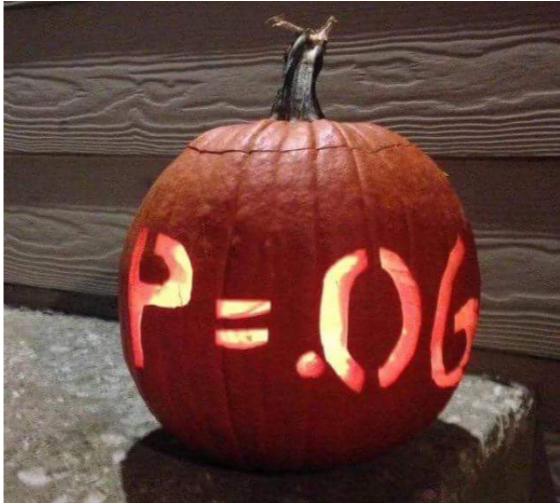


SAY IT WITH NUMBERS: QUANTITATIVE ANALYSIS

ALEXANDER SEREBRENIK



PLEASE TELL US WHAT YOU THINK



RECAP: both Nathan and I will do our best to make this course interesting for you. Please do not wait till the end of the course to provide us feedback.

DATA ANALYSIS

3

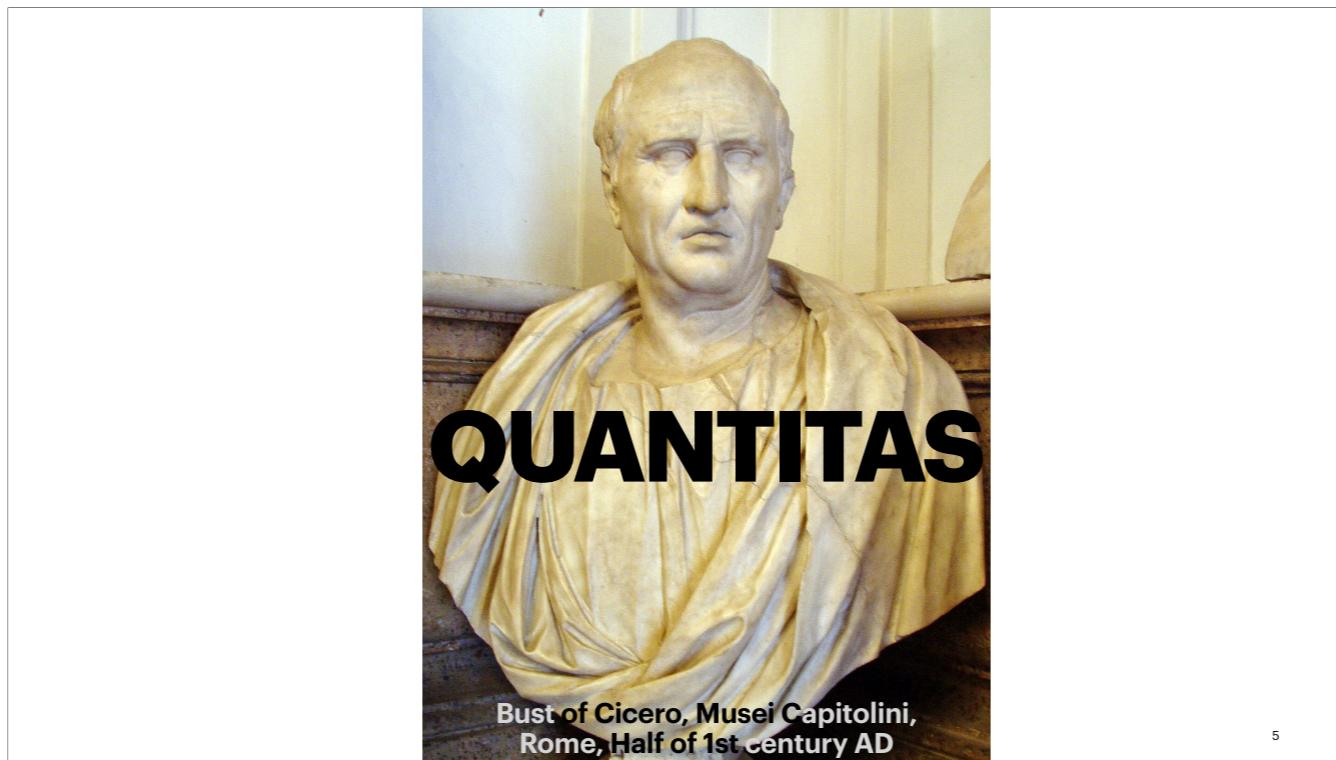
Today we start discussing data analysis techniques. Without data analysis all the data we have been carefully collecting so far will remain useless!

SUMMARY OF THE STUDIES

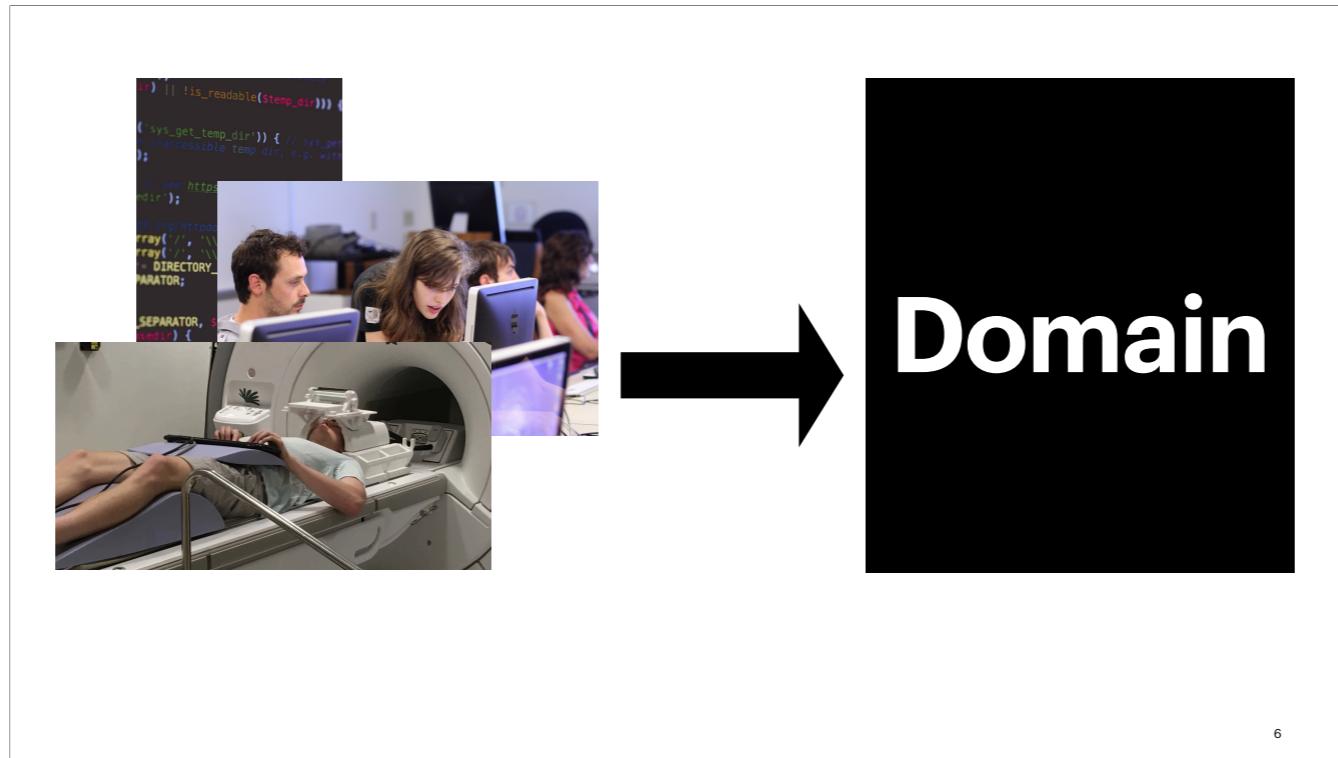
| | App reviews | Good day/ Typical day | Code and prose | Asking for help | Gender and GitHub |
|-----------------|--|--|---|--|--|
| Data source | Apple app store | Experiences of developers | Brain activity of developers | Communication and development | GitHub |
| Data collection | Archival data analysis (Repository mining), sampling | Interviews, surveys | Controlled experiment, post- experiment survey | Ethnography | Archival data analysis (Repository mining), sampling, survey |
| Data analysis | Open coding | Open coding | Statistical analysis, visualisation | Informal | Statistical analysis machine learning |
| Beneficiaries | Developers | Managers | Researchers | Tool builders | Developers, women in particular |
| Recommendation | Focus on the most impactful complaints | Make good days typical and atypical days good | Surveys should be augmented with objective measures | A tool showing who can help and when to contact them | Join projects that use different programming languages |

4

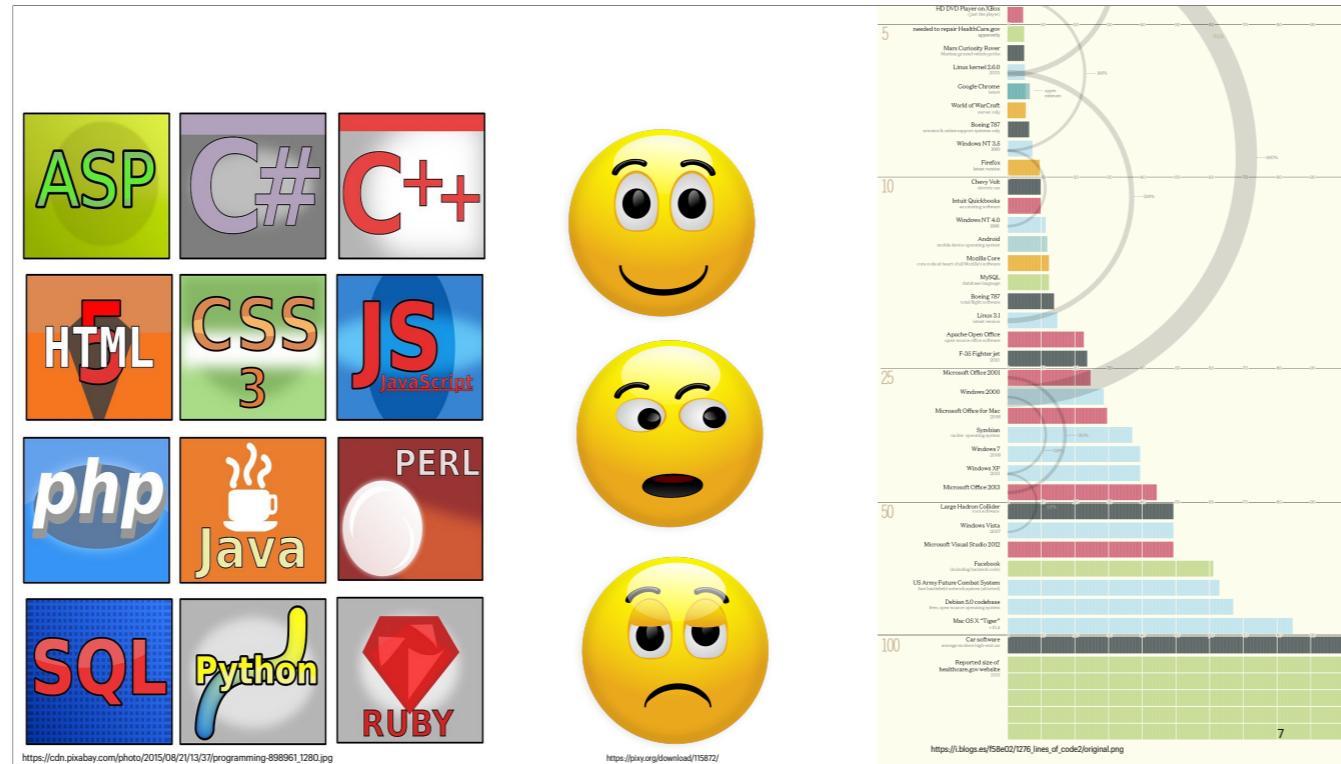
You might remember this slide from the first lecture. As you see there are different types of data analysis: here open coding, statistical analysis, visualisation, machine learning and an informal analysis have been used.



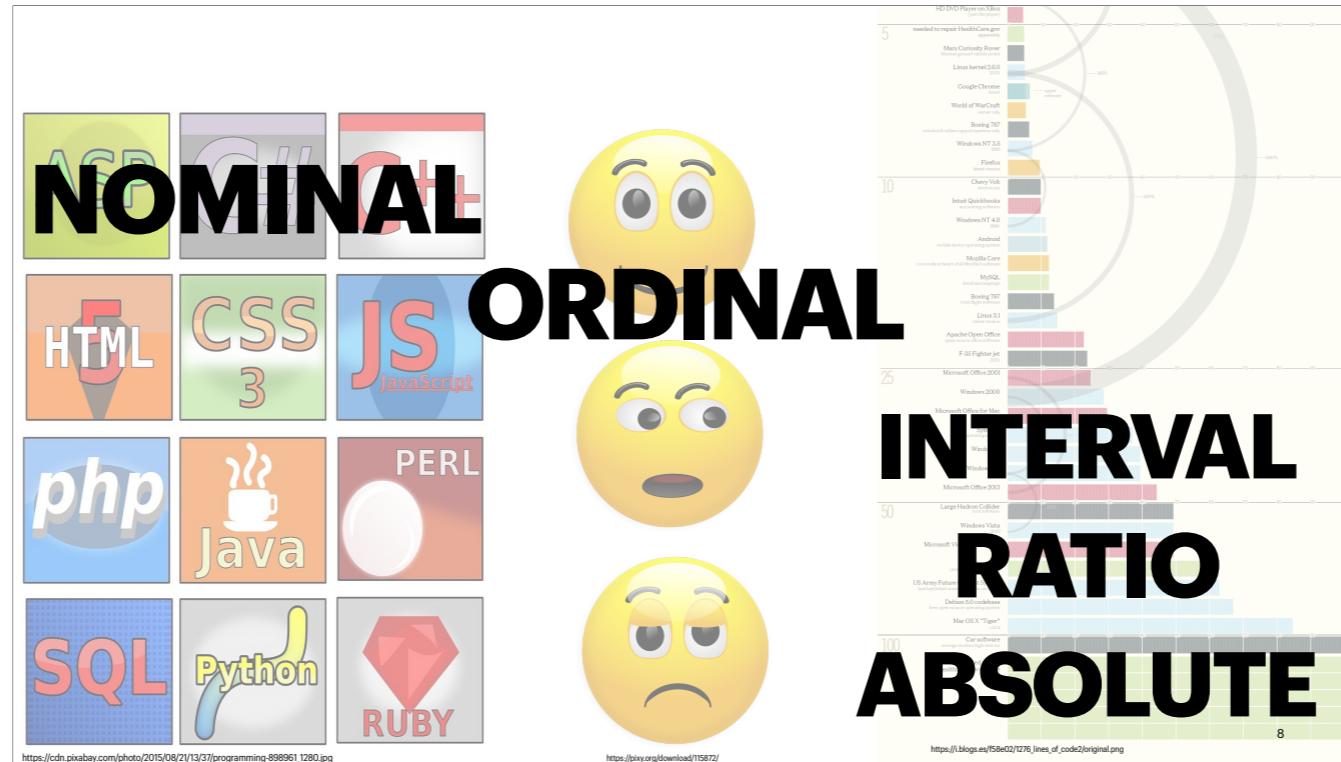
So today we are going to talk about quantitative analysis, and this will also include discussion of several statistical techniques. The word “quantitative” comes from Latin *quantitas*, coined by Cicero /sɪsərōʊ/ SISS-ə-roh from *quantus* (“how much”) + *-tās* (“-ity”). A calque of Ancient Greek ποσότης (posótēs).



Quantitative analysis starts with measuring: we have already seen different kinds of measures, such as interview or survey questions, or information extracted from software repositories. These measures come in multiple shapes and one of the most important differences is the domain. Domain is, in general, a collection of values that we can count: do not forget that *quantus* means “how much”. Beware: the values can be numerical but are not necessarily so.



Here are examples of three domains: programming languages, positive/neutral/negative score from the sentiment analysis experiment, and sizes of different codebases in million lines of code.

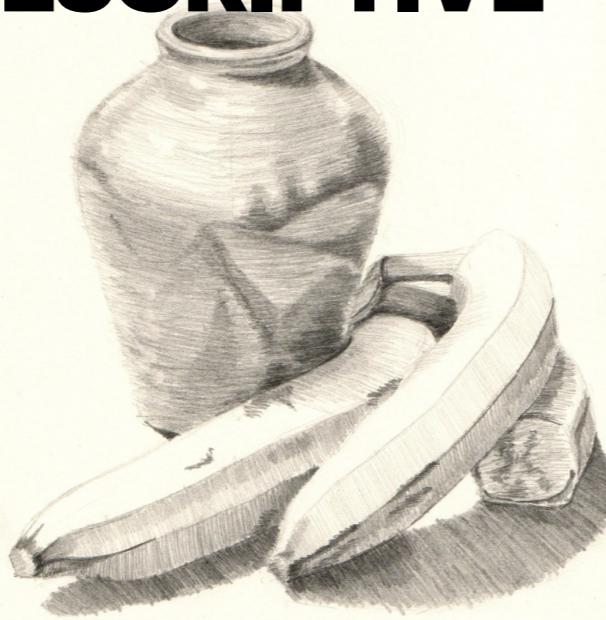


Nominal domain (or nominal scale) boils down to labels. We can only say that two programming languages are the same or different, there is no “natural” order.

Ordinal domain (or ordinal scale) induces an ordering over the values. This ordering is somehow natural: we know that neutral is between negative and positive, and “disagree” is between “strongly disagree” and “mildly disagree”. However, we cannot argue that the “distance” between negative and neutral is the same as between neutral and positive. There is no “distance”.

Interval, ratio, absolute are numbers. There is a difference between the three. Interval has the notion of “distance” and ratio has not only a “distance” but also a clear zero and a clear 100% or 1. For example % of women in a project or % of commits performed by Alice. Absolute is an absolute measurement, for example, the number of developers. However, for many practical purposes, interval, ratio and absolute scales are “just numbers”, and the analyses we perform usually will not require distinguishing between these three scales.

DESCRIPTIVE



[https://cspushe.com/photos/4e163/still-life_bananas_vase_drawing_pencil_art_artwork-751451.jpg#l](https://cspushe.com/photos/4e163/still-life_bananas_vase_drawing_pencil_art_artwork-751451.jpg)

INFERENTIAL



The next question is what do we want to do with our measurements: do we merely want to report what we see, i.e., describe the two and half bananas that we have observed during the study, or do we want to derive conclusions about the broader context, i.e., bananas in general. Recall our discussion of sampling: we study a small group of objects or people (sample) with the intention of inferring conclusions about a much larger group of objects or people (population). This is why most of the time we are interested in inferential statistics. However, before performing inferences we tend to describe the sample we have taken.

QUESTION

Debian 3.0 (published two years after Debian 2.2), groups 4,579 packages of source code with almost 105 MSLOC.

- (A) ORDINAL, DESCRIPTIVE**
- (B) ORDINAL, INFERENTIAL**
- (C) NUMERICAL (ABSOLUTE), DESCRIPTIVE**
- (D) NUMERICAL (ABSOLUTE), INFERENTIAL**

GREGORIO ROBLES, JESÚS M. GONZÁLEZ-BARAHONA, MARTÍN MICHL MAYR, JUAN JOSÉ AMOR MINING LARGE SOFTWARE COMPILED OVER TIME: ANOTHER PERSPECTIVE OF SOFTWARE EVOLUTION. MSR 2006: 3-9

10

C - numerical, descriptive

| | Nominal | Ordinal | Interval, Ratio, Absolute |
|-------------|---------|---------|------------------------------|
| Descriptive | | | |
| Inferential | | | |

11

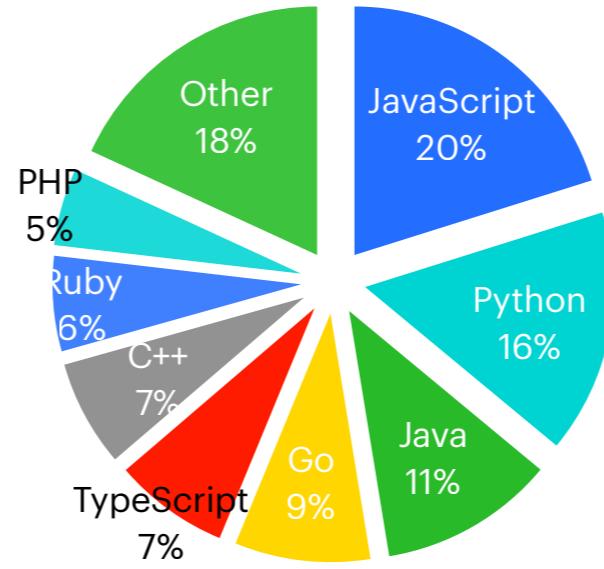
In the remainder of the lecture we are going to fill in the following table.

NOMINAL



https://cdn.pixabay.com/photo/2015/08/21/13/57/programming-898961_1280.jpg

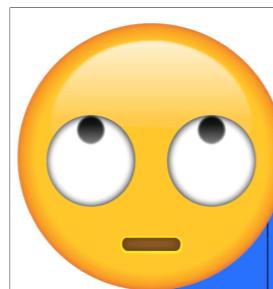
% PULL REQUESTS ON GITHUB IN JULY-SEPT 2020



https://madnight.github.io/github/#/pull_requests/2020/3

12

When it comes to describing nominal data one most often uses frequencies and percentages. In addition to pie-charts of any kind nominal data is often represented as a table with the ordering induced by frequency.



| | Anger | Disgust | Sadness | Surprise | Lack of Awareness | Lack of Emotion | Not Possible | Other |
|-------|-------|---------|---------|----------|-------------------|-----------------|--------------|-------|
| Men | 47 | 24 | 9 | 24 | 17 | 65 | 25 | 13 |
| Women | 53 | 19 | 5 | 13 | 10 | 32 | 7 | 9 |

WESLEY BRANTS, BONITA SHARIF, ALEXANDER SEREBRENIK: ASSESSING THE MEANING OF EMOJIS FOR EMOTIONAL AWARENESS - A PILOT STUDY. WWW (COMPANION VOLUME) 2019: 419-423 ¹³

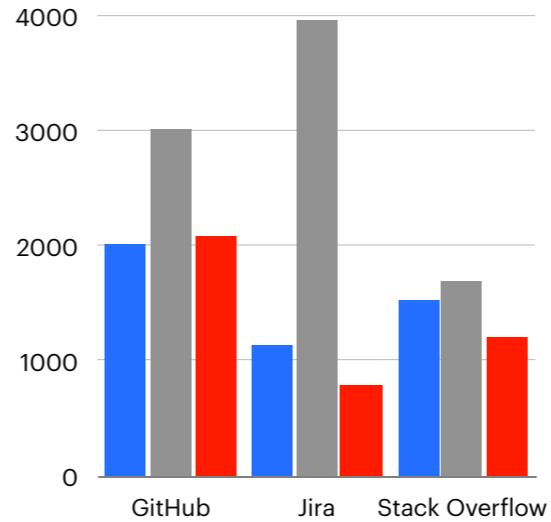
If we have two nominal variables it is a common practice to present a table with the two variables corresponding to rows and columns. This kind of tables are known as contingency tables, cross-tabulation or shorter crosstabs.

