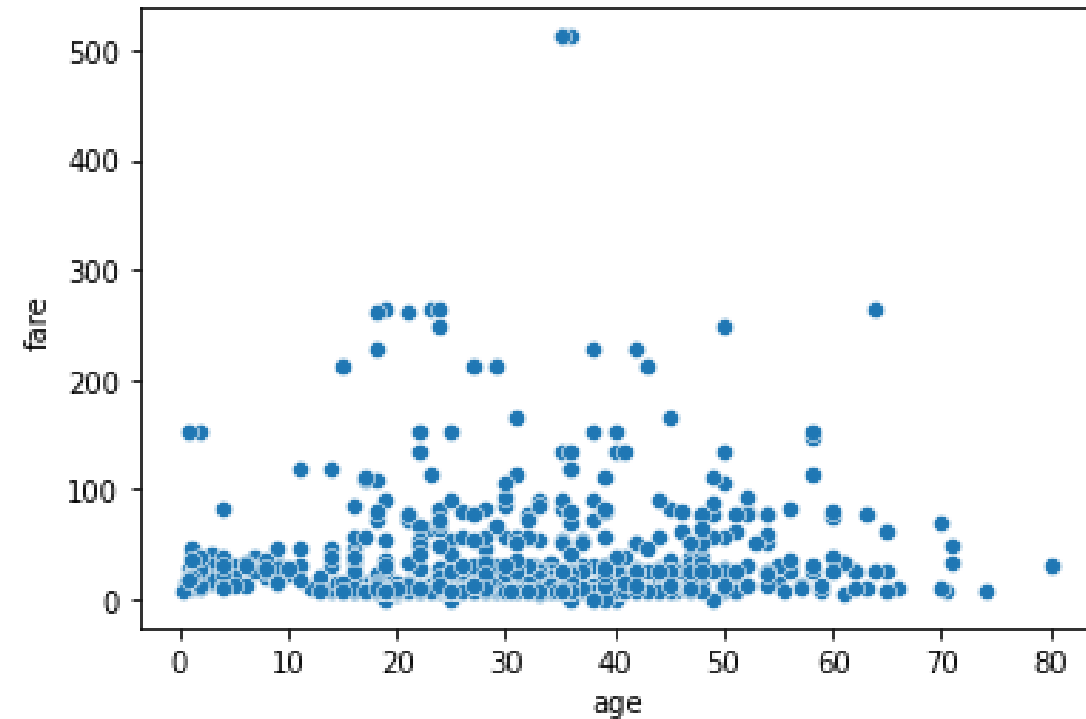
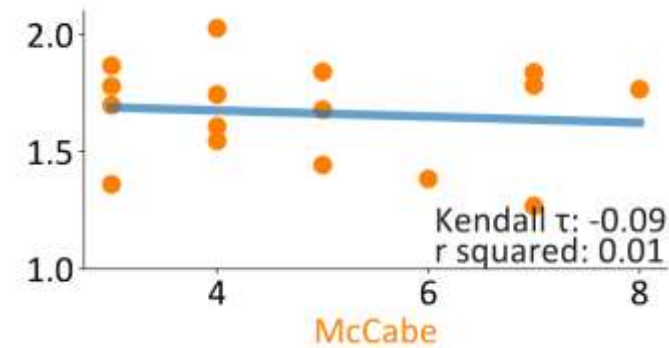
`scipy.stats.spearmanr(a, b)`

$\rho=0.13$

Correlations: Kendall's tau

- Accepts non-normal distributed data
- Assumes at least ordinal data
 - Ordinal-ordinal
 - Ordinal-interval/ratio
- Deals well with outliers and repeated values