In this study we have asked an open question “What emotion is this emoji displaying?”.
 There have been no non-binary participants

ORDINAL



Positive Neutral Negative



14

Pretty much the same happens for the ordinal scale. The difference is, however, is that since the ordinal scale is ordered so in addition to pie charts and tables, bar charts are quite common. These are three popular datasets for sentiment analysis (these are not complete datasets of GitHub, Jira or Stack Overflow, but datasets extracted from these websites).

| | | SentiStrength | | |
|------|----------|---------------|---------|----------|
| | | Negative | Neutral | Positive |
| NLTK | Negative | 17 | 39 | 25 |
| | Neutral | 15 | 96 | 34 |
| | Positive | 6 | 20 | 43 |

ROBBERT JONGELING, SUBHAJIT DATTA, ALEXANDER SEREBRENIK: CHOOSING YOUR WEAPONS: ON SENTIMENT ANALYSIS TOOLS FOR SOFTWARE ENGINEERING RESEARCH. ICSME 2015: 531-535⁵

Here we see two different sentiment analysis tools (SentiStrength and NLTK) that are used to assess the same theoretical construct (polarity of the sentiment). In this case we can also talk about agreement **between** the tools. Please note that agreement can be determined for nominal variables as well, as long as both variables are supposed to measure the same theoretical construct!

| | | SentiStrength | | |
|------|----------|---------------|---------|----------|
| | | Negative | Neutral | Positive |
| NLTK | Negative | 17 | 39 | 25 |
| | Neutral | 15 | 96 | 34 |
| | Positive | 6 | 20 | 43 |

$$p_o = \frac{17 + 96 + 43}{17 + 39 + 25 + 15 + 96 + 34 + 6 + 20 + 43} \approx 53\%$$

16

ROBBERT JONGELING, SUBHAJIT DATTA, ALEXANDER SEREBRENIK: CHOOSING YOUR WEAPONS: ON SENTIMENT ANALYSIS TOOLS FOR SOFTWARE ENGINEERING RESEARCH. ICSME 2015: 531-535

Observed agreement is percentage of data points where the tools agree.

The problem is however that observed agreement can be high merely by chance. If we toss two fair coins, then the observed agreement will be 50% but this does not say much.

So how much agreement would we expect if the agreement was only due to chance, i.e., the evaluation of NLTK and SentiStrength would have been completely independent?

In total 295 data points

| | | SentiStrength | | | $p_o \simeq 53\%$ |
|------|----------|---------------|---------|----------|-------------------|
| | | Negative | Neutral | Positive | |
| NLTK | Negative | 17 | 39 | 25 | |
| | Neutral | 15 | 96 | 34 | |
| | Positive | 6 | 20 | 43 | |

$p_e = p_{neg} + p_{neu} + p_{pos}$

$$p_{neg} = \frac{17 + 39 + 25}{17 + 39 + 25 + 15 + 96 + 34 + 6 + 20 + 43} \times \frac{17 + 15 + 6}{17 + 39 + 25 + 15 + 96 + 34 + 6 + 20 + 43} \simeq 3.5\%$$

17

ROBBERT JONGELING, SUBHAJIT DATTA, ALEXANDER SEREBRENIK: CHOOSING YOUR WEAPONS: ON SENTIMENT ANALYSIS TOOLS FOR SOFTWARE ENGINEERING RESEARCH. ICSME 2015: 531-535

To compute the expected agreement we consider NLTK and SentiStrength to be independent. For example, for the expected agreement on the negative class we take the percentage of negative values computed by SentiStrength and multiply it by the percentage of negative values computed by NLTK.

| | | SentiStrength | | | $p_o \simeq 53\%$ |
|------|----------|---------------|---------|----------|-------------------|
| | | Negative | Neutral | Positive | |
| NLTK | Negative | 17 | 39 | 25 | |
| | Neutral | 15 | 96 | 34 | |
| | Positive | 6 | 20 | 43 | |

$p_e = p_{neg} + p_{neu} + p_{pos}$

$p_{neg} \simeq 3.5\% \quad p_{neu} \simeq 26\% \quad p_{pos} \simeq 8\%$

$p_e \simeq 37.5\%$

ROBBERT JONGELING, SUBHAJIT DATTA, ALEXANDER SEREBRENIK: CHOOSING YOUR WEAPONS: ON SENTIMENT ANALYSIS TOOLS FOR SOFTWARE ENGINEERING RESEARCH. ICSME 2015: 531-535

18

Continuing in the same way we get the expected agreement for the two remaining classes. Ultimately the expected agreement is the sum of the agreement for the three classes.

| | | SentiStrength | | | |
|------|----------|---------------|---------|----------|--|
| | | Negative | Neutral | Positive | |
| NLTK | Negative | 17 | 39 | 25 | |
| | Neutral | 15 | 96 | 34 | |
| | Positive | 6 | 20 | 43 | |

$p_o \simeq 53\%$
 $p_e \simeq 37.5\%$
 $\kappa = \frac{p_o - p_e}{1 - p_e} \simeq 24.8\%$

| no | slight | fair | moderate | substantial | almost perfect |
|----------|--------|--------|----------|-------------|----------------|
| ≤ 0 | 1-20% | 21-40% | 41-60% | 61-80% | 81-100% |

ROBBERT JONGELING, SUBHAJIT DATTA, ALEXANDER SEREBRENIK: CHOOSING YOUR WEAPONS: ON SENTIMENT ANALYSIS TOOLS FOR SOFTWARE ENGINEERING RESEARCH. ICSME 2015: 531-535

19

To compare the observed agreement and the expected agreement we compute the so-called Cohen's kappa. In our case the agreement is merely **fair**...

QUESTION

| | | B | |
|---|-----|-----|----|
| | | Yes | No |
| A | Yes | 45 | 15 |
| | No | 25 | 15 |

(A) $\kappa \simeq 28\%$

(B) $\kappa \simeq 7\%$

(C) $\kappa \simeq -11\%$

(D) $\kappa \simeq 13\%$

20

$$p_o = (45+15)/(45+15+25+15) = 0.6$$

$$p_e = p_A + p_B$$

$$p_A = 60/100 * 70/100 = 0.42$$

$$p_B = 40/100 * 30/100 = 0.12$$

$$p_e = 0.54$$

$$\kappa = (0.6 - 0.54)/(1 - 0.54) = 0.06/0.46 = 3/23 \sim 13\% \Rightarrow D$$

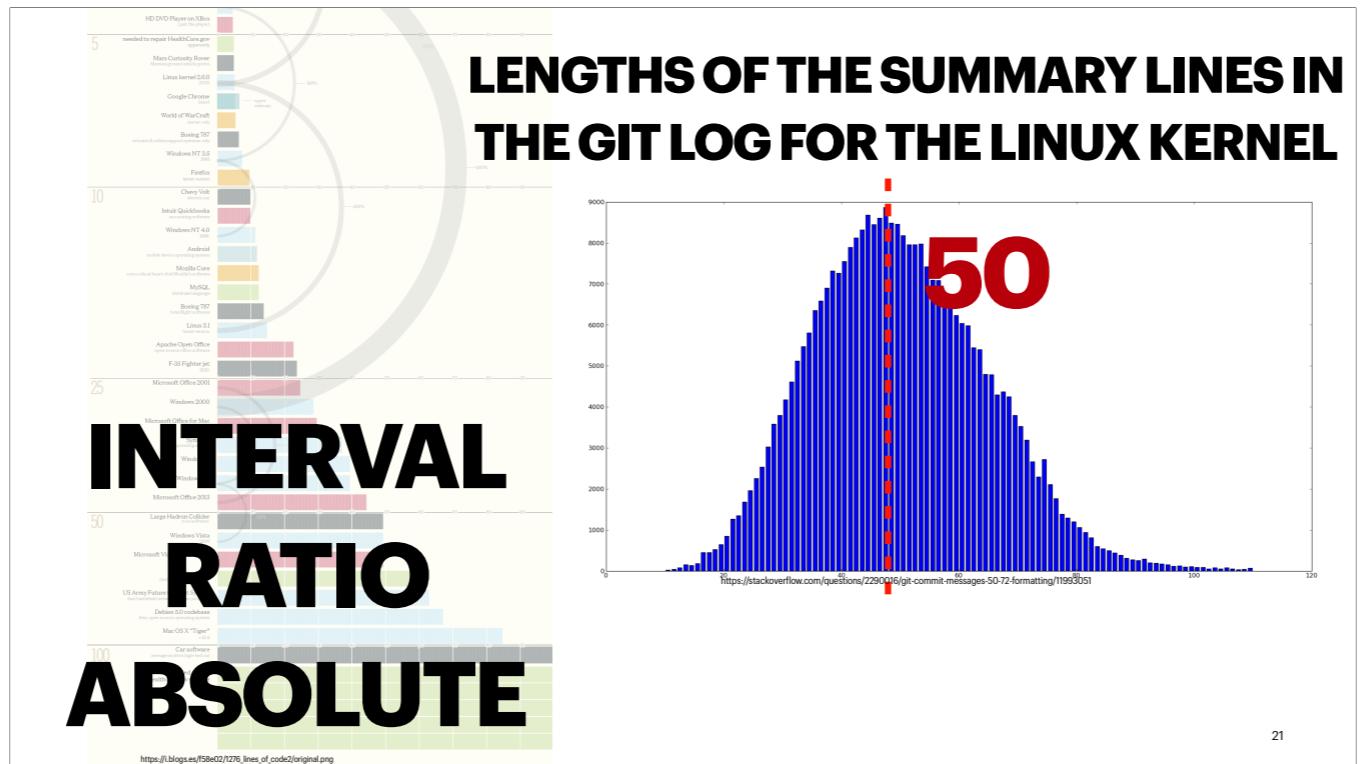
15 45

15 25

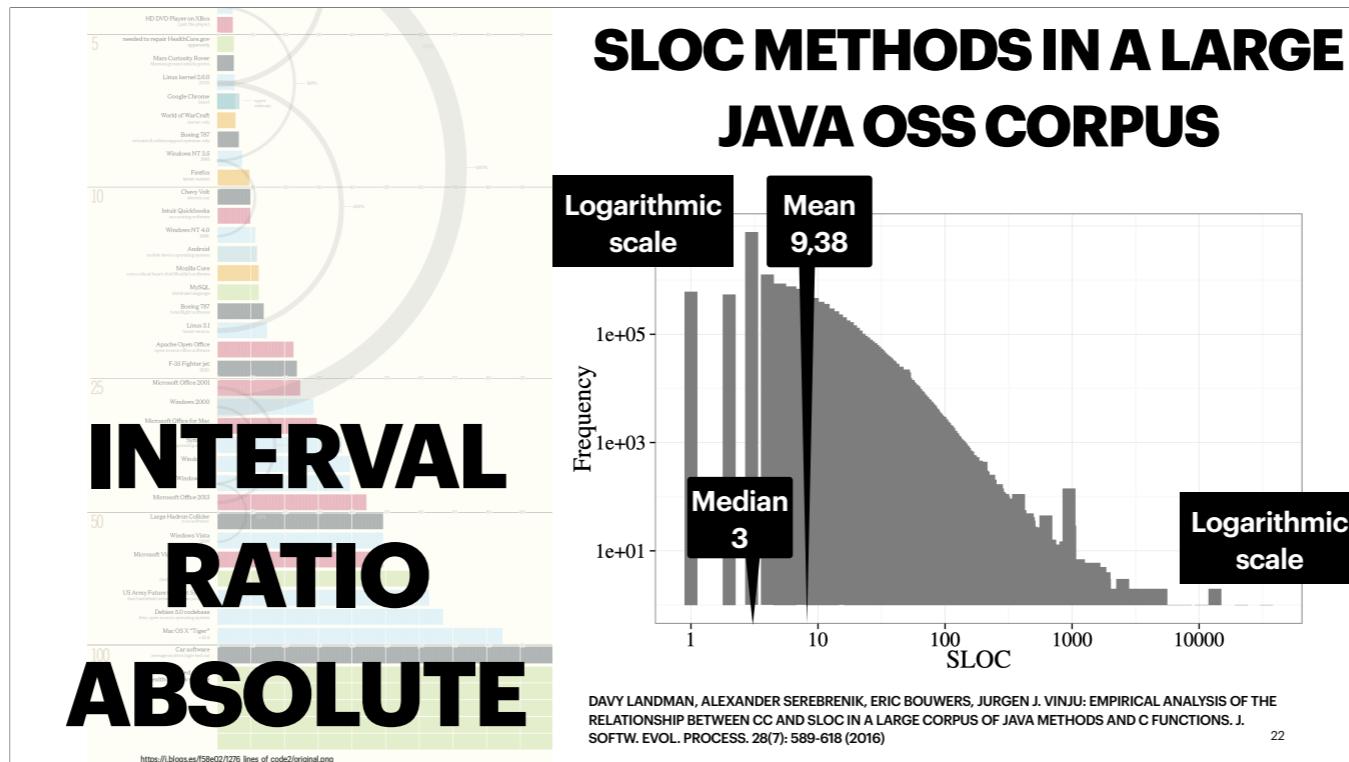
$$p_o = 0.4$$

$$p_e = 30/100 * 60/100 + 40/100 * 70/100 = 0.18 + 0.28 = 0.46$$

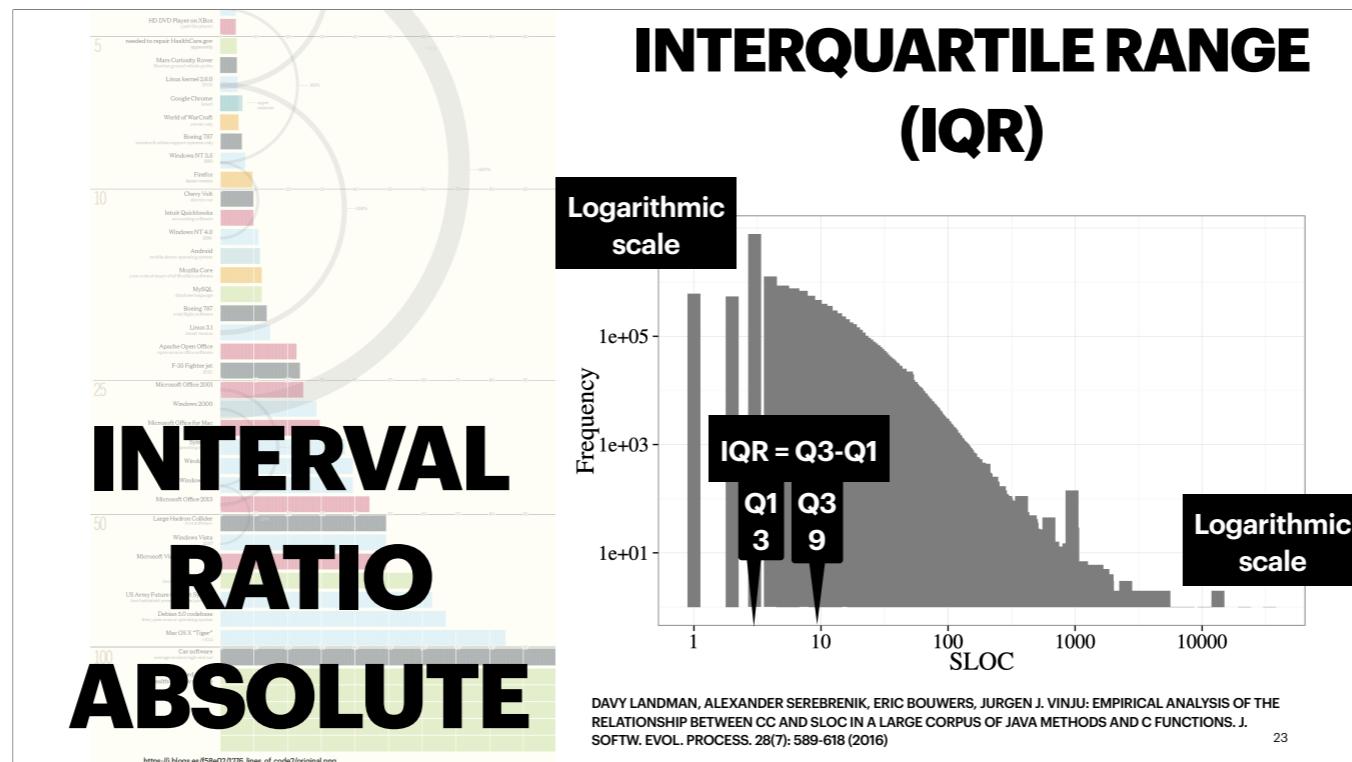
$$\kappa = (0.4 - 0.46)/(1 - 0.46) = -0.06/0.54$$



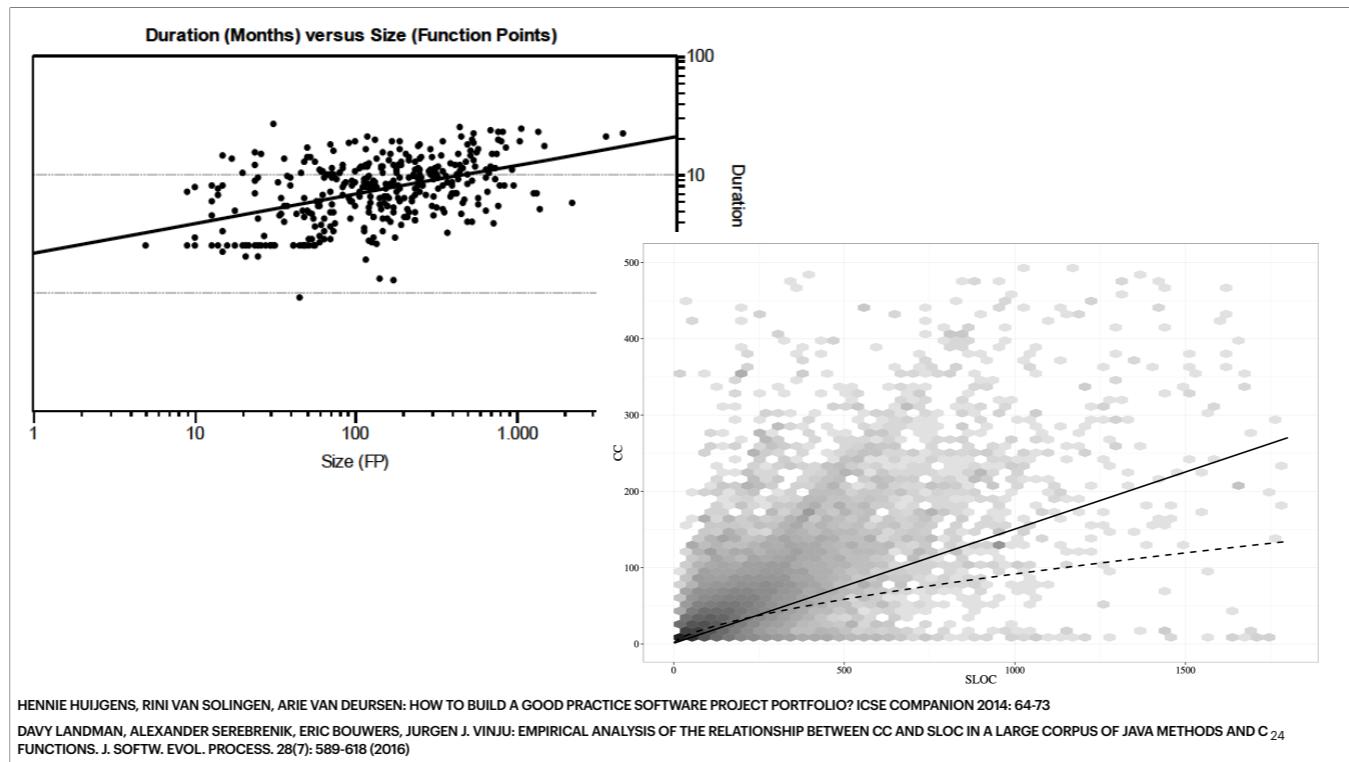
Finally, when it comes to numerical values, one can compute means and standard deviations. Here we see a distribution that looks symmetric and close to “normal”.



However, most software engineering measurements do not look nice and symmetric. Here we see the distribution of SLOC per method in a large corpus of open-source Java programs. Please note the log-log scale, i.e., logarithmic scale on both axes). This means that the mean is strongly affected by presence of very large values and does not adequately represent central tendency. Median is a more adequate measure in this case. For the same argument, standard deviation is not reliable, a better option is IQR.

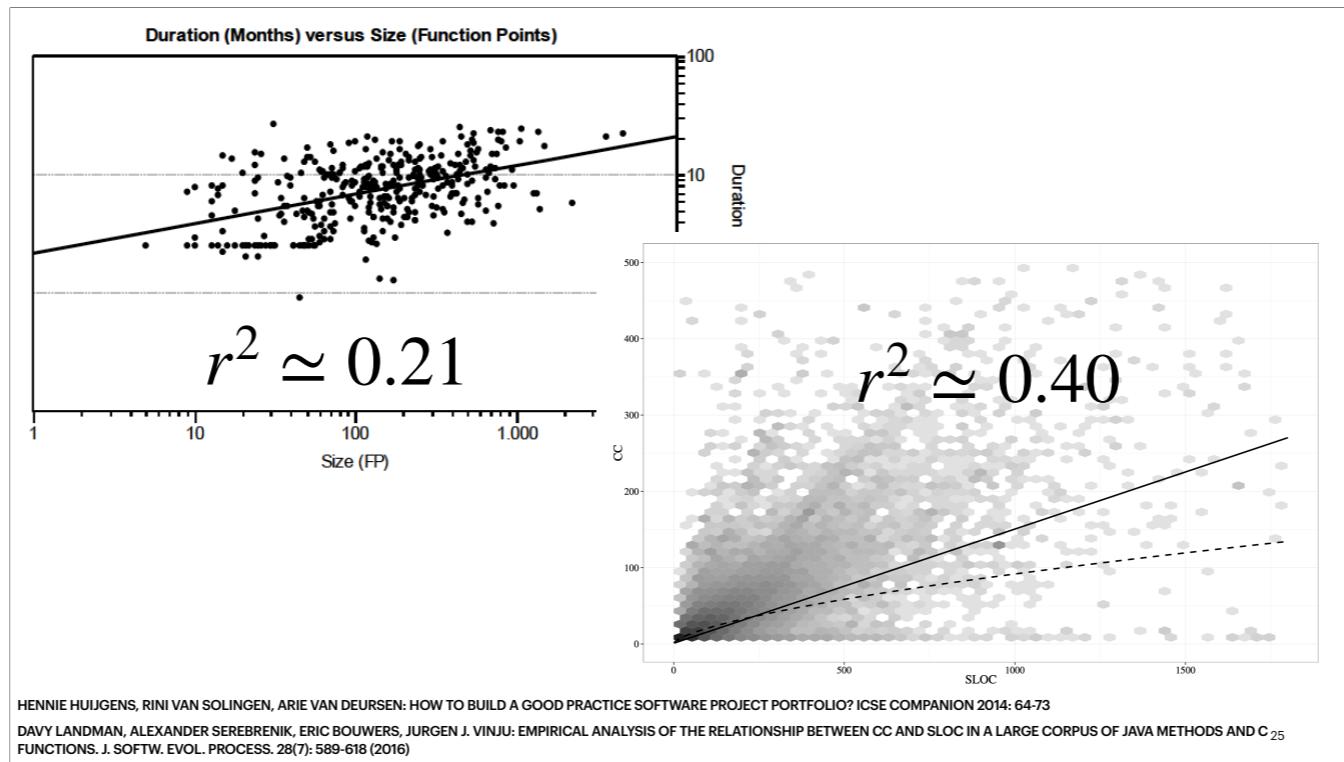


Interquartile range (IQR) is the difference between Q3 and Q1. Q1 is the value such that 25% of the values do not exceed it; Q3 is the value such that 75% of the values do not exceed it. This means that between Q1 and Q3 there are 50% of the values.

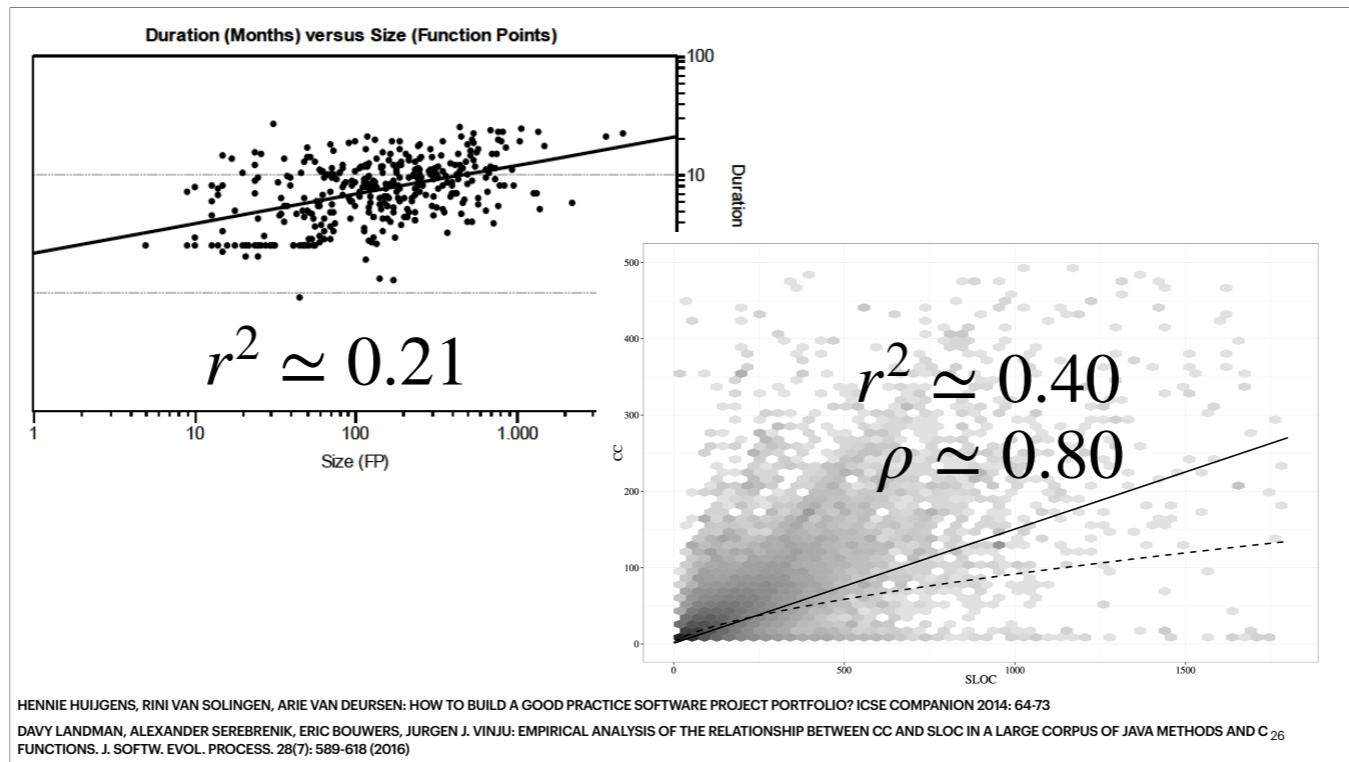


Similarly, to the sentiment analysis tools and variables on the ordinal scale, if we have two numeric variables we might like to evaluate their “agreement”. A common way of doing so is to show a scatter plot. On the left we see a scatter plot showing the amount of functionality in the system on the x-axis and duration of the project on the y-axis, please note the logarithmic scale on both axes. The problem with the traditional scatter plot is that multiple data points can overlap, and one cannot really see different points. A slightly better solution is a hexagon plot on the right where every hexagon represents multiple data points with close values; the colour of the hexagon represents the number of data points.

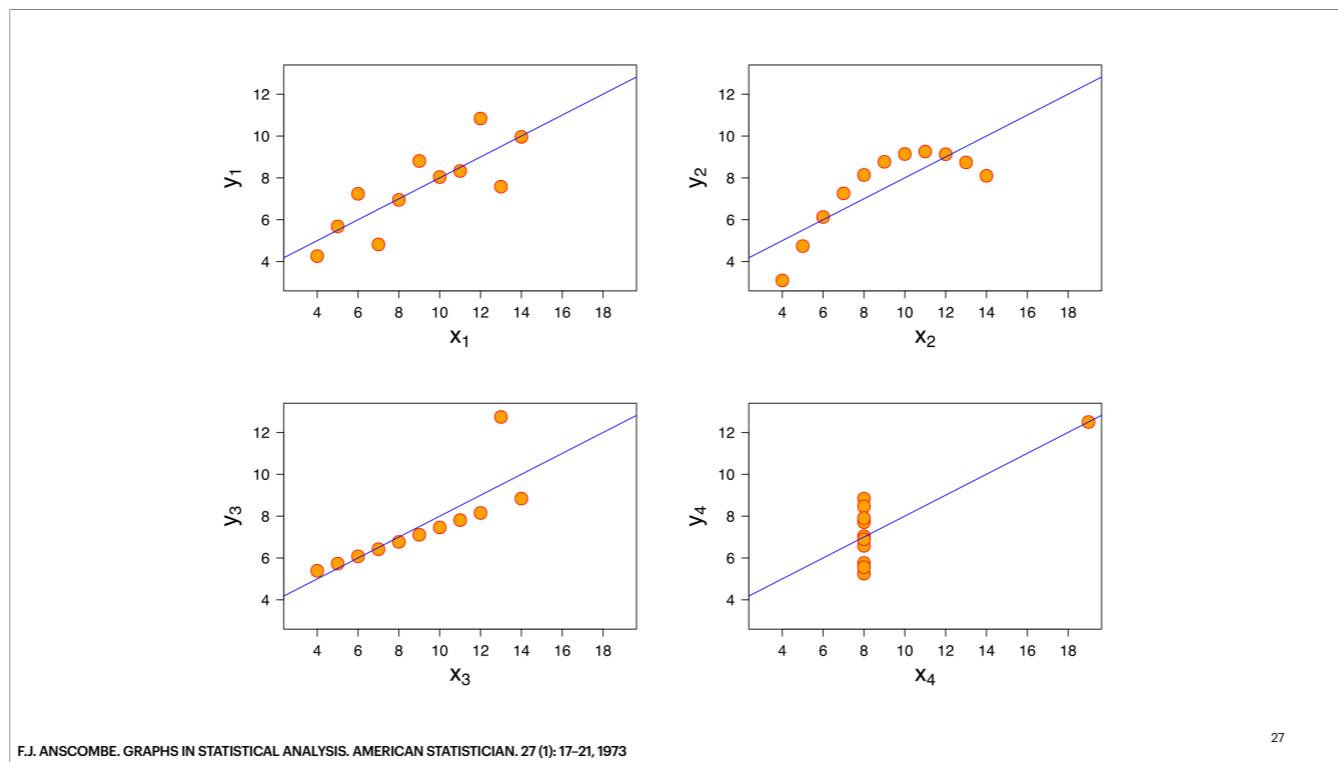
Please ignore the lines on for the moment.



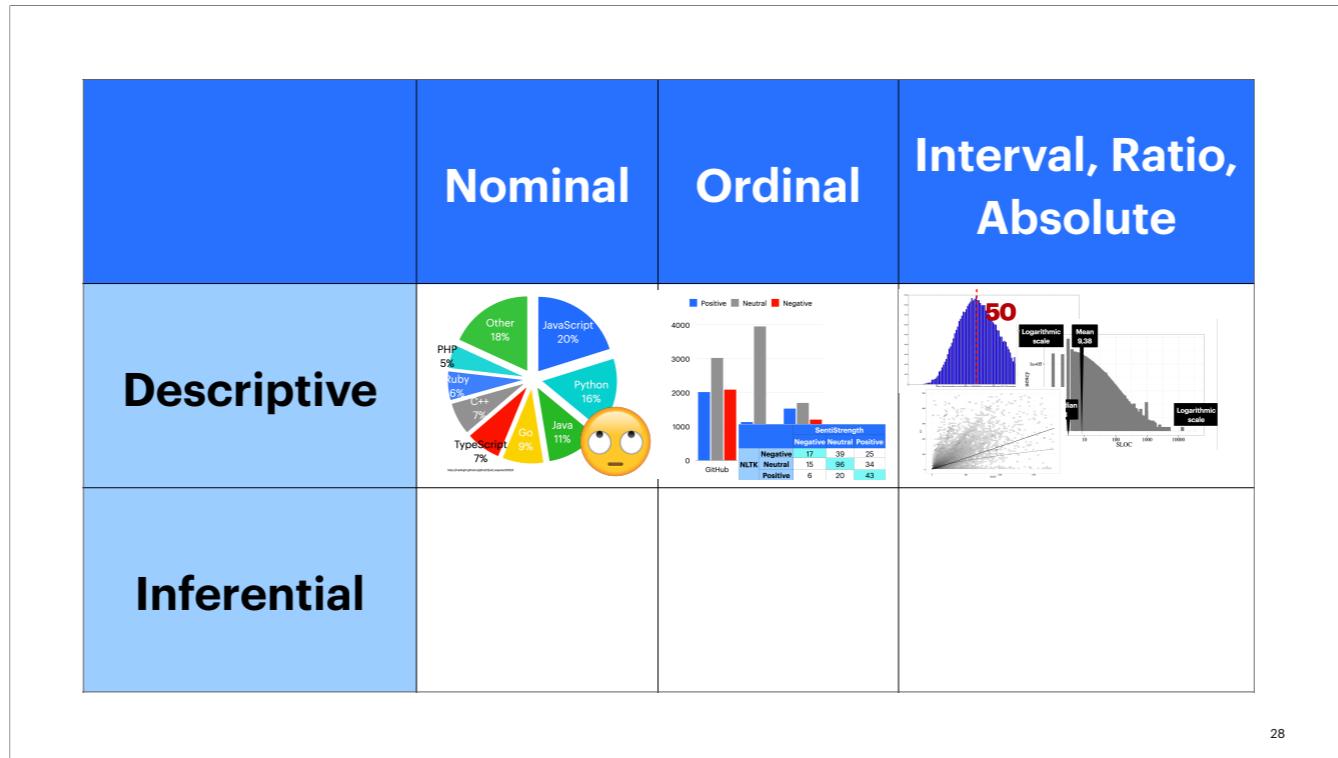
Furthermore, one can characterise this “agreement” numerically. This measure is known as correlation, and the Pearson product-moment correlation coefficient (PPMCC). I hope that you have seen it in your basic statistics course but just to remind you, the Pearson coefficient measures **linear correlation** between two variables X and Y. It has a value between +1 and -1. People often report r^2 instead of r itself.



If there is no reason to assume that the relation is linear, then one can (a) transform one or both variables using logarithm and then checking whether the relation would be meaningfully linear, and (b) use a different correlation coefficient, Spearman's ρ or Kendall's τ .

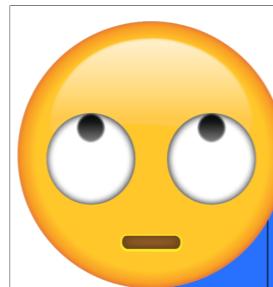


Be careful, however - merely computing the Pearson correlation is not enough. One has to inspect the scatterplot. These four examples have been constructed by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers and other influential observations on statistical properties. The Pearson's correlation coefficient is almost the same - to 3 decimal places - for all four examples.



28

This what we have seen so far: descriptive statistics. The next step is inferential statistics. The focus of inferential statistics is on inferring properties of a population based on a sample, for example by testing hypotheses and deriving estimates. There are numerous tests, and it is clearly not possible to discuss all of them during this lecture. I would recommend you to check advanced books on data analysis to decide which statistical test to use. Whatever test you have decided to carry out, it is important to understand what kind of assumptions each one of the tests makes.



| | Anger | Disgust | Sadness | Surprise | Lack of Awareness | Lack of Emotion | Not Possible | Other |
|-------|-------|---------|---------|----------|-------------------|-----------------|--------------|-------|
| Men | 47 | 24 | 9 | 24 | 17 | 65 | 25 | 13 |
| Women | 53 | 19 | 5 | 13 | 10 | 32 | 7 | 9 |

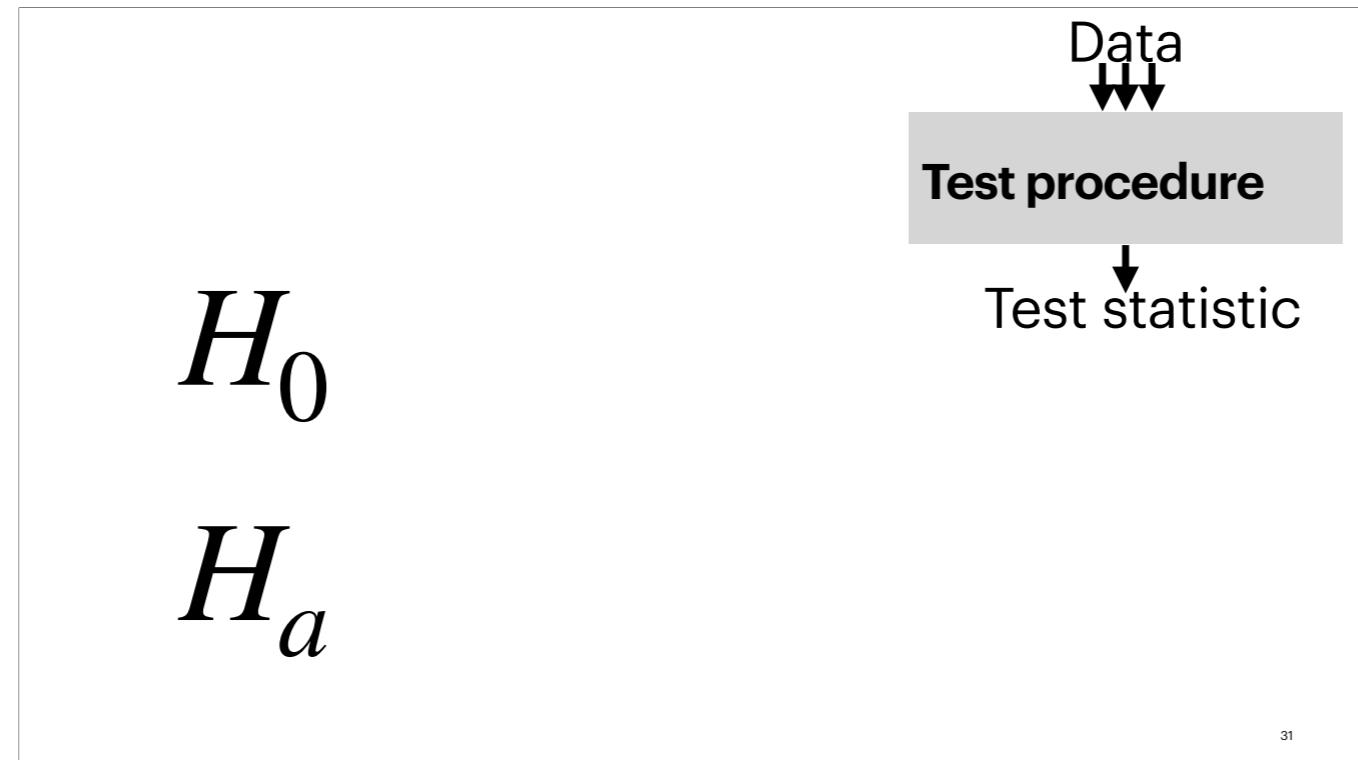
WESLEY BRANTS, BONITA SHARIF, ALEXANDER SEREBRENIK: ASSESSING THE MEANING OF EMOJIS FOR EMOTIONAL AWARENESS - A PILOT STUDY. WWW (COMPANION VOLUME) 2019: 419-423 ²⁹

We have already seen this table before. Inspecting this table we see that 24 men interpret this emoji as disgust vs 19 women, 24 men interpret it as surprise vs 13 women, etc. However, there are more men in total: 224 men vs 148 women. So can we say that this emoji is interpreted different by people of different genders in the population in general?

H_0 H_a

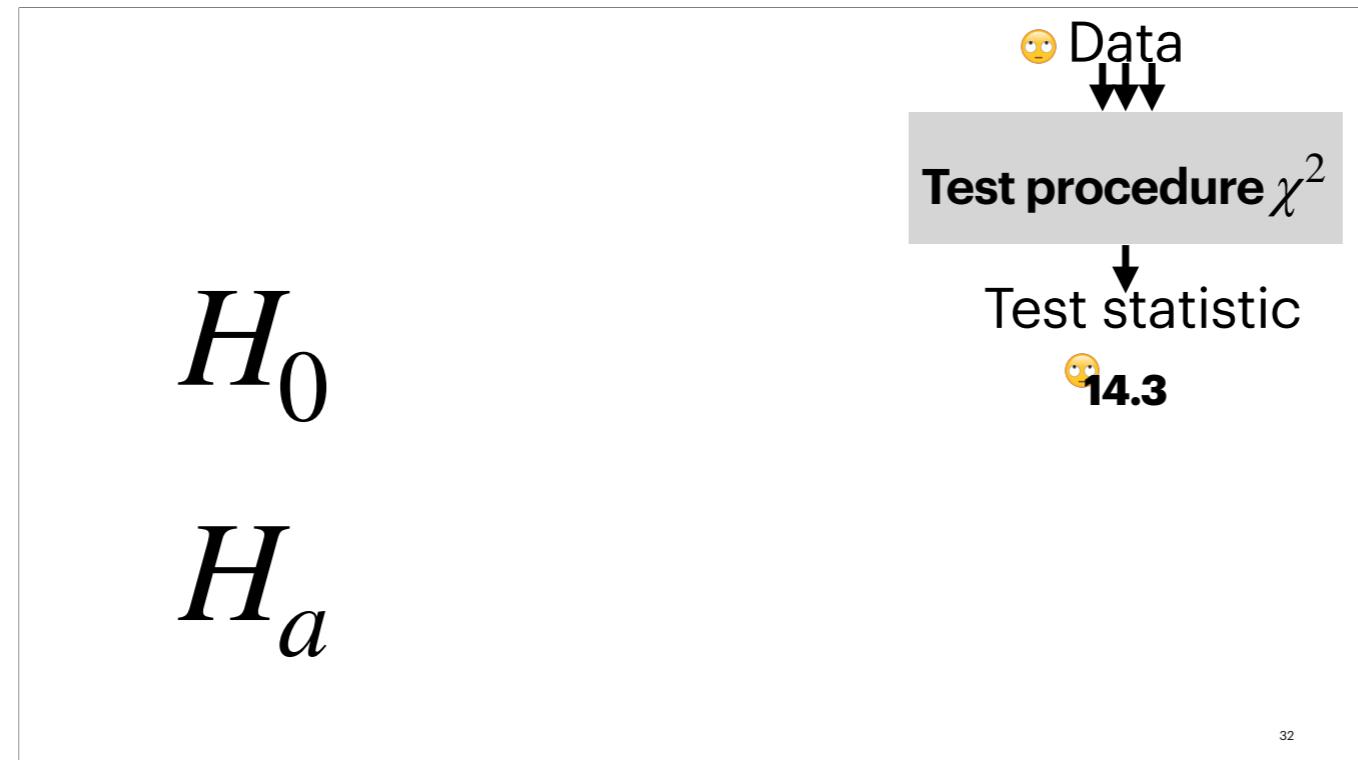
30

To perform statistical testing we need to formulate two hypotheses H_0 and H_a . The null hypothesis is typically the default hypothesis that assumes no relationship between variables in the population from which the sample is selected. A null hypothesis is contrasted with an alternative hypothesis, and the two hypotheses are distinguished on the basis of the sample data. For example, in our case the null hypothesis for this test is that the interpretation of the emoji is independent from the gender. The alternative hypothesis would be that the interpretation of the emoji does depend on the gender, i.e., that there are gender-related differences in interpretation of this emoji.

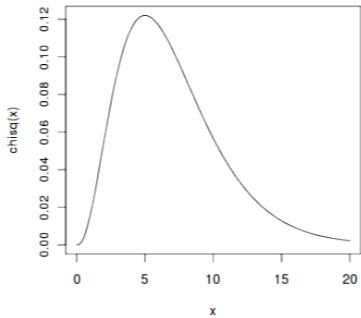


31

Statistical test consists of several elements. One of them is a test procedure that given the data produces a test statistic, a single value that can be used for the actual testing.



In this case our test is the chi-square test of independence. For our example the test statistic is 14.31.

H_0 H_a 

>Data
↓↓

Test procedure χ^2

Test statistic

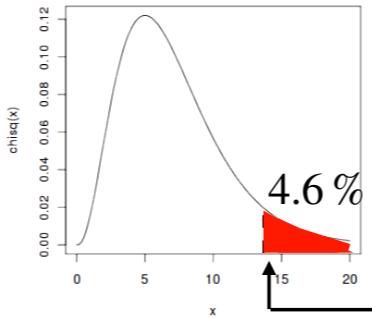
14.3

33

Another element of the statistical test is the distribution: under the assumption that the null hypothesis is true, we know what distributions the test statistic follows. The shape and the parameters of the distribution are specific to the test. In this case the distribution is the chi-square distribution with 7 degrees of freedom (we will not discuss what does this exactly mean).

H_0
 H_a

If H_0 is true, it is unlikely that the test statistic is here



>Data
↓↓

Test procedure χ^2

Test statistic

14.3

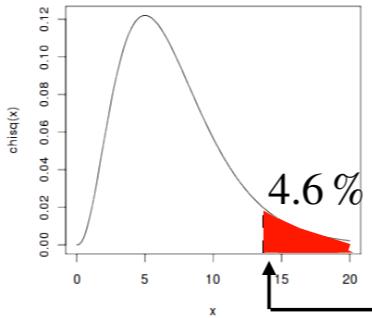
34

Then we check where does the test statistic fall on the distribution plot. The area under the curve to the right of the value of the test statistic is the p-value. This p-value indicates how likely is it the value of the test statistic has been obtained under the assumption that the null hypothesis is true. In our case, this value is 4.6%.

Since in this case the value of the test-statistic is in the 5% range, the p-value is lower than the customary 5% threshold, and we can reject the null hypothesis. In this case this would mean that we can accept the alternative hypothesis that the distribution of the interpretations of the emoji depends on gender.

H_0
 H_a

If H_0 is true, it is unlikely that the test statistic is here

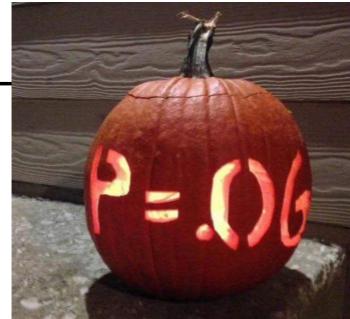


>Data
↓↓

Test procedure χ^2

Test statistic

14.3



35

Since in this case the value of the test-statistic is in the 5% range, the p-value is lower than the customary 5% threshold, and we can reject the null hypothesis. In this case this would mean that we can accept the alternative hypothesis that the distribution of the interpretations of the emoji depends on gender.

5% is the common threshold, so if the p-value of 6% is the nightmare of a quantitative researcher :)

QUESTION

A p-value indicates...

**(A) THE PROBABILITY THAT THE
NULL HYPOTHESIS IS TRUE**

**(B) THE PROBABILITY OF
OBTAINING THE RESULTS (OR ONE
MORE EXTREME) IF THE NULL
HYPOTHESIS IS TRUE**

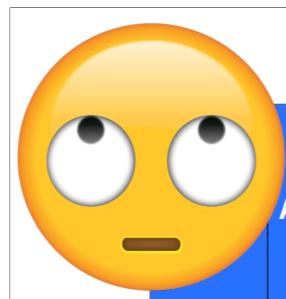
**(C) THE PROBABILITY THAT THE
ALTERNATIVE HYPOTHESIS IS TRUE**

**(D) THE PROBABILITY OF
OBTAINING THE RESULTS (OR ONE
MORE EXTREME) IF THE
ALTERNATIVE HYPOTHESIS IS TRUE**

BUT

37

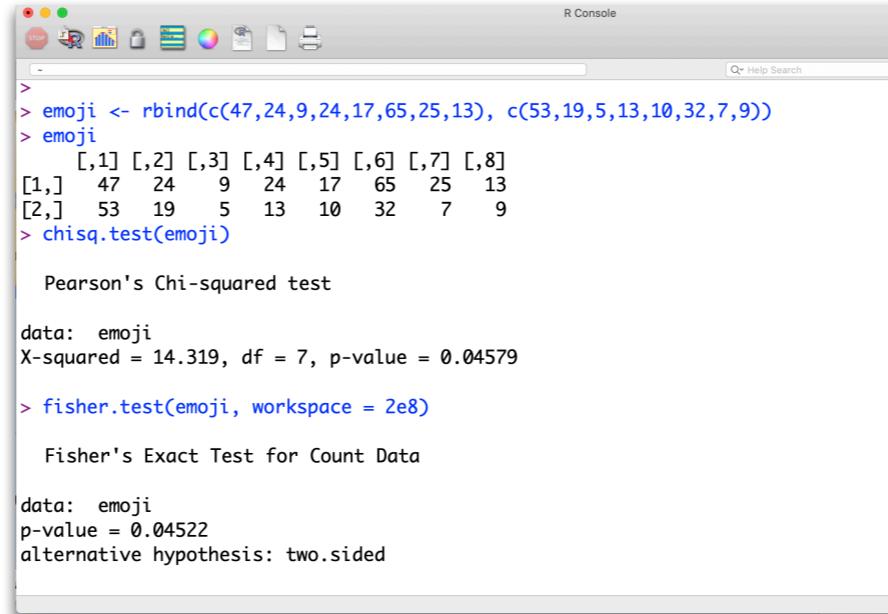
Do you remember that I stressed the importance of checking the assumptions of the test?



| | Anger | Disgust | Sadness | Surprise | Lack of Awareness | Lack of Emotion | Not Possible | Other |
|-------|-------|---------|---------|----------|-------------------|-----------------|--------------|-------|
| Men | 47 | 24 | 9 | 24 | 17 | 65 | 25 | 13 |
| Women | 53 | 19 | 5 | 13 | 10 | 32 | 7 | 9 |

WESLEY BRANTS, BONITA SHARIF, ALEXANDER SEREBRENIK: ASSESSING THE MEANING OF EMOJIS FOR EMOTIONAL AWARENESS - A PILOT STUDY. WWW (COMPANION VOLUME) 2019: 419-423 38

One of the assumptions of the chi-square test is that the values in the cells should be 10 or more (and the expected values should be 5 or more). If this assumption is violated the chi-square test is imprecise. There is an alternative, Fisher's exact test. However, Fisher's test is computationally expensive and traditionally has only been applied for small tables and small numbers. Luckily, we have computers today. Ideally, the results of the hypothesis testing for the chi-square test and Fisher's exact test should be the same.