`scipy.stats.kendalltau(a, b)`



tau=0.09

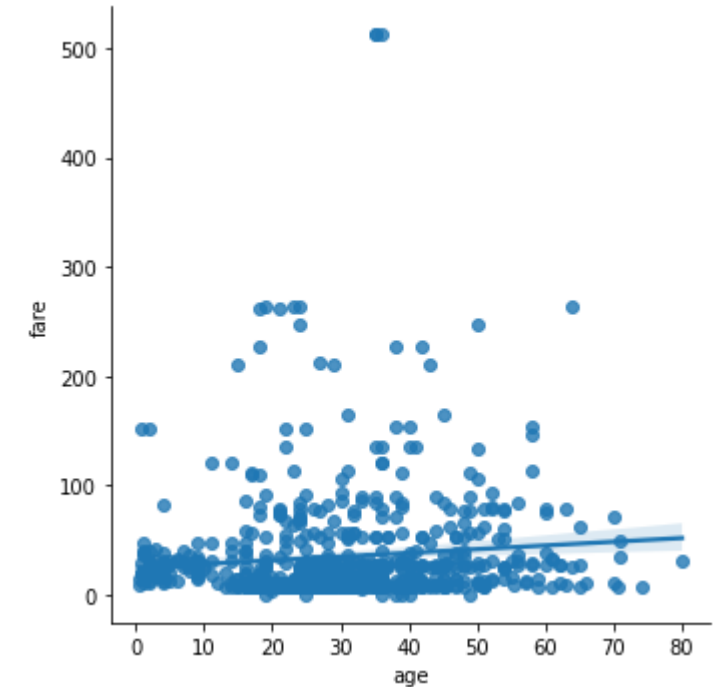
Correlations and p-Values

- Depending on the correlation, there are different significance tests
- “Significance” means the correlation is most likely different from 0
- Often not accurate for small data sets
- → in SE, p-values of correlations are typically not reported

Regression

- Regression is predicting a variable based on some input
- For example, linear regression: $y = b \cdot x + a$

`scipy.stats.linregress(a, b)`



- They are similar to correlations, but imply causality
 - However, that is not automatically given. It must be part of the experiment design

When conducting multiple significance tests, the significance level needs to be adapted.

For example:

- 1 t-test at p-level of 0.05 the chance of type 1 error is: 0.05
- 2 t-test at p-level of 0.05 the chance of type 1 error is: 0.10
- 3 t-test at p-level of 0.05 the chance of type 1 error is: 0.14
- ...

Some statistical tests correct for multiple testing (e.g., ANOVA). Otherwise, it can be manually corrected to reduce chances for type 1 errors.

Two main approaches are Bonferoni and false-discovery rate (FDR)

Bonferoni correction	FDR correction
Correct p-level: $\alpha/\text{number of tests}$	FDR: $\text{false positives}/(\text{false positives}/\text{true positives})$
For example, $0.05/5 = 0.01$	
Very strict	Less strict

Questions?

Mini Test 10

Q 1: Data scale type matters when choosing the correct statistical test.

- a. True
- b. False



Q 2: Data distribution (normality) matters for choosing the correct statistical test.

- a. True
- b. False

Q 3: Fisher's exact test is preferable over chi-squared test for small n.

- a. True
- b. False

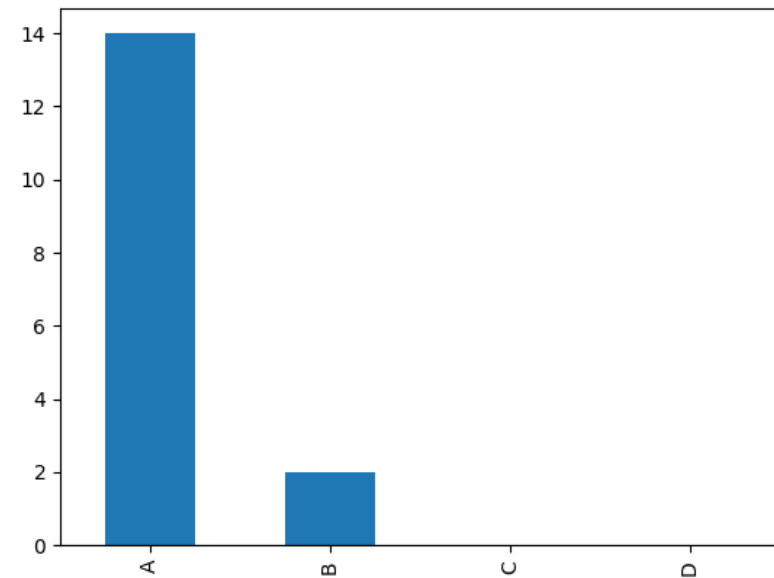
Q 4: Which of the following is NOT a correlation coefficient.