The screenshot shows an R console window with the following content:

```

> emoji <- rbind(c(47,24,9,24,17,65,25,13), c(53,19,5,13,10,32,7,9))
> emoji
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]   47   24    9   24   17   65   25   13
[2,]   53   19    5   13   10   32    7    9
> chisq.test(emoji)

Pearson's Chi-squared test

data: emoji
X-squared = 14.319, df = 7, p-value = 0.04579

> fisher.test(emoji, workspace = 2e8)

Fisher's Exact Test for Count Data

data: emoji
p-value = 0.04522
alternative hypothesis: two.sided

```

39

This is R, my favourite statistical package: it is open source, has a strong community surrounding it, meaning that lots and lots of statistical tests are implemented. If you do not like R, you can always use other statistical packages such as SPSS, or libraries such as scikit of Python.

As you see, the results of both tests are very similar: 4.58% for the chi-square test and 4.52% for the Fisher's exact test. This gives us more confidence that our conclusion about gender-related differences in the interpretation of the Face With Rolling Eyes emoji.

Another point here is “workspace = 2e8”. As opposed to chi-square that immediately produces the results, for Fisher's test I had to allocate extra memory and wait a couple of seconds for the p-value to be computed.



40

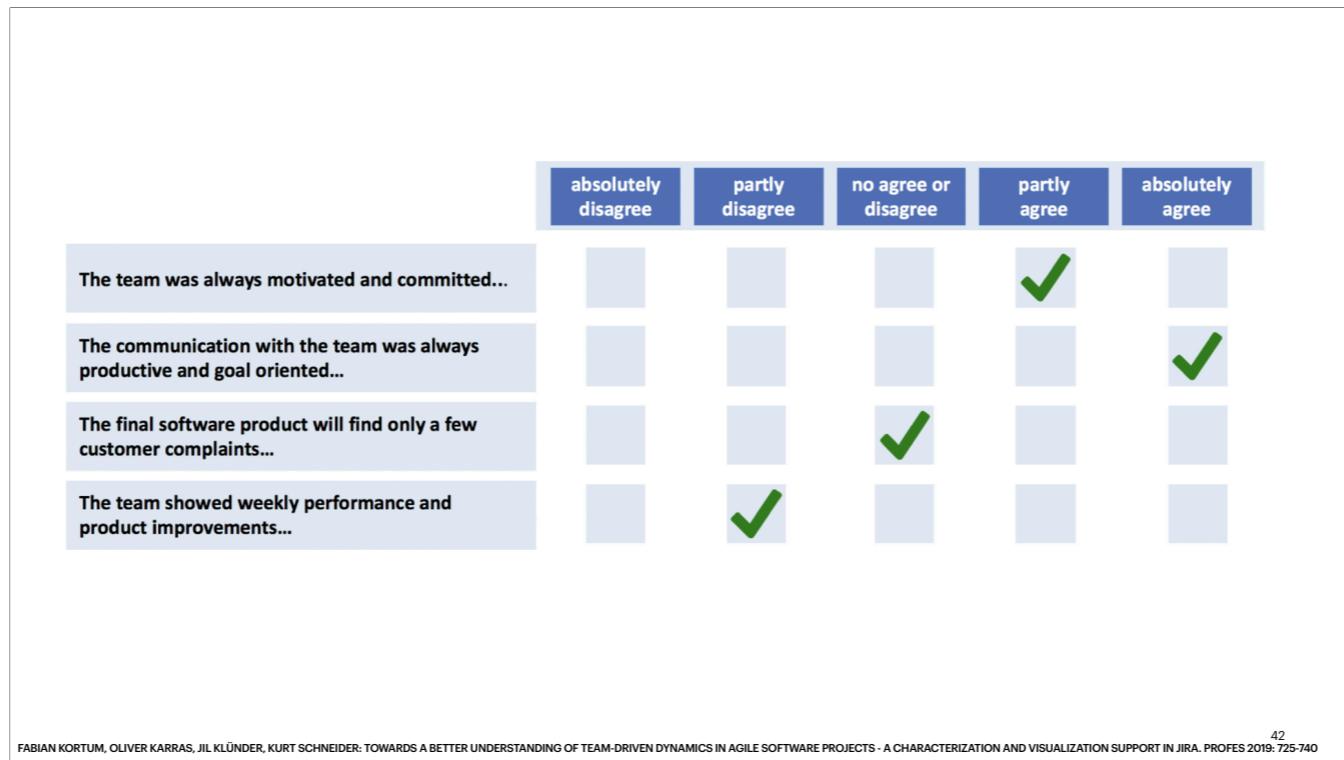
Where does this magic 5% come from and what can we do about it?

Usually 5% is associated with the risk level being labeled as “important”, loss of not more than 100 thousand euros, no lives being lost, but some injuries.

Still the choice of the experimental design influences our ability to detect phenomena, and we should not stare blindly at the data. It makes sense to combine quantitative analyse with qualitative analysis, or maybe to combine different research methods together such as repository mining and a survey.

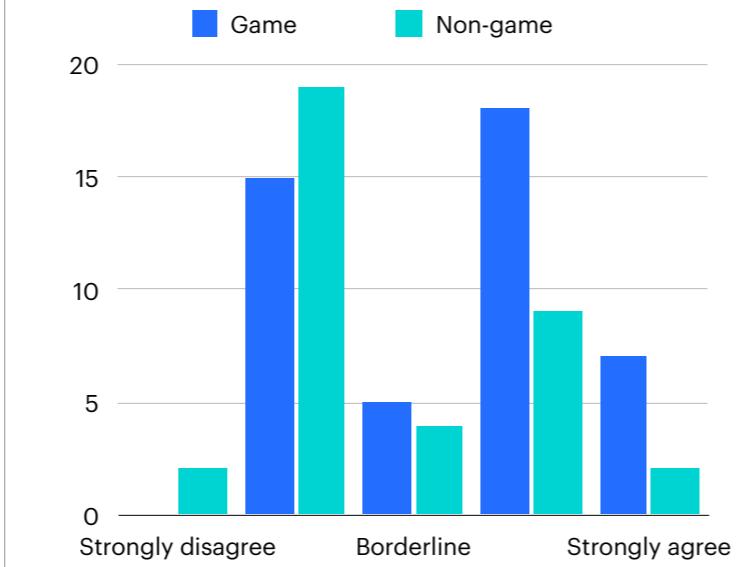


Let us move next to the ordinal scale.



Most often ordinal scale data comes from surveys. As you might remember, surveys often employ attitude questions calling participants to agree or disagree with one or more statements provided by the researchers. One usually refers to this kind of questions as questions on the Likert scale. This scale is named after its inventor, psychologist Rensis Likert. [Attn: it should be /'lɪk.ərt/ LIK-ərt[1] but commonly mispronounced /'laɪ.kərt/ LY-kərt]

Whether requirements are met in my software is highly subjective.

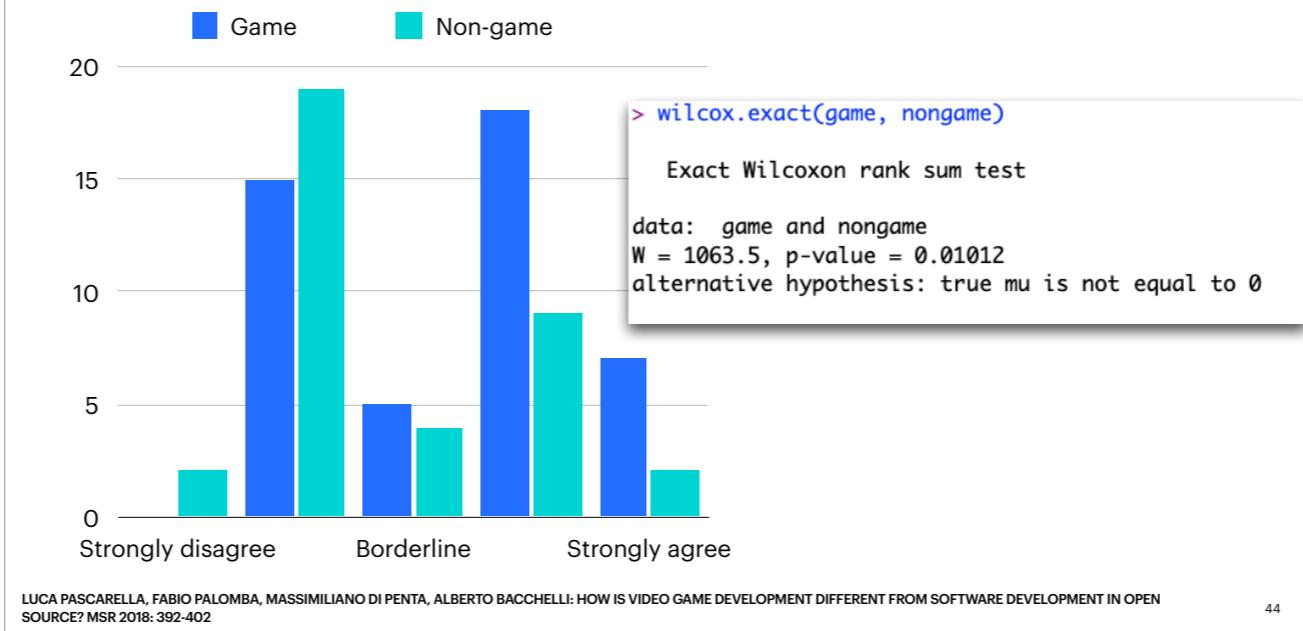


LUCA PASCARELLA, FABIO PALOMBA, MASSIMILIANO DI PENTA, ALBERTO BACCHELLI: HOW IS VIDEO GAME DEVELOPMENT DIFFERENT FROM SOFTWARE DEVELOPMENT IN OPEN SOURCE? MSR 2018: 392-402

43

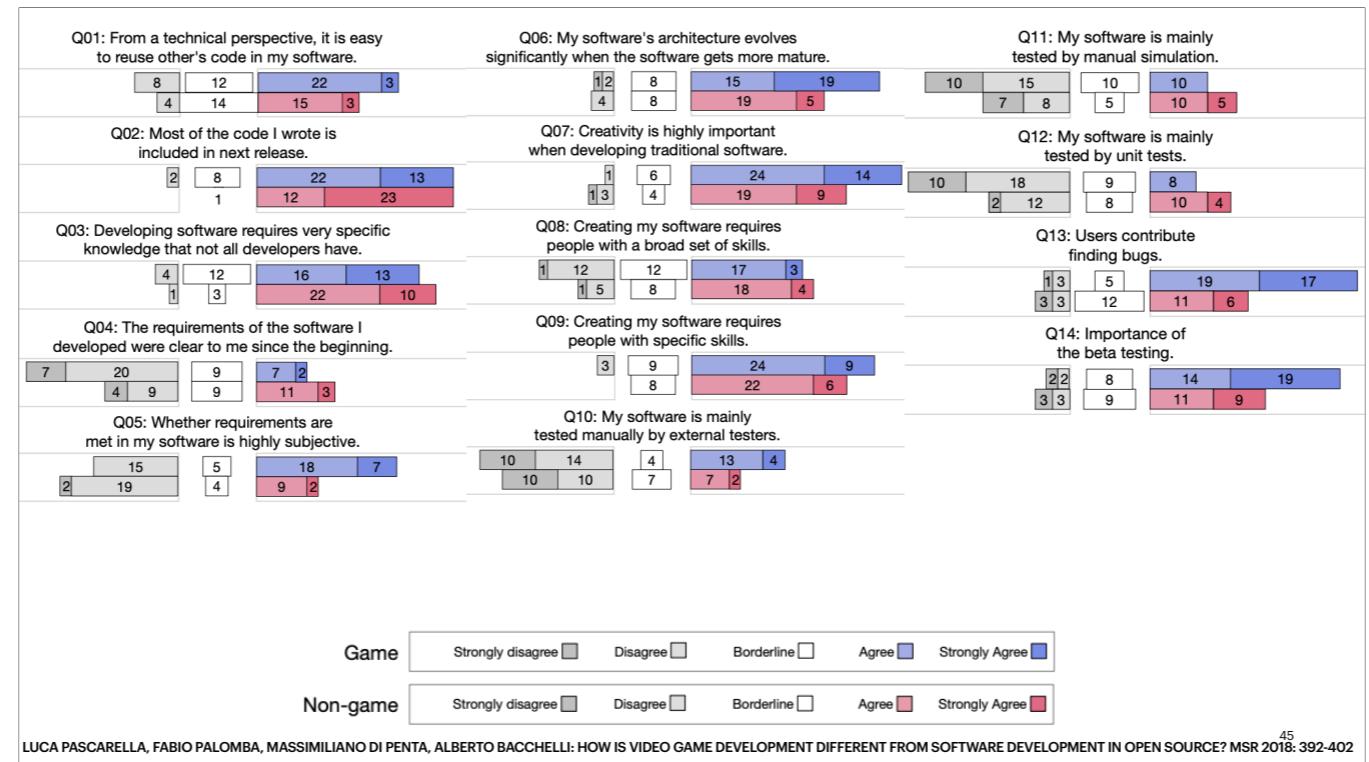
Similarly to the example with the interpretation of the emoji, we often want to compare the answers provided by two different groups of people. For example, Luca Pascarella and his co-authors wanted to compare the answers provided by video-game developers and software developers that are not game developers. The answers are shown on the slide, but what does this mean for the population? Do two samples represent two different populations or can they be derived from the same population?

Whether requirements are met in my software is highly subjective.

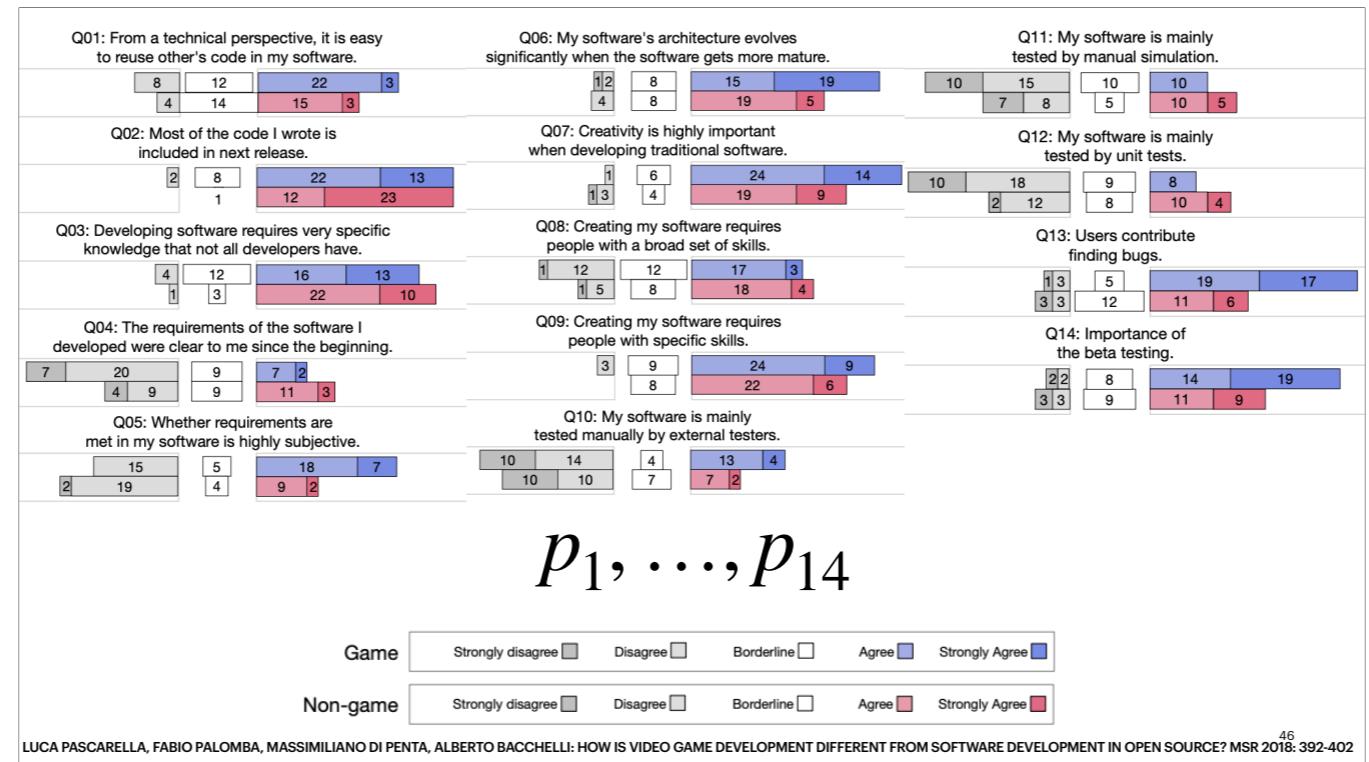


In terms of the process we are going to do the same as with the chi-square test. We will take as the null hypothesis that there are no differences between game developers and non-game developers, and we take presence of differences as an alternative hypothesis. The test we use is due to Frank Wilcoxon (2 September 1892 – 18 November 1965) who was a chemist and statistician, known for the development of several statistical tests. The p-value is lower than the common threshold of 0.05 meaning that we can reject the null hypothesis and claim that there is a difference...

Moreover, there is even more good news! The Wilcoxon test can also be applied to numerical values.



However, usually we have multiple questions. For example, the question on the previous slide comes from a series of 14 questions.



Of course, we can compute fourteen p-values corresponding to the Wilcoxon test for each one of the questions. It seems natural to select the p-values lower than the common threshold and report those findings as being the relevant differences between developers of video games and developers of other kinds of software.

THE MORE INFERENCES ARE MADE, THE MORE LIKELY ERRONEOUS INFERENCES ARE TO OCCUR.

47

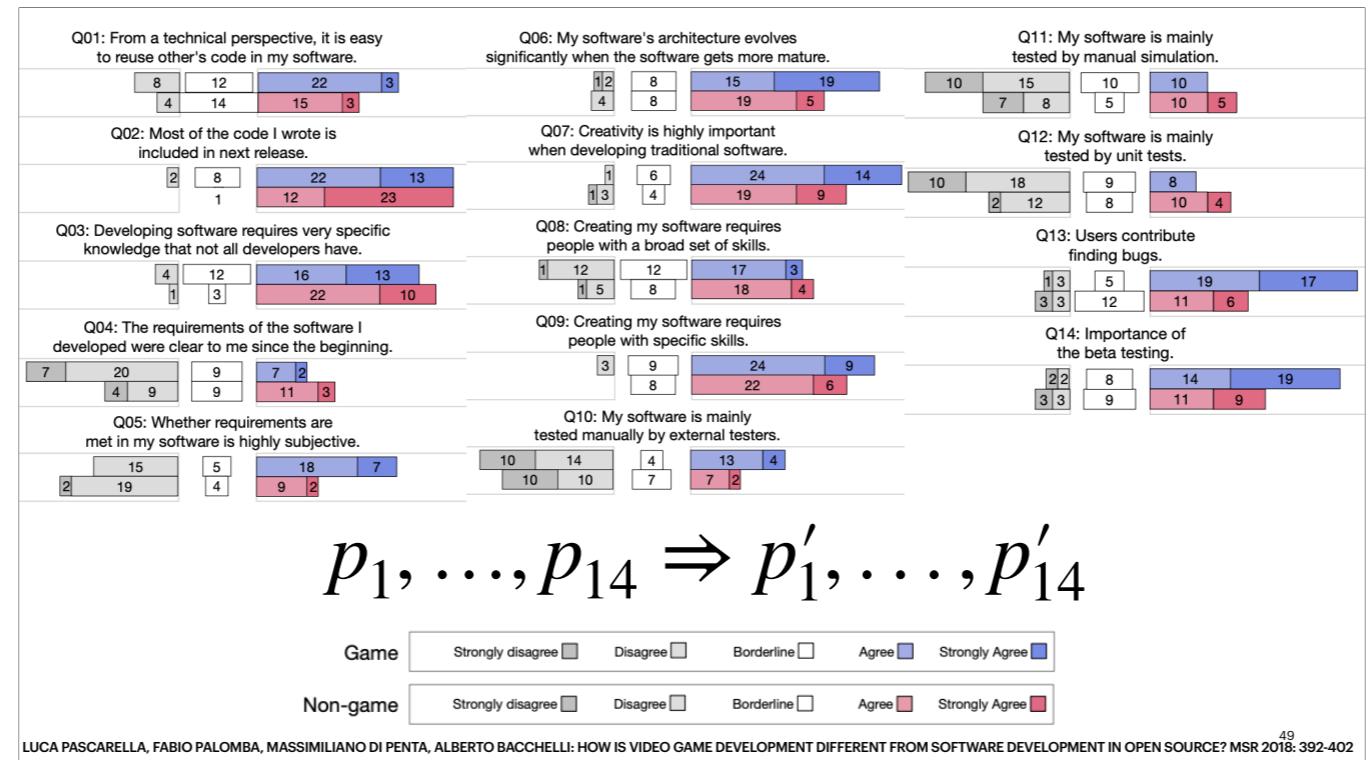
Unfortunately, the more inferences are made, the more likely erroneous inferences are to occur. Even if we have 5% threshold there could still be a 1 in 20 chance that this result was purely a statistical fluke, i.e., there was no difference but our statistical hypothesis testing still claimed there is one.

THE MORE INFERENCES ARE MADE, THE MORE LIKELY ERRONEOUS INFERENCES ARE TO OCCUR.

$$1 - 0.95^{14} \simeq 0.51$$

48

The probability of having no false positive in 14 tests in the study of Pascarella et al. is $0.95^{14} \sim 0.4876\%$, so the chance that there is at least one false positive is slightly more than a half.

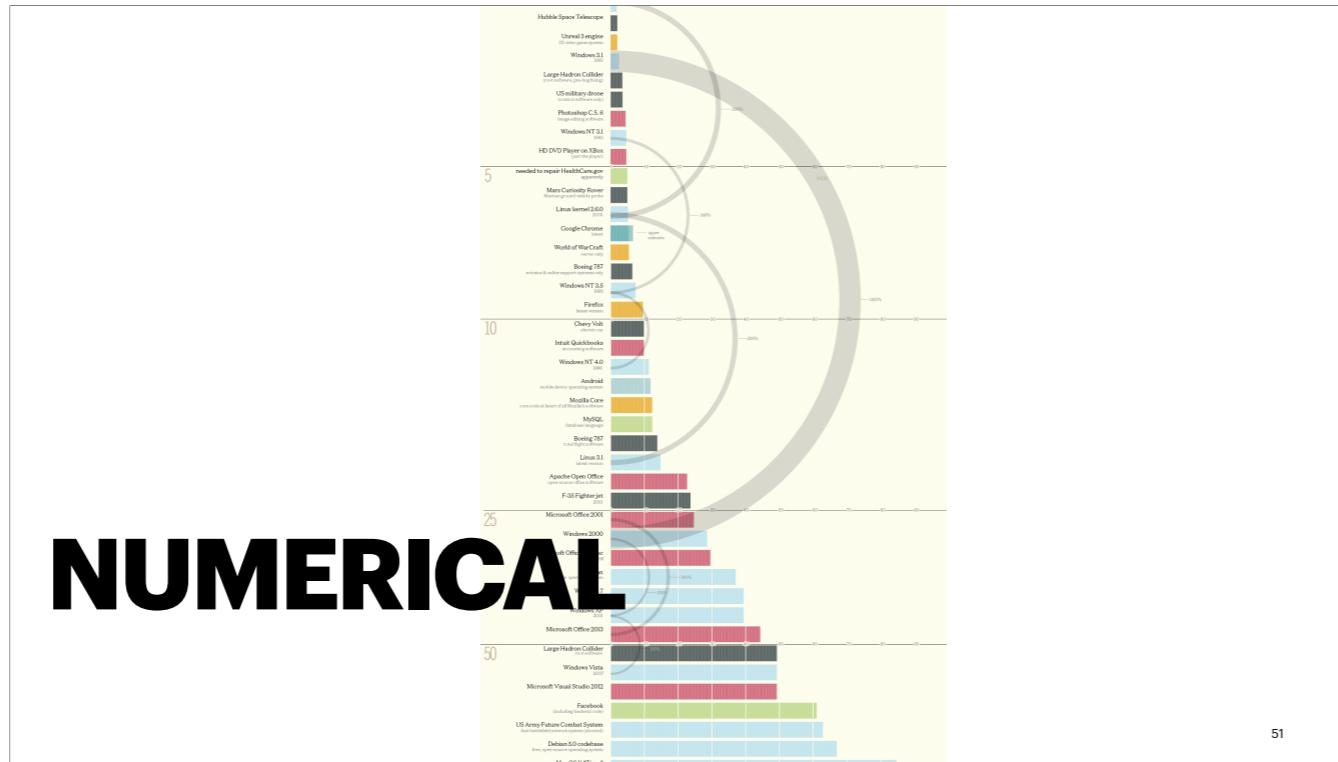


Luckily, there exist statistical methods that **increase** the p-values to control for false discovery rate. There are multiple such techniques, Pascarella et al. have chosen the Benjamini-Hochberg correction.

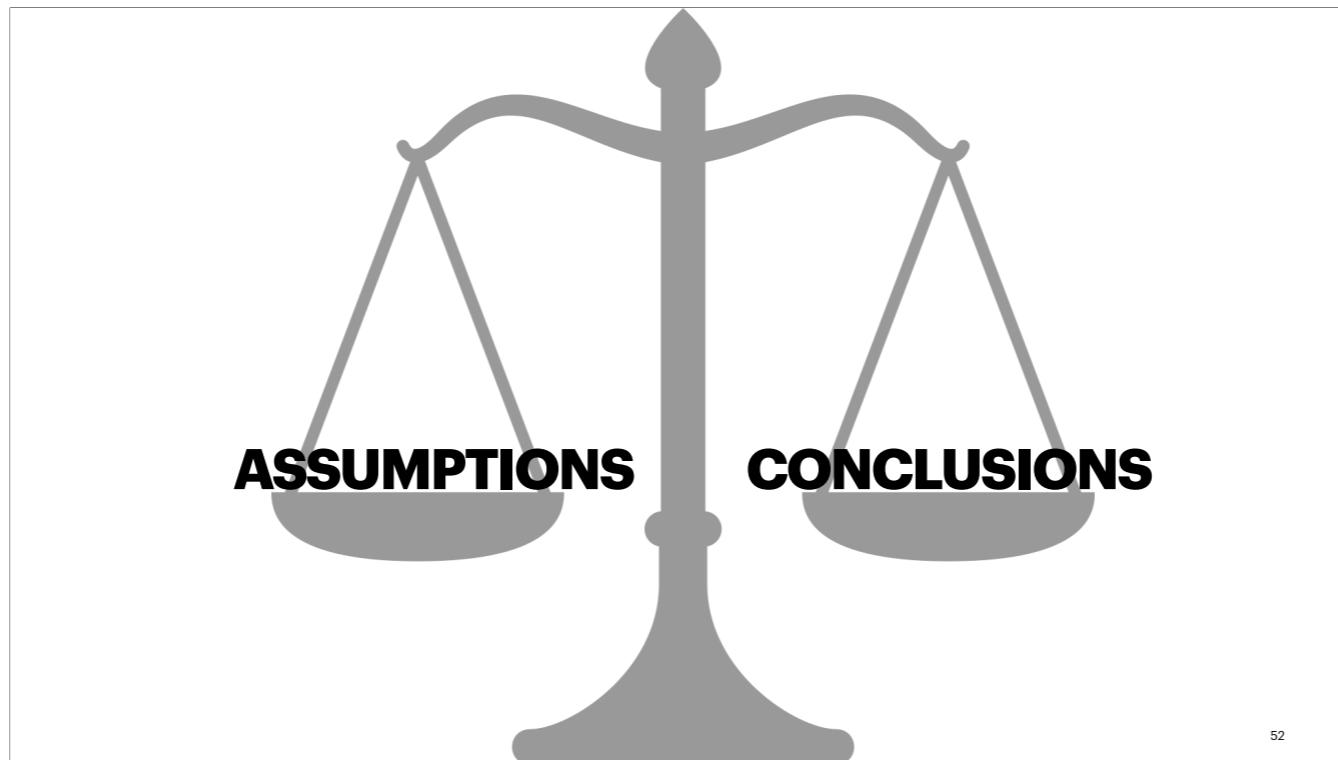
| Q | p | p adjusted | Q | p | p adjusted |
|----------|-----------------|---------------|-----------|----------------|---------------|
| 1 | 0,979 | 0,979 | 8 | 0,122 | 0,191 |
| 2 | 0,000492 | 0,0069 | 9 | 0,822 | 0,885 |
| 3 | 0,172 | 0,241 | 10 | 0,433 | 0,505 |
| 4 | 0,0412 | 0,0961 | 11 | 0,106 | 0,185 |
| 5 | 0,010 | 0,0354 | 12 | 0,00564 | 0,0354 |
| 6 | 0,028 | 0,0785 | 13 | 0,00866 | 0,0354 |
| 7 | 0,341 | 0,435 | 14 | 0,076 | 0,152 |

LUCA PASCARELLA, FABIO PALOMBA, MASSIMILIANO DI PENTA, ALBERTO BACCHELLI: HOW IS VIDEO GAME DEVELOPMENT DIFFERENT FROM SOFTWARE DEVELOPMENT IN OPEN SOURCE? MSR 2018: 392-402 ⁵⁰

To give you an impression of the impact of the adjustment, the following table shows the p-values computed by the Wilcoxon test and the values after the Benjamini-Hochberg adjustment. You might notice that before the adjustment we had six values are lower than 0.05 (boldface), after the adjustment only four.



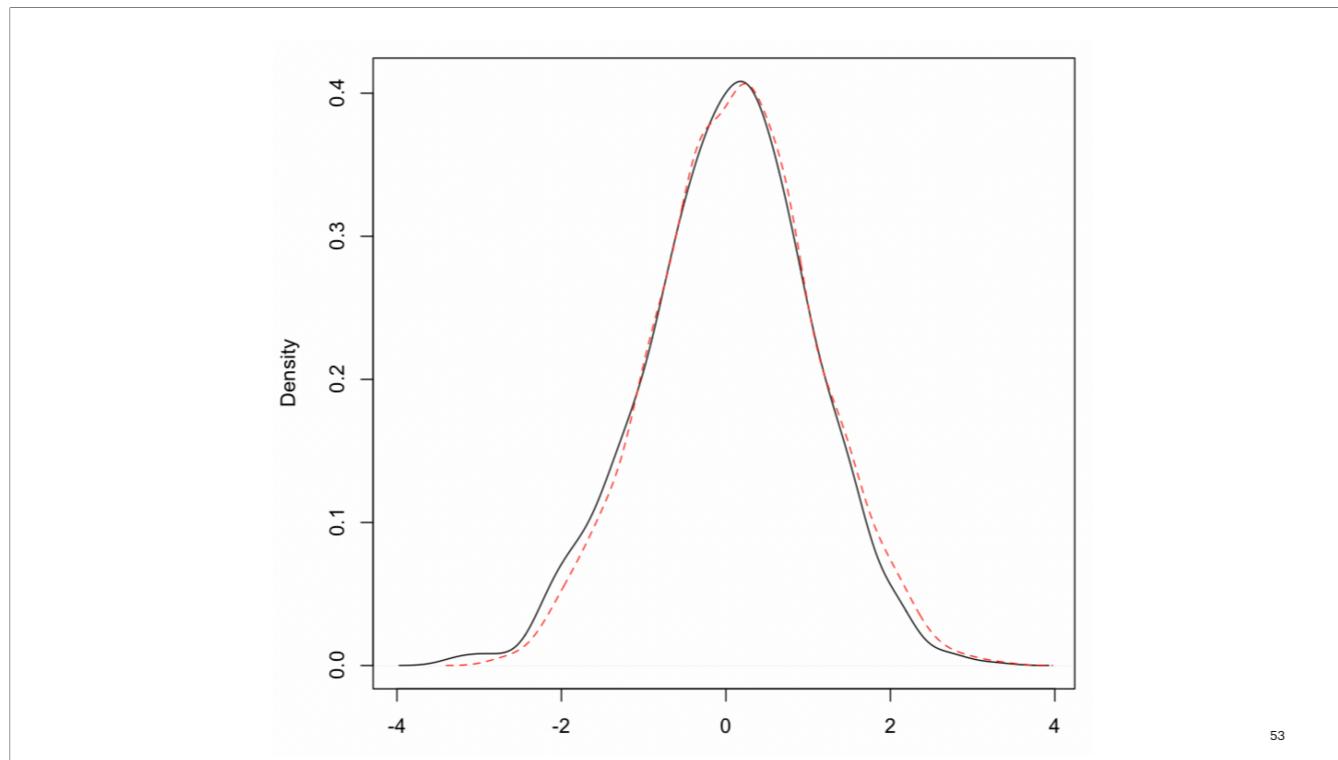
Finally, let us move to the numerical section. I have already mentioned that the Wilcoxon test can be applied to numerical data. We are going to take a closer look at comparison of two distributions.



52

In general, statistical analysis balances assumptions we can make about the data and power of the conclusions we can derive. We distinguish between

- **Parametric statistics** which assumes that sample data comes from a population that can be adequately modeled by a probability distribution that has a fixed set of parameters. This is the most constrained option, the assumptions are the strongest ones and we can infer the strongest conclusions. However, these assumptions are not always valid.
- **Non-parametric:** The assumptions made about the process generating the data are much less than in parametric statistics and may be minimal
- There is also **semi-parametric** which is somehow half-way: it combines a statistical model that has parametric and nonparametric components.



Here we see two very similar normal distributions. The black solid line represents the density plot of the normal distribution with the mean 0; the red dashed line - the normal distribution with the mean 0.11. In both cases the standard deviation is 1.

```
x = rnorm(1000)
y = rnorm(1000, mean=0.11)
> t.test(x,y)
```

Welch Two Sample t-test

```
data: x and y
t = -2.0166, df = 1997.2, p-value = 0.04388
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.176312309 -0.002455622
sample estimates:
mean of x mean of y
0.02514615 0.11453012
```

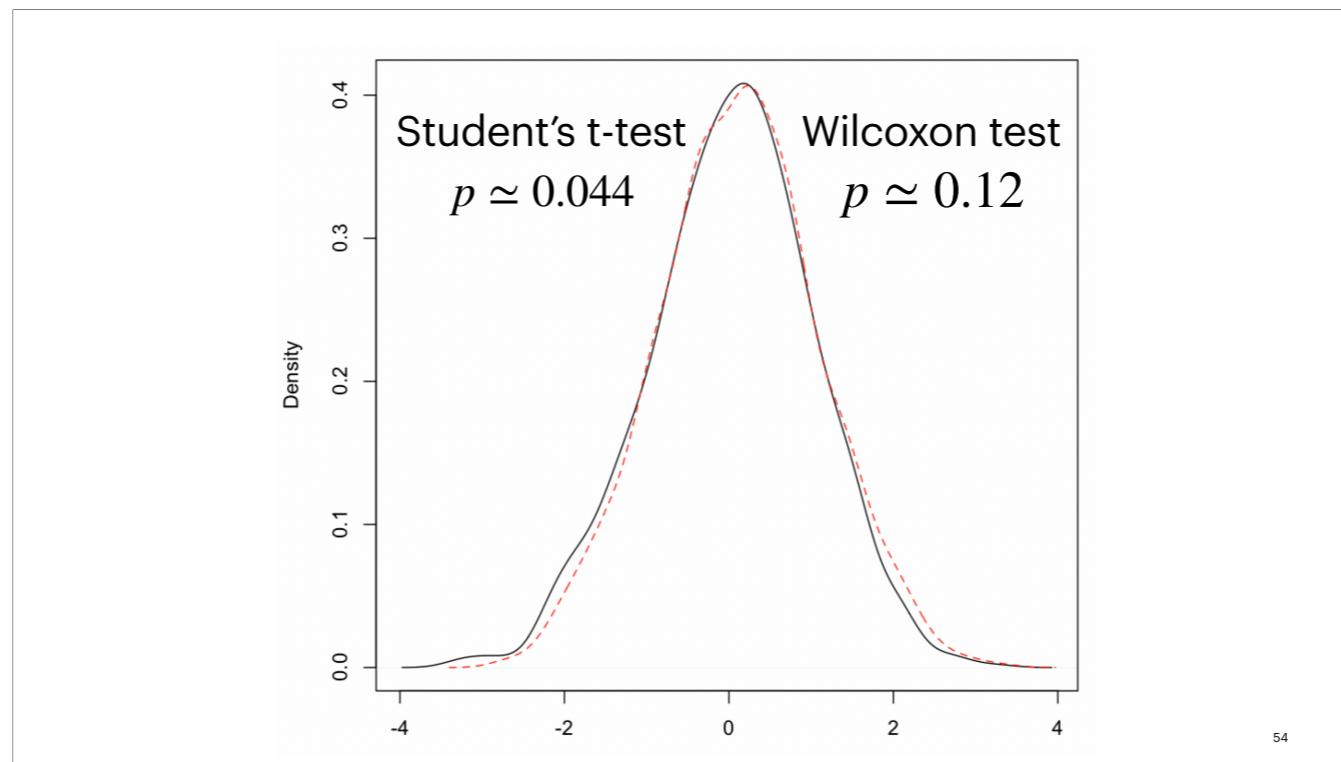
```
> wilcox.test(x,y)
```

Wilcoxon rank sum test with continuity correction

data: x and y

W = 479709, p-value = 0.1161

alternative hypothesis: true location shift is not equal to 0



Student's t test is can answer the same question as the Wilcoxon test we have seen before but it is only applicable if both distributions compared can be assumed to be normal. If one uses the traditional threshold of 0.05 then the Student's t-test is powerful enough to distinguish between the populations represented by these samples, while the Wilcoxon test cannot do this.

```
x = rnorm(1000)
y = rnorm(1000, mean=0.11)
> t.test(x,y)
```

Welch Two Sample t-test

```
data: x and y
t = -2.0166, df = 1997.2, p-value = 0.04388
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-0.176312309 -0.002455622
sample estimates:
mean of x mean of y
```

0.02514615 0.11453012

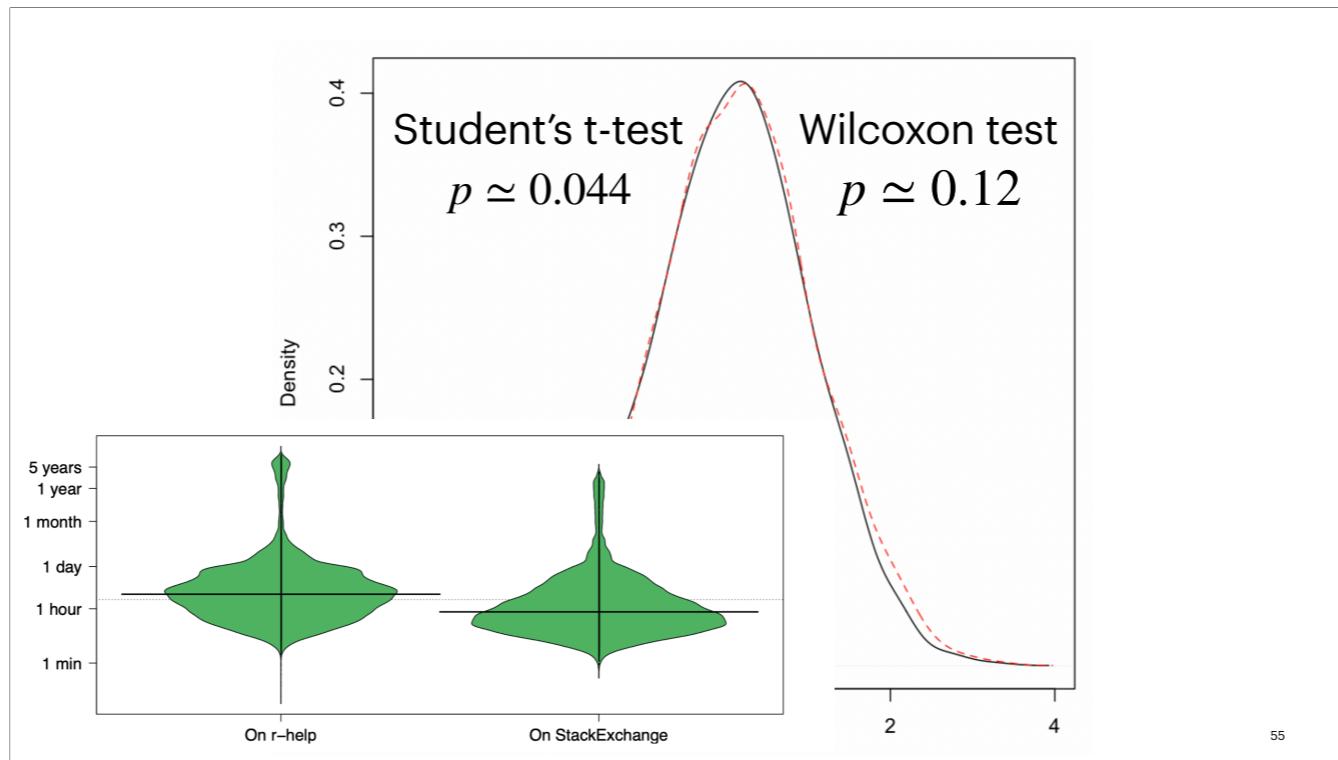
> wilcox.test(x,y)

Wilcoxon rank sum test with continuity correction

data: x and y

W = 479709, p-value = 0.1161

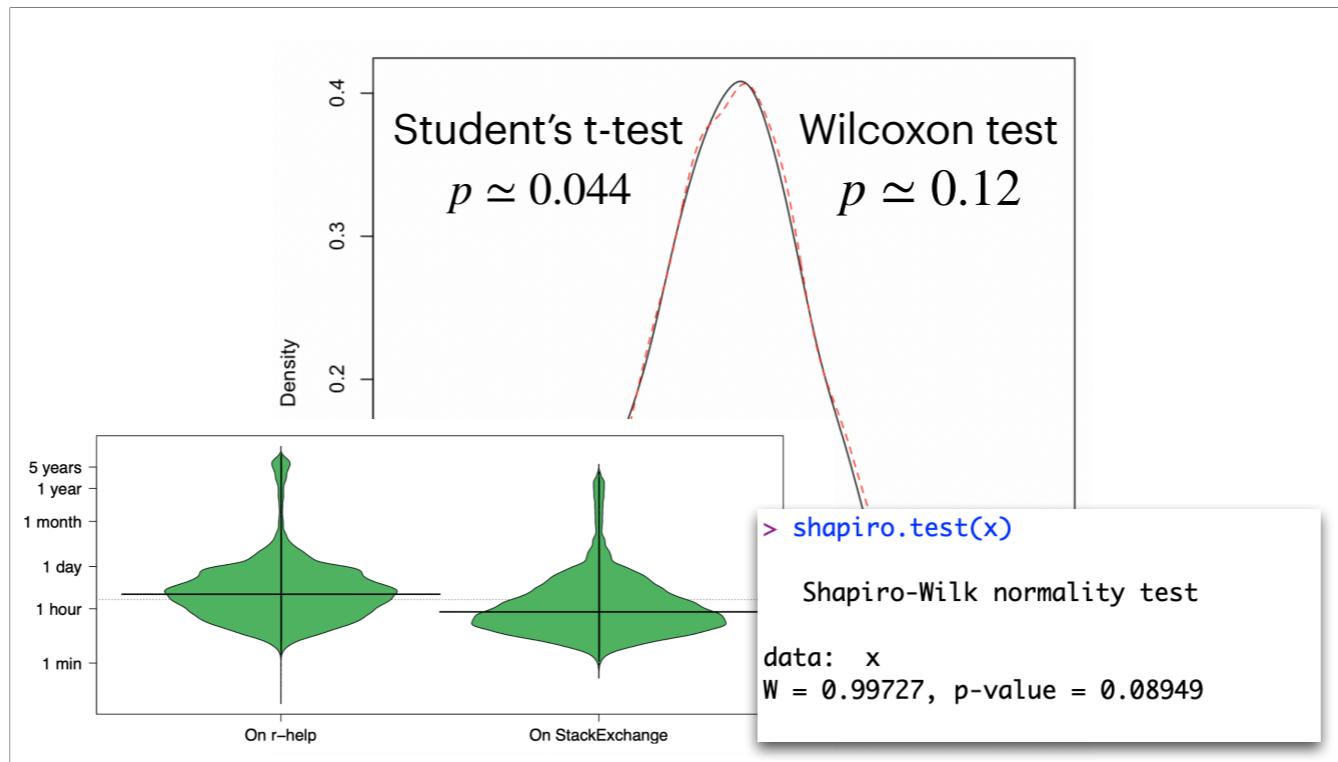
alternative hypothesis: true location shift is not equal to 0



However, in software engineering we often use the Wilcoxon test. In this paper of Bogdan Vasilescu et al. the authors have studied the community surrounding R. In this study the authors compute for each developer active both on Stack Exchange about R and on the mailing list about R, group the time intervals between their first answer within a thread and the thread start (for all threads for which they provide answers), on the one hand, and between their first answer to a StackExchange question and the question date (for all questions they answer), on the other hand. Then, using the Wilcoxon test we compare two large groups of r-help and StackExchange time deltas obtained by concatenating the intervals for all “r-help and StackExchange” members.

So why Wilcoxon and not Student's t-test?

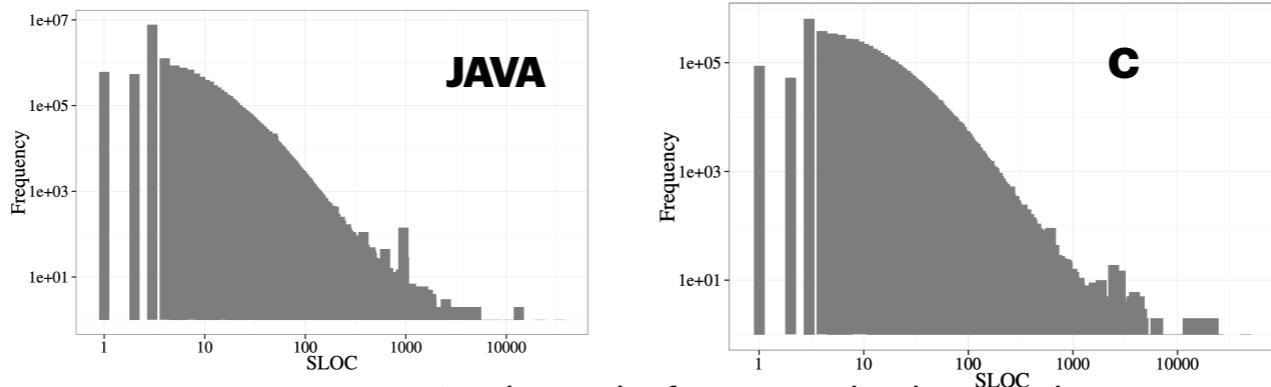
If you slant your head you see that the shape of these beans does not look like a bell curve, in particular you see a long “tail”. This is why we suspect that the normality assumption of the Student's t-test is not satisfied and we have to use the Wilcoxon test.