- a. Kendall's tau
- b. Pearson's r
- c. Spearman's rank
- d. Bonferroni's r

Mini Test 10

Q 1: Data scale type matters when choosing the correct statistical test.

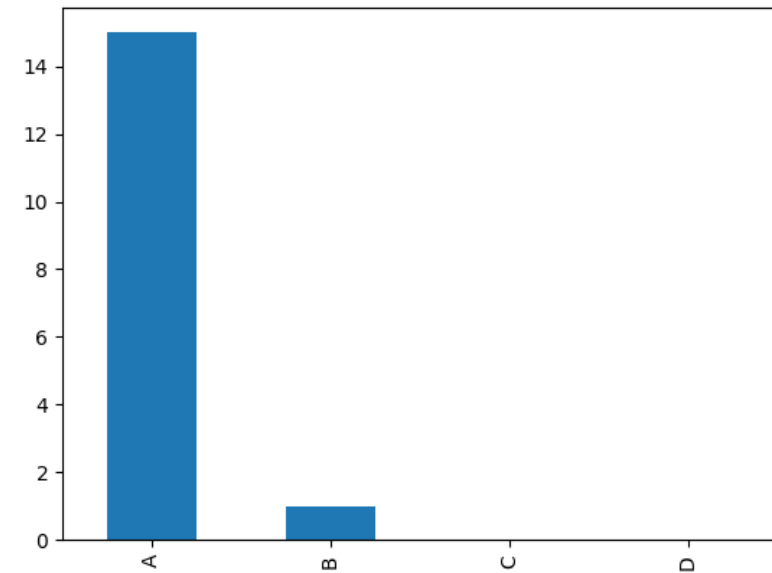
- a. True
- b. False



Mini Test 10

Q 2: Data distribution (normality) matters for choosing the correct statistical test.

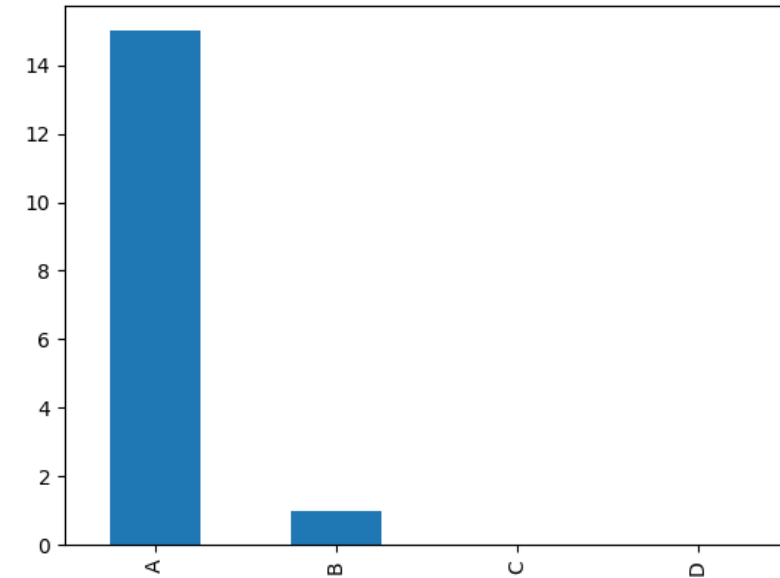
- a. True
- b. False



Mini Test 10

Q 3: Fisher's exact test is preferable over chi-squared test for small n.

- a. True
- b. False



Mini Test 10

Q 4: Which of the following is NOT a correlation coefficient.

- a. Kendall's tau
- b. Pearson's r
- c. Spearman's rank
- d. Bonferroni's r