Of course, merely looking at the picture is often not good enough. This is the purpose of normality tests. There is an entire series of such tests, for example, Kolmogorov-Smirnov's test, Shapiro-Wilk test and many others. On the slide you see the application of the Shapiro-Wilk test to the data of the black curve.

The null hypothesis of the Shapiro-Wilk test is that the distribution is normal, so if the p-value exceeds the threshold of 0.05 then the null hypothesis cannot be rejected and the parametric techniques such as Student's t-test can be applied. In a way, normality tests are “special”: usually we like when the p-values are small, but not in case of normality tests!

QUESTION



To compare size (SLOC) of Java methods with the size (SLOC) of C functions one shall use

(A) STUDENT'S T-TEST

(B) WILCOXON TEST

DAVY LANDMAN, ALEXANDER SEREBRENIK, ERIC BOUWERS, JURGEN J. VINJU: EMPIRICAL ANALYSIS OF THE RELATIONSHIP BETWEEN CC AND SLOC IN A LARGE CORPUS OF JAVA METHODS AND C FUNCTIONS. J. SOFTW. EVOL. PROCESS. 28(7): 589-618 (2016)

57

B - Wilcoxon test



However, sometimes merely comparing two groups is not enough. This is particularly the case if we are interested in analysing processes that are related to the passing of time. Time is a commonly studied variable in empirical software engineering research, in particular, in the subfield of empirical software engineering related to software evolution. If you are interested in software evolution, please consider taking 2IMP25 Software evolution in Q3.

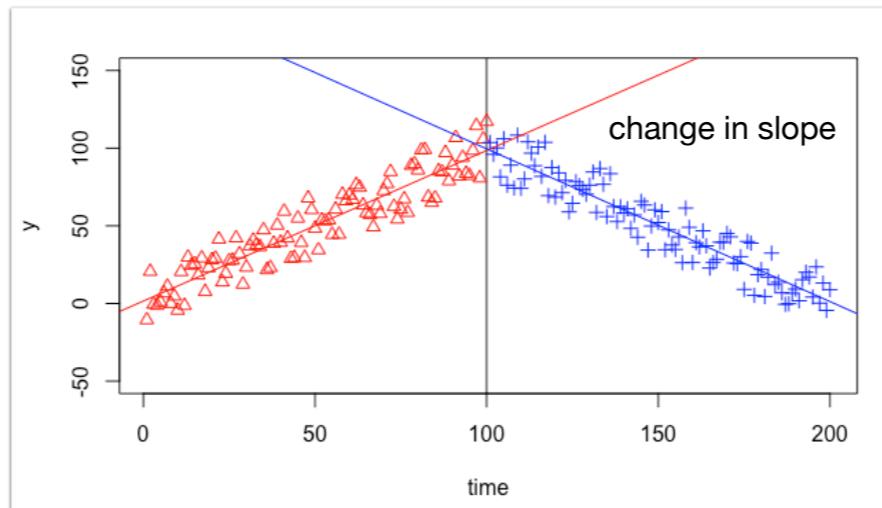
Interventions are common in software engineering

- SVN → git
- push → pull request
- ? → continuous integration
- ? → bot

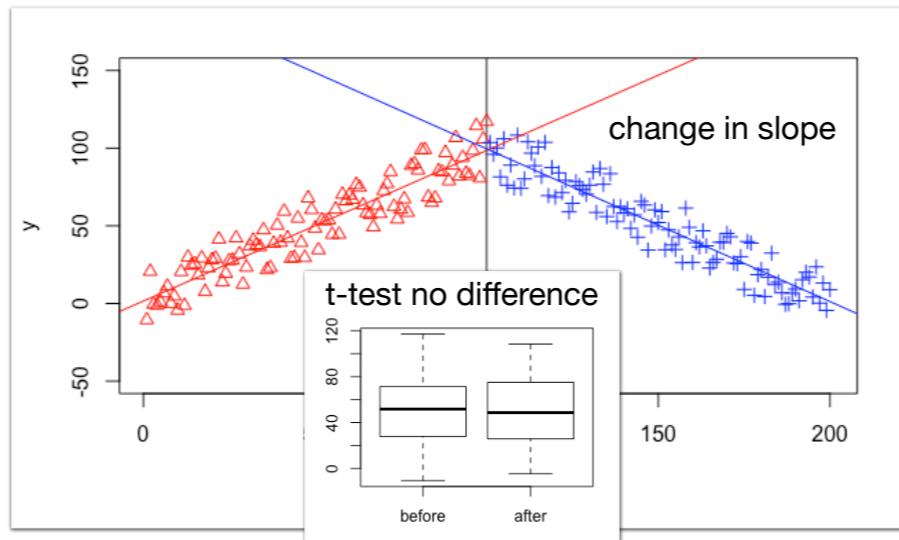
How to measure effects
of the intervention?

This is particularly true since software systems do not just evolve “naturally”. In particular, interventions are common in software engineering.

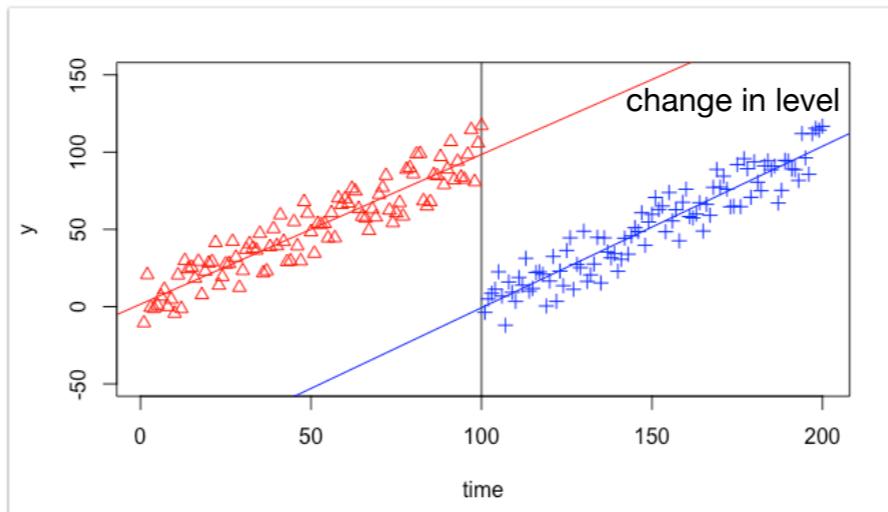
Evaluating the effects of an intervention: *before vs. after*



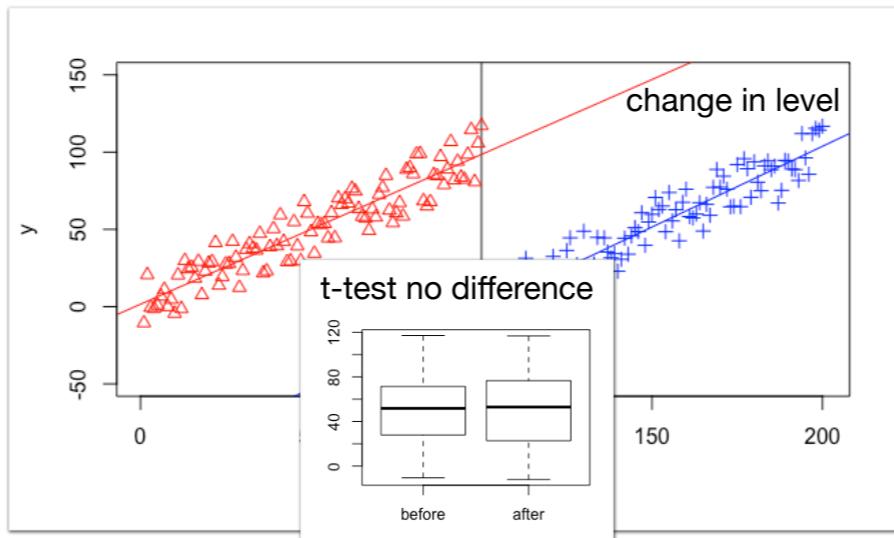
Evaluating the effects of an intervention: *before* vs. *after*



Evaluating the effects of an intervention: *before vs. after*

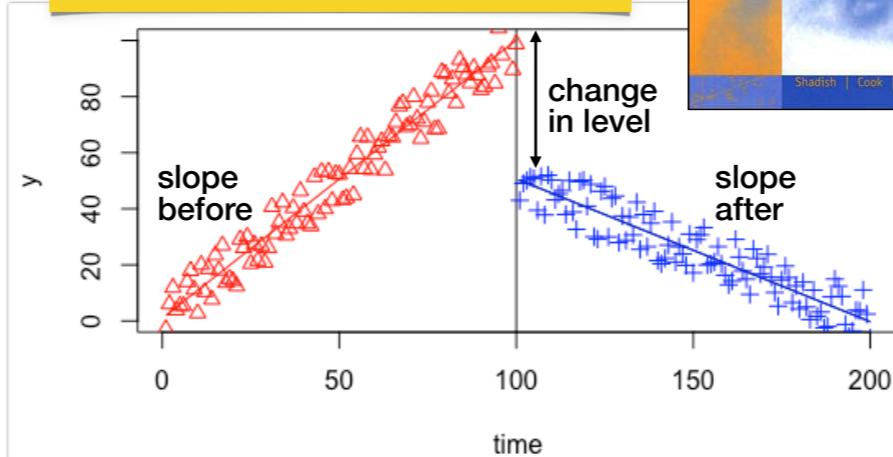


Evaluating the effects of an intervention: *before* vs. *after*

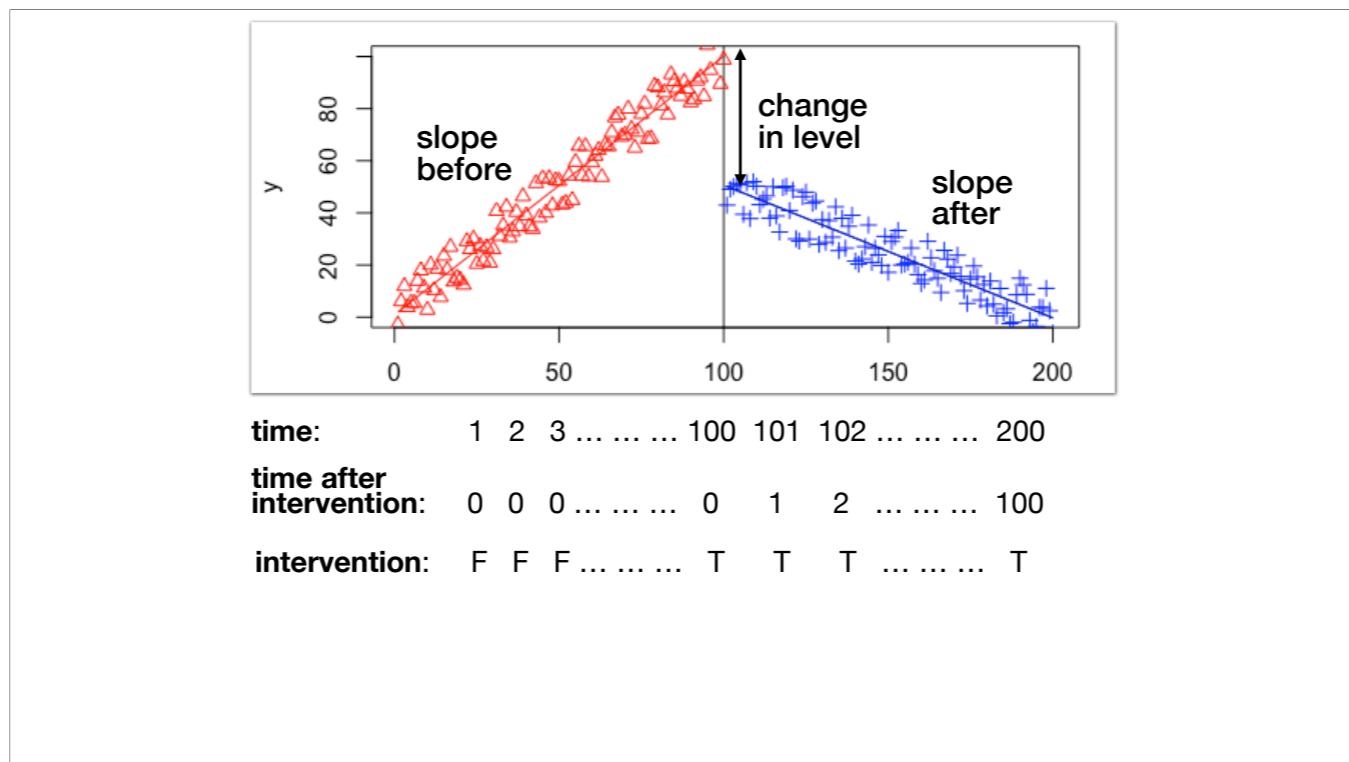


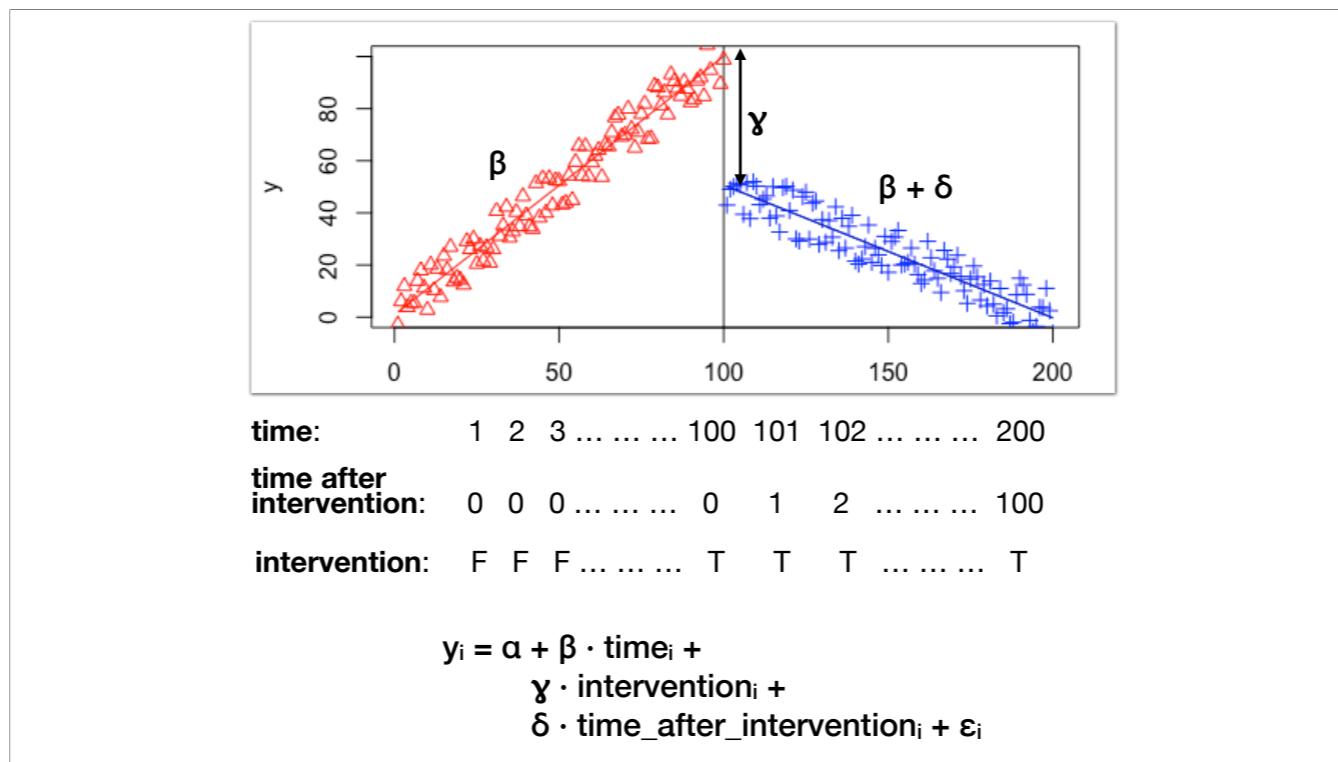
Interrupted time series

Multiple regression w/
controls for confounds

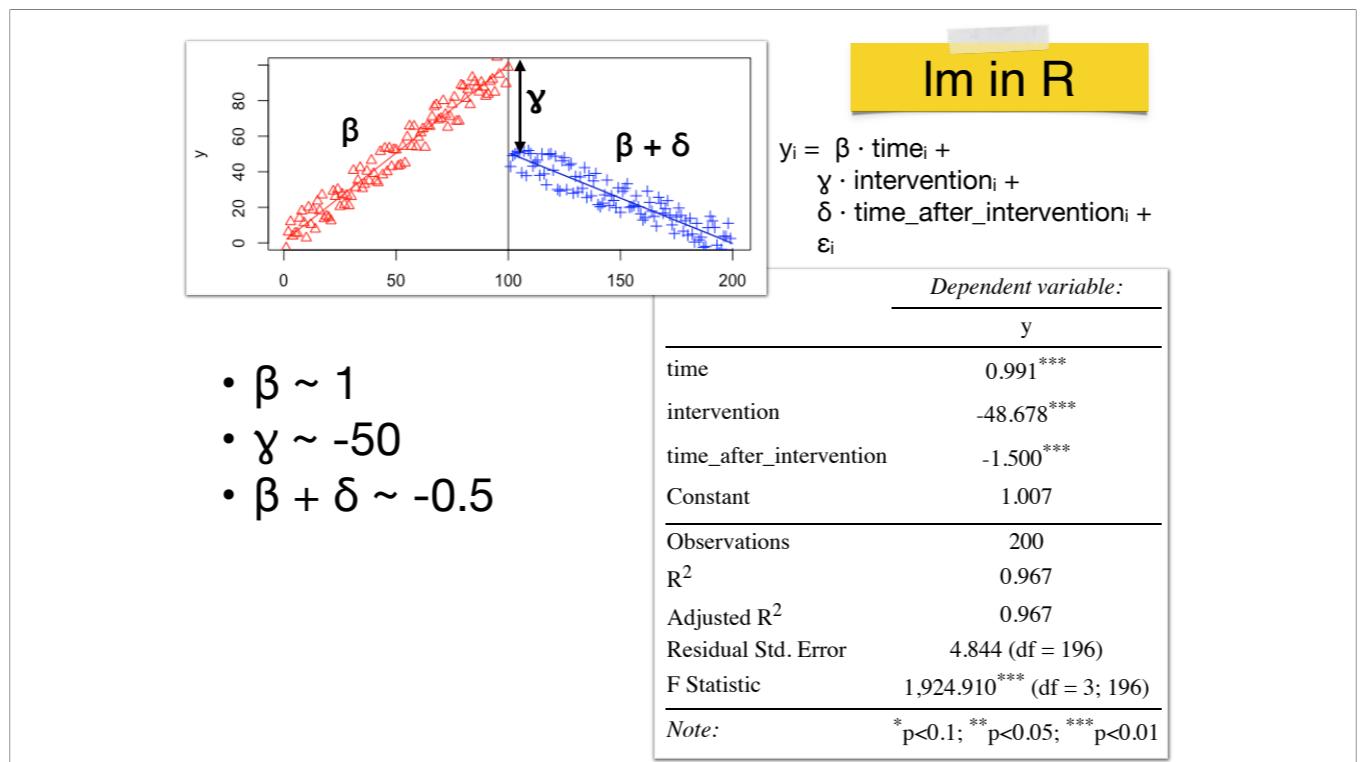


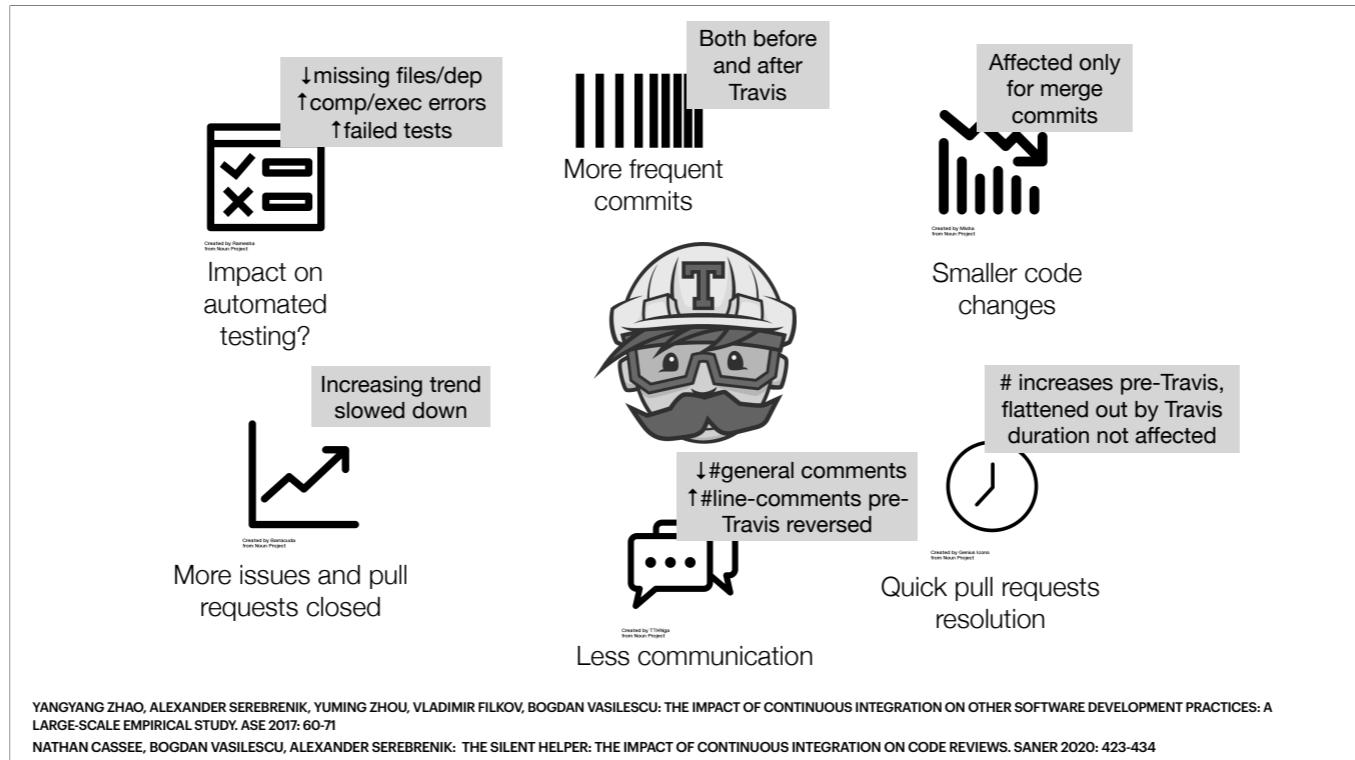
To answer this kind of questions would require comparing the situation before the introduction of CI and after the introduction of CI. In particular, we are interested in changes in trends. There are multiple ways of analysing such data, we focus on one called regression discontinuity design (RDD).



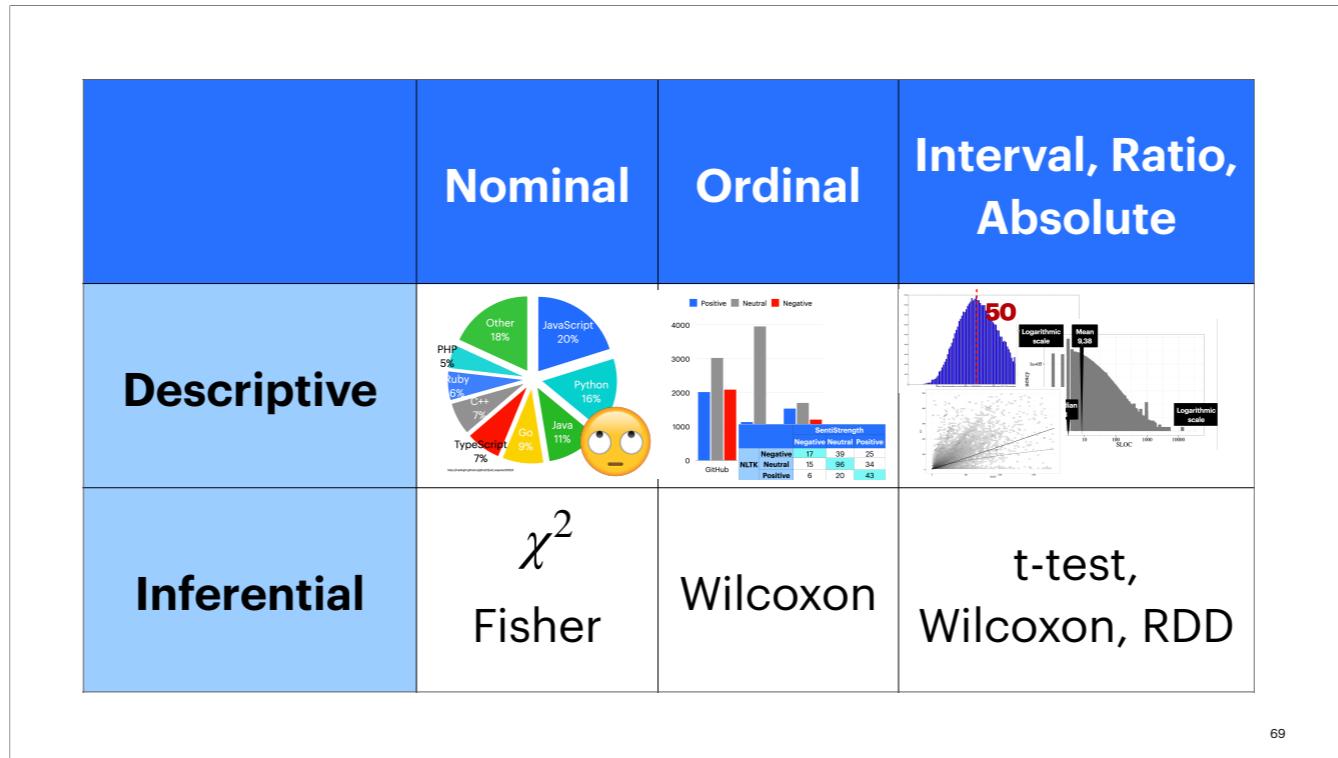


To answer this kind of questions would require comparing the situation before the introduction of CI and after the introduction of CI. In particular, we are interested in changes in trends. RDD



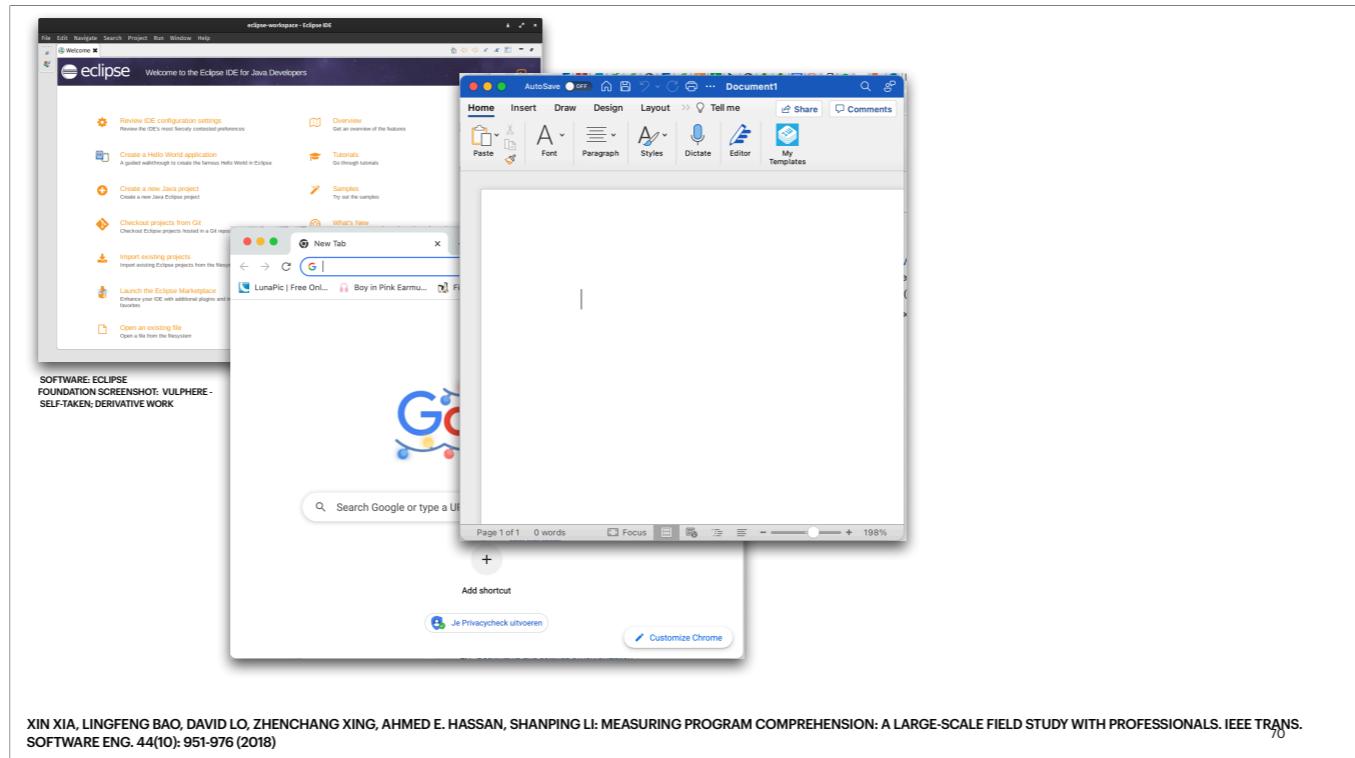


Impact of the adoption of Travis CI on software development practices in GitHub



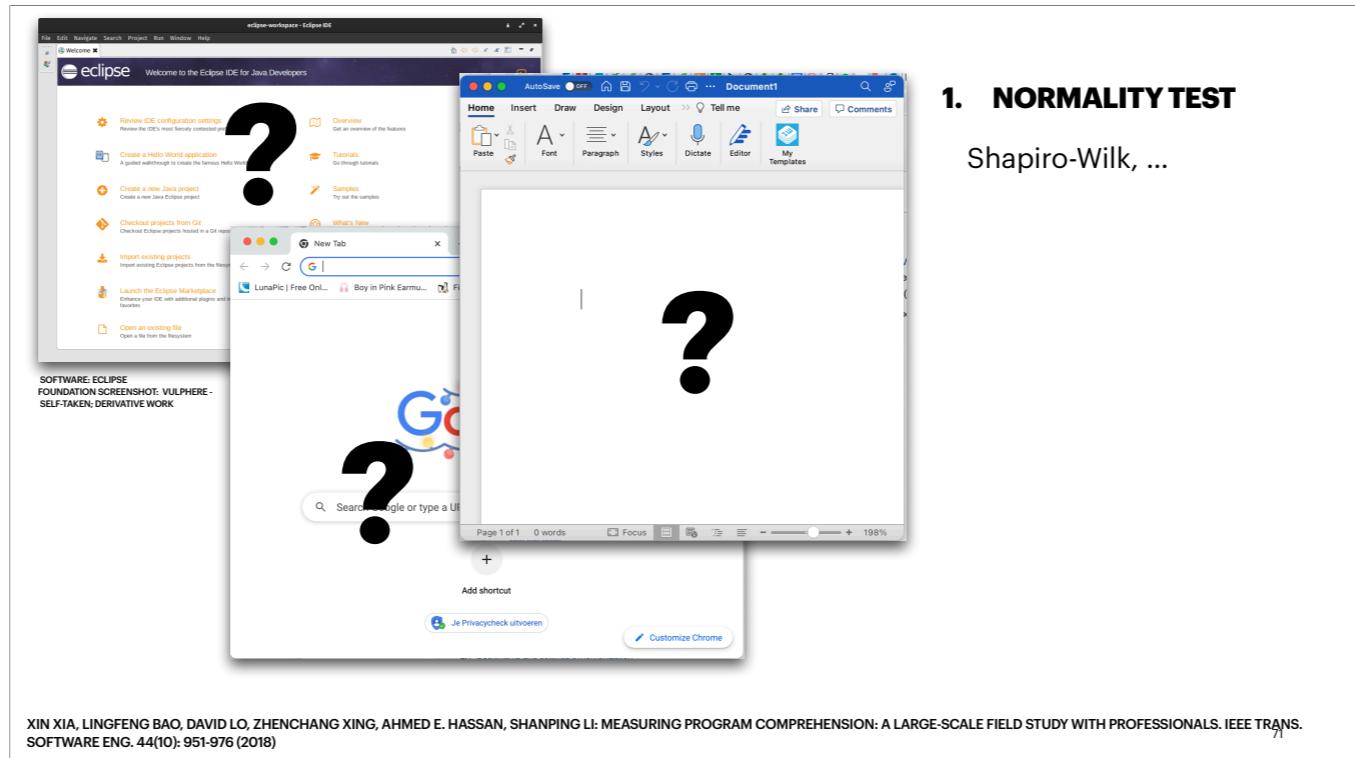
69

This is a brief summary of the techniques we have seen during this lecture.



XIN XIA, LINGFENG BAO, DAVID LO, ZHENCHANG XING, AHMED E. HASSAN, SHANPING LI: MEASURING PROGRAM COMPREHENSION: A LARGE-SCALE FIELD STUDY WITH PROFESSIONALS. IEEE TRANS. SOFTWARE ENG. 44(10): 951-976 (2018) ⁷⁰

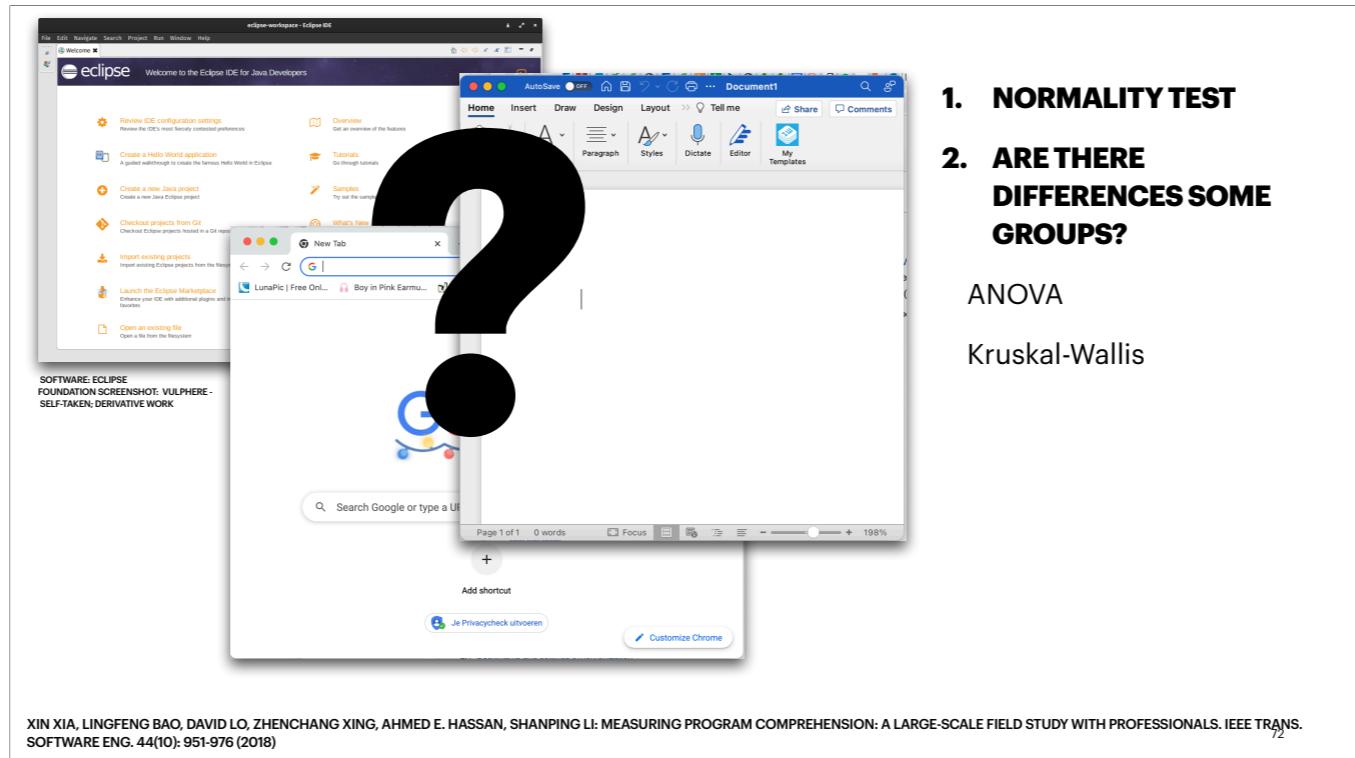
Xin Xia and his co-authors have conducted a field study at two Chinese companies and observed 83 professional developers working on seven projects. For each one of them they registered the time developers spend during program comprehension tasks, and percentages of this time spent working with an IDE, a web-browser and a text editor (for documents). The researchers wanted to understand whether there are any statistically significant differences between the percentages of the time.



1. NORMALITY TEST

Shapiro-Wilk, ...

In the same way as before, we start with a normality-test such as Shapiro-Wilk. We perform such a test for each group.



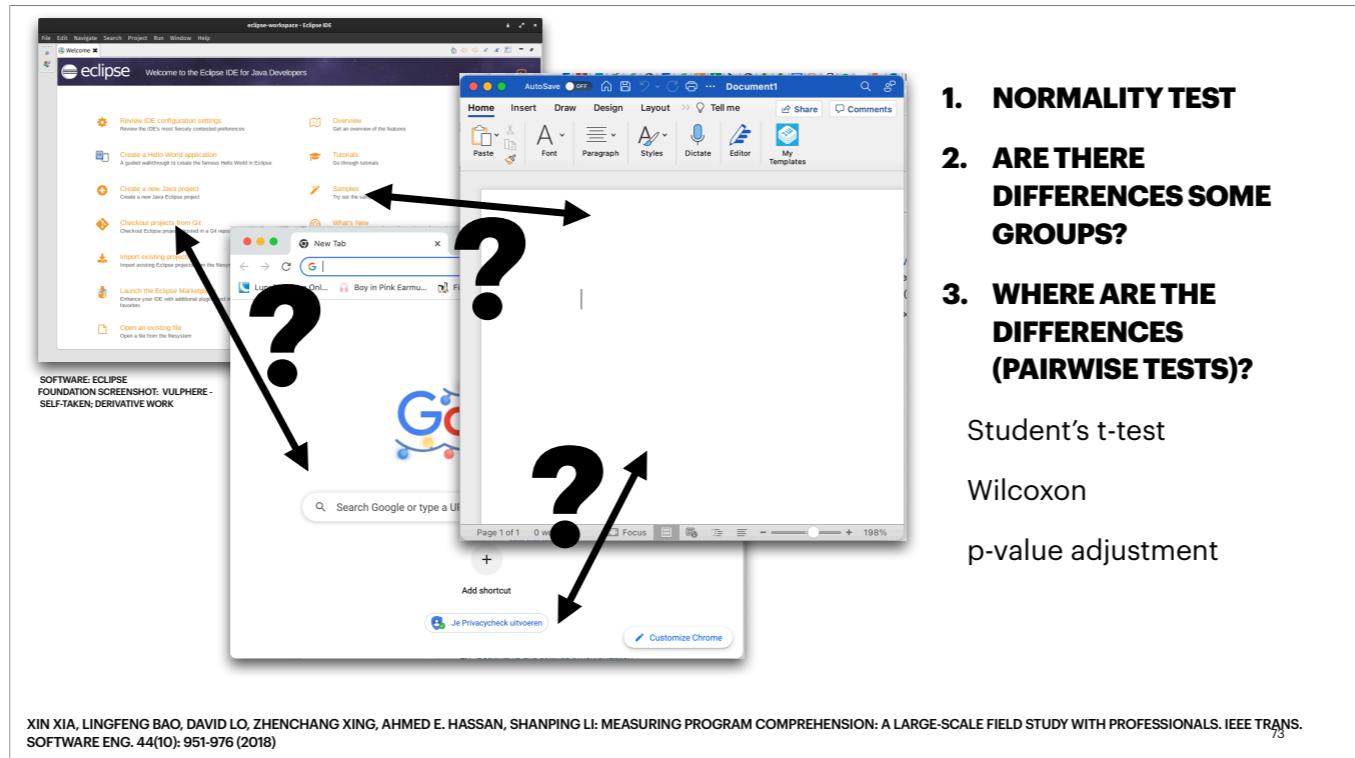
1. NORMALITY TEST

2. ARE THERE DIFFERENCES SOME GROUPS?

ANOVA

Kruskal-Wallis

Then we check whether there are differences between some groups. If the distribution on each one of the groups is normal, then we can use ANOVA; otherwise we need to use Kruskal-Wallis test. In the study of Xin Xia the distributions turned out to be normal, so they used ANOVA.



- 1. NORMALITY TEST**
- 2. ARE THERE DIFFERENCES SOME GROUPS?**
- 3. WHERE ARE THE DIFFERENCES (PAIRWISE TESTS)?**

Student's t-test

Wilcoxon

p-value adjustment

If the results of the second step suggest that there are some differences between the groups (the p-values are lower than the threshold) then we perform pairwise tests: if the distributions were normal we use pairwise t-tests and if this was not the case - pairwise Wilcoxon tests. And we also adjust the p-values obtained.

In the study of Xia et al the authors have performed three pairwise t-tests with a Bonferroni correction.

1. NORMALITY TEST

2. ARE THERE DIFFERENCES SOME GROUPS?

3. WHERE ARE THE DIFFERENCES (PAIRWISE TESTS)?

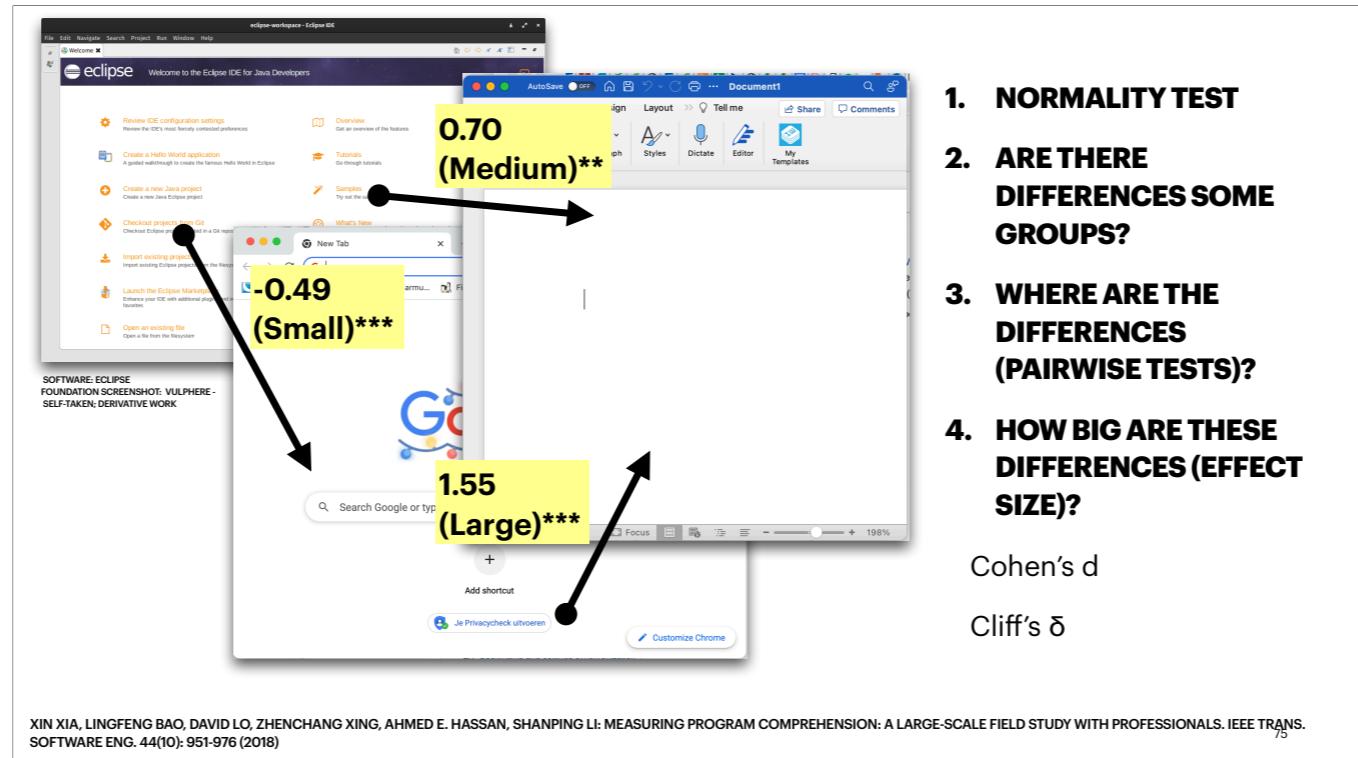
4. HOW BIG ARE THESE DIFFERENCES (EFFECT SIZE)?

Cohen's d

Cliff's δ

XIN XIA, LINGFENG BAO, DAVID LO, ZHENCHANG XING, AHMED E. HASSAN, SHANPING LI: MEASURING PROGRAM COMPREHENSION: A LARGE-SCALE FIELD STUDY WITH PROFESSIONALS. IEEE TRANS. SOFTWARE ENG. 44(10): 951-976 (2018)

Not every statistically significant difference is meaningful in practice: to capture this intuition several measures have been introduced to describe how “big” the differences are. For each one of these measures there is a standardised interpretation scale. For example, if the absolute value of Cohen’s d is smaller than 0.2 it is called “small” and if it is larger than 2 then it is called “huge”. Cohen’s d is used if the t-test has been used at the previous step, Cliff’s delta follows the Wilcoxon test.



- 1. NORMALITY TEST**
- 2. ARE THERE DIFFERENCES SOME GROUPS?**
- 3. WHERE ARE THE DIFFERENCES (PAIRWISE TESTS)?**
- 4. HOW BIG ARE THESE DIFFERENCES (EFFECT SIZE)?**

Cohen's d

Cliff's δ

These are the values of Cohen's d reported in the paper by Xia et al. Asterisks indicate statistical significance: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$. The sign is determined by the difference in the means and hence sensitive to the order of the comparison and this is why I have changed the shape of one of the end points: Browser vs Text has a positive effect size means that developers spend a larger share of their comprehension time in the browser; IDE vs Browser has a negative effect size meaning that developers spend a small share of their time in the IDE.

This means that developers spend least time on program comprehension activities when using text editors and most time when using web browsers.



- 1. NORMALITY TEST**
- 2. ARE THERE DIFFERENCES SOME GROUPS?**
- 3. WHERE ARE THE DIFFERENCES (PAIRWISE TESTS)?**
- 4. HOW BIG ARE THESE DIFFERENCES (EFFECT SIZE)?**

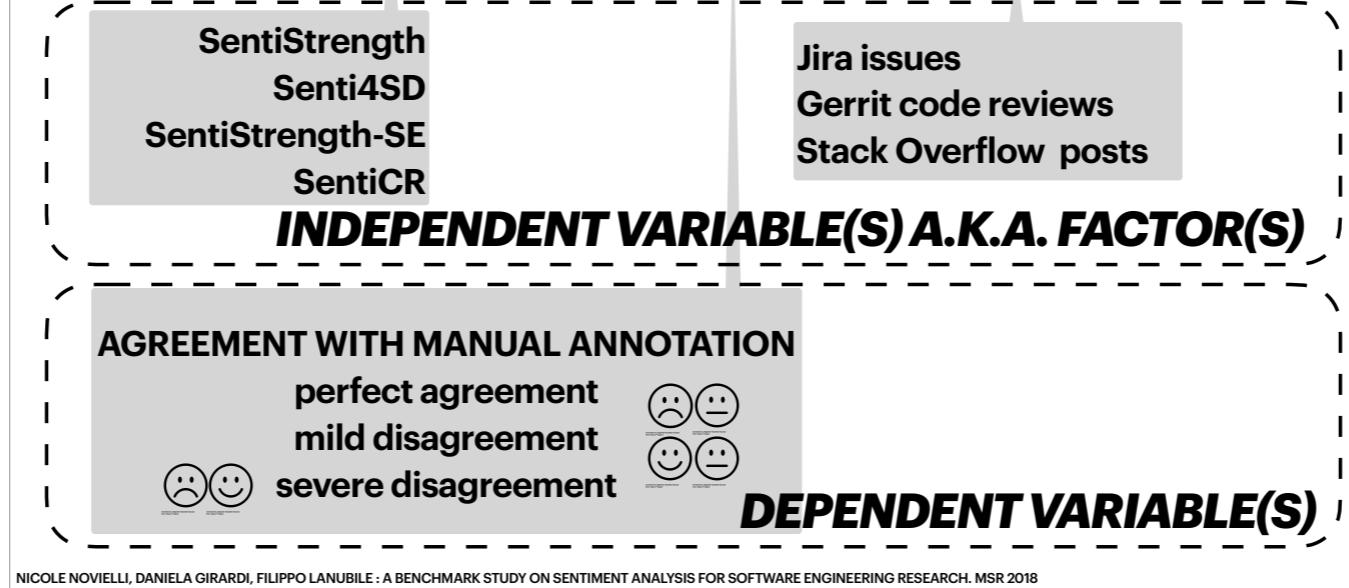
76

While this is a common method it is not perfect. For example, while not very common in actual studies, it is possible that the second step claims that there are differences between some groups but the follow-up pairwise tests do not detect them. Alternatively, it is possible that the second step does not find any differences but the follow-up pairwise comparisons would have found them. Moreover, the pairwise comparisons do not guarantee transitivity. Konietzschke et al have proposed more advanced methods allowing one to combine steps 2, 3 and 4 and ensuring transitivity. Their popularity in software engineering remains limited.



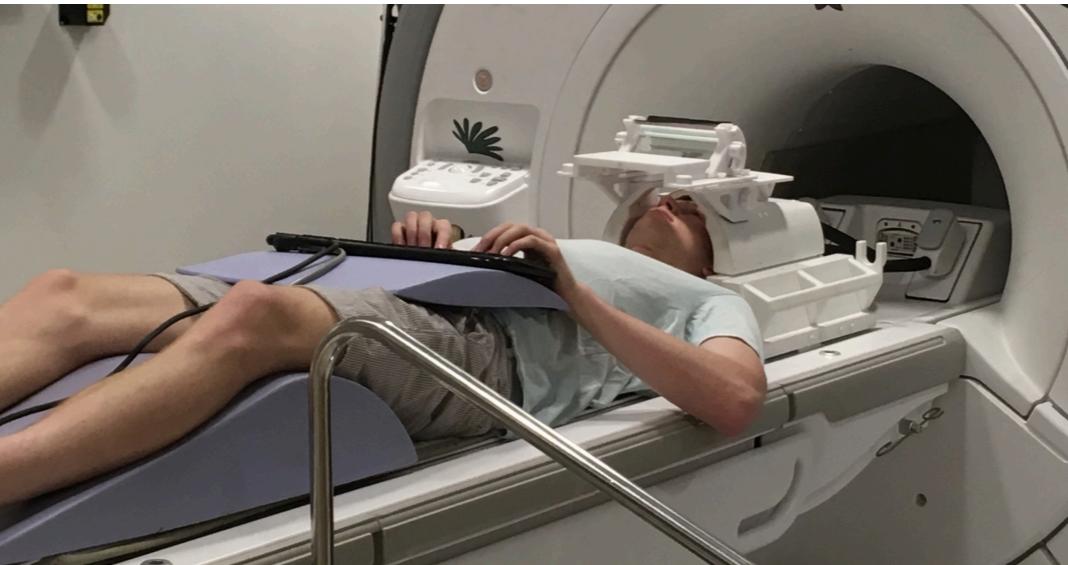
But whether we use more traditional multi-step approaches or more advanced approaches such as Konietschke's `nparcomp`, we are limited by the number of groups we want to consider. These approaches will never work for tens of people, or hundreds or thousands of projects! Still, we need some kind of mechanism capable of taking into account differences between individuals and projects... This kind of mechanisms are known as mixed-effects models. Before discussing mixed effects let me briefly recap a couple of notions you might be familiar with.

WHICH TOOL IS THE BEST FOR WHICH DATASET?



In this case we want to understand how the choice of a sentiment analysis tool and the choice of the dataset influences the agreement with manual annotations. The independent variables are the input, the dependent variables are the output, something that is expected to be influenced by the independent variable(s). In this case we have two independent variables, a.k.a. factors: tool and data source; and three dependent variables (% of perfect agreement, % mild disagreement and % severe disagreement).

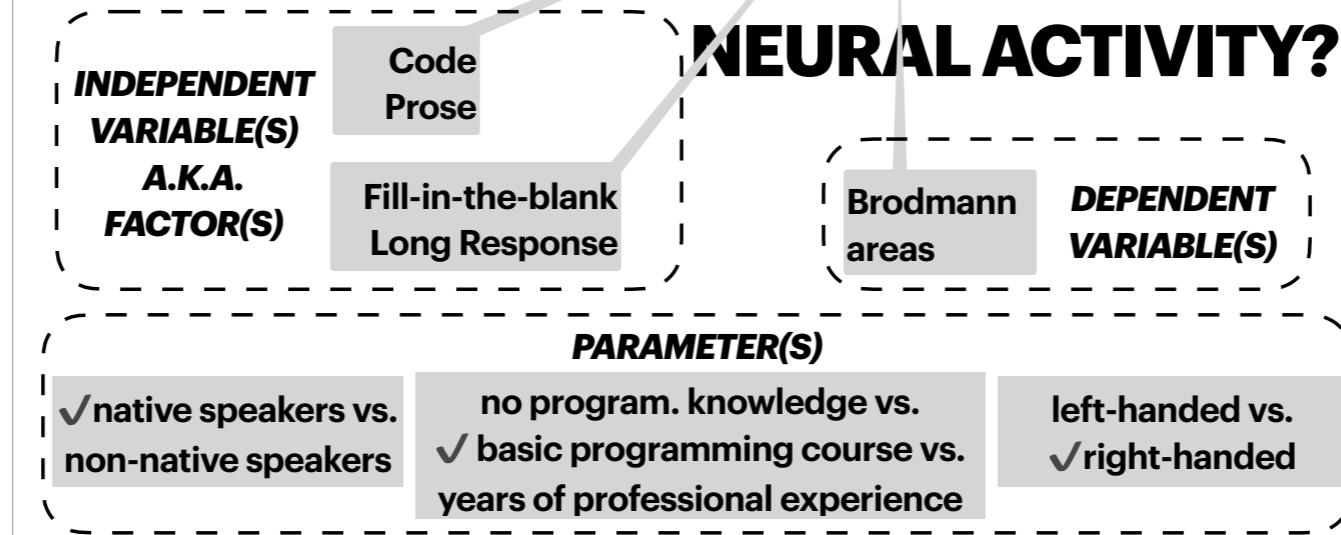
EXPERIMENT



RYAN KRUEGER, YU HUANG, XINYU LIU, TYLER SANTANDER, WESTLEY WEIMER, KEVIN LEACH. NEUROLOGICAL DIVIDE - AN FMRI STUDY OF PROSE AND CODE WRITING. INT CONF SOFTWARE ENGINEERING 2020

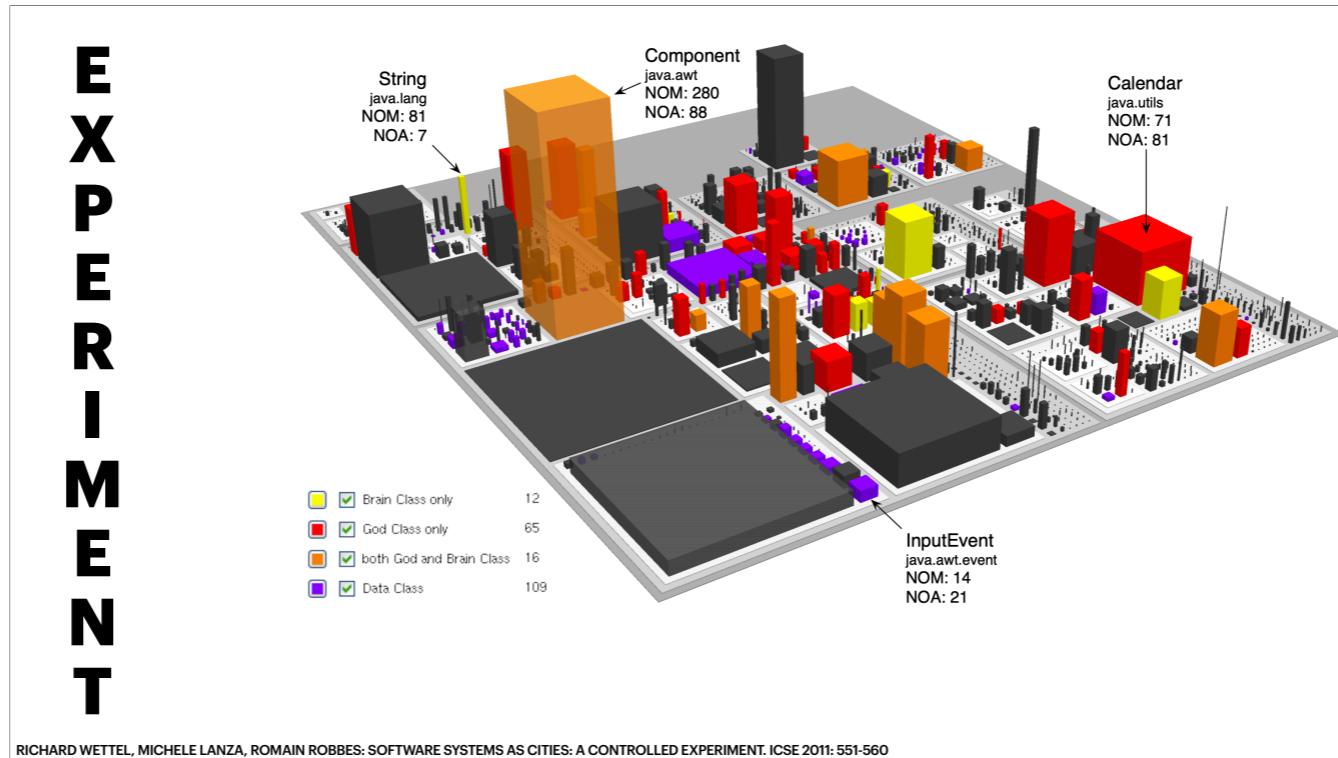
You might remember this example from the first lecture. The authors used fMRI to study whether code writing and prose writing are similar in terms of brain area activation.

DO DIFFERENT TYPES OF WRITING EXHIBIT THE SAME PATTERNS OF NEURAL ACTIVITY?



This another category of variables, we call them parameters. We do not want them to influence the dependent variables. Hence, we fix the values of these variables during the experiment. In the study of Kruger et al. this means that the authors have *only* included native speakers, that have were students having a basic programming course as a background, and that are right-handed. This means that conclusions of this study cannot be generalised to non-native speakers, professionals or left-handed individuals.

EXPERIMENT

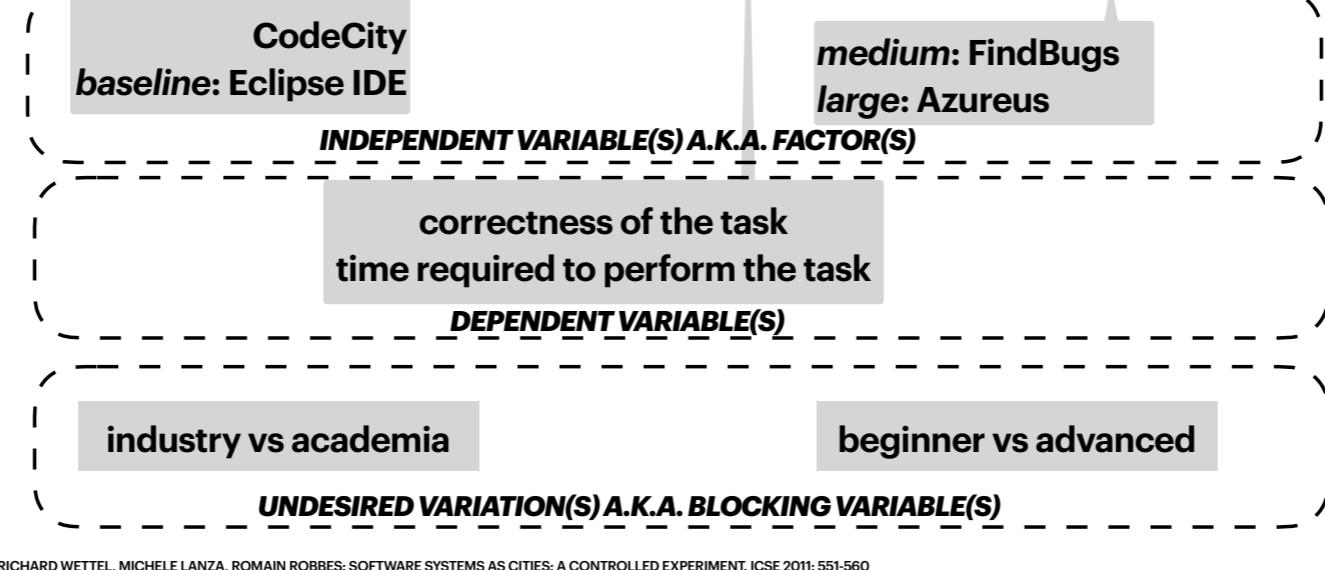


This image is a visualisation of a software system. The visualisation technique is known as CodeCity.

- The Number Of Methods is mapped on the height of the buildings,
- the Number Of Attributes on the base size,
- the color is presence of specific code smells

This looks pretty of course but is this also useful?

WHICH TOOL IS THE BEST FOR WHICH SYSTEM?



We can expect that people having more experience will be more successful in performing the tasks both in terms of correctness and in terms of time. One might consider these variables as parameters, and, e.g., focus only on beginner programmers from academia. However, this is not always possible. For example, it might be hard to find sufficiently many subjects with similar characteristics, or it might not be possible to get very similar projects on which to apply the different alternatives.

An aeronautics software development laboratory aims to identify the best two of four possible programming languages (Pascal, C, PL/M and FORTRAN) in terms of productivity, which are to be selected to implement two versions of the same flight control application so that if one fails the other comes into operation. There are 12 programmers and 30 modules with similar functionalities to flight control applications for the experiment.

The individual productivity of each programmer differs, which could affect the experiment productivity.

- (A) DEPENDENT: PROGRAMMING LANGUAGE, BLOCKING: PROGRAMMERS**
- (B) INDEPENDENT: PRODUCTIVITY, PARAMETER: SAME APPLICATION**
- (C) PARAMETER: SAME FUNCTIONALITY, BLOCKING: PROGRAMMERS**
- (D) INDEPENDENT: PROGRAMMERS, BLOCKING: SAME FUNCTIONALITY**

NATALIA JURISTO, ANA M. MORENO. BASICS OF SOFTWARE ENGINEERING EXPERIMENTATION. 2001

C - the detailed answer on the next slide

An aeronautics software development laboratory aims to identify the best two of four possible **programming languages** (Pascal, C, PL/M and FORTRAN) in terms of **productivity**, which are to be selected to implement two versions of **the same flight control application** so that if one fails the other comes into operation. There are **12 programmers** and 30 **modules** with **similar functionalities** to flight control applications for the experiment.

The individual productivity of each programmer differs, which could affect the experiment productivity.

independent variable(s), parameter(s), blocking variable(s) and dependent variable(s)

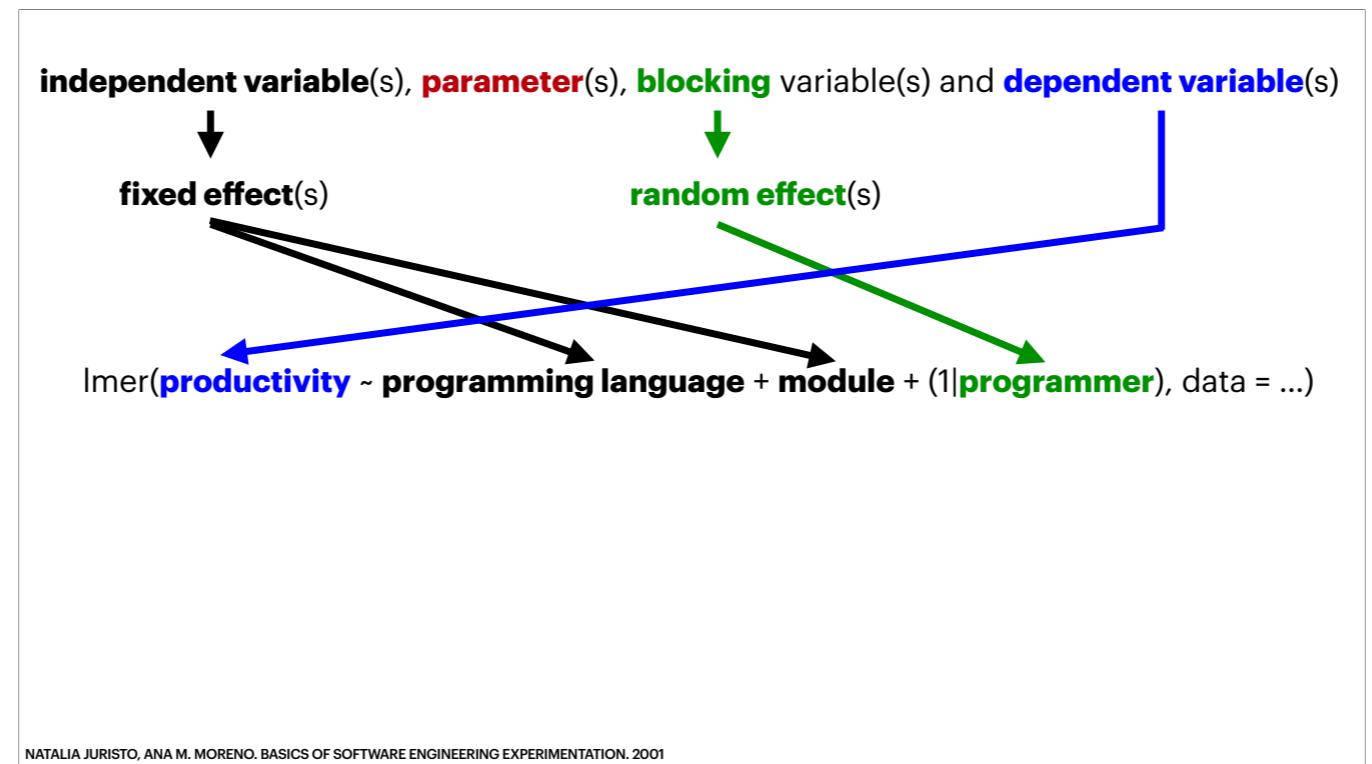
NATALIA JURISTO, ANA M. MORENO. BASICS OF SOFTWARE ENGINEERING EXPERIMENTATION. 2001

Productivity careful: they are looking for some kind of average productivity.

Dependent variable: mean productivity in terms of months/person, for example.

Programmers are blocking since their individual productivity differs and we cannot control it.

A unitary experiment would involve the implementation of one of the modules by one of the subjects in a given language.

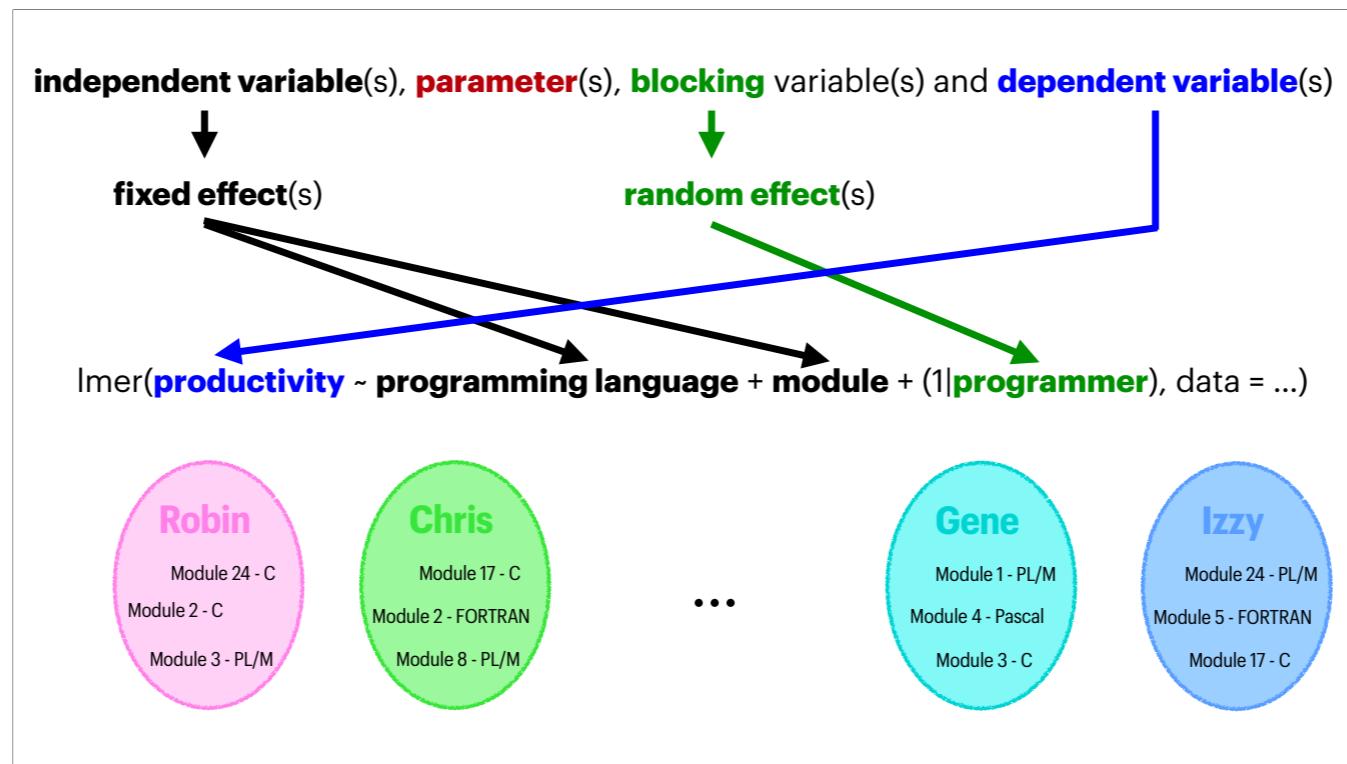


and now we can get back to mixed effects modelling

Since the parameters are fixed we do not include them in the modelling.

This is a linear mixed-effects model; there are also other kinds of models.

The code is in R, I am not sure how this would work with other statistical tools or Python...



How is this related to groups of observations? We can see blocking variables/random effects as groups of observations. For example, “programmer” is a group of observations corresponding to all the implementation tasks performed by the programmer. In the same way we can see a project as group of commits or a code review as a group of comments. We can even have more nesting if we add more random effects!

```
lmer(log(opened_pr + 0.5) ~  
    time  
    + intervention  
    + time_after_intervention  
    + age_at_bot  
    + log(total_number_pr_authors)  
    + log(commits)  
    + (1 | name)  
    + (1 | lang),  
    data=data)
```

MAIRIELI SANTOS WESSEL, ALEXANDER SEREBRENIK, IGOR WIESE, IGOR STEINMACHER, MARCO AURÉLIO GEROSA: EFFECTS OF ADOPTING CODE REVIEW BOTS ON PULL REQUESTS TO OSS PROJECTS. ICSME 2020: 1-11

Moreover, we can also combine this approach with RDD. For example, Mairieli Wessel and her coauthors wanted to understand the impact of bots on the opening of pull requests. name is project name, lang is the dominant programming language, RDD-related variables are typeset in magenta

Experience Sampling

In which activity are you involved?

How do you feel now?

| | Annoyed | Calm | Excited | In control | |
|---|---------|------|---------|------------|---------|
| 1 | | | | | 2 |
| 3 | | | | | 4 |
| 5 | | | | | Pleased |

My productivity is: Very low Below average Average Above average Very high

Notes (optional)

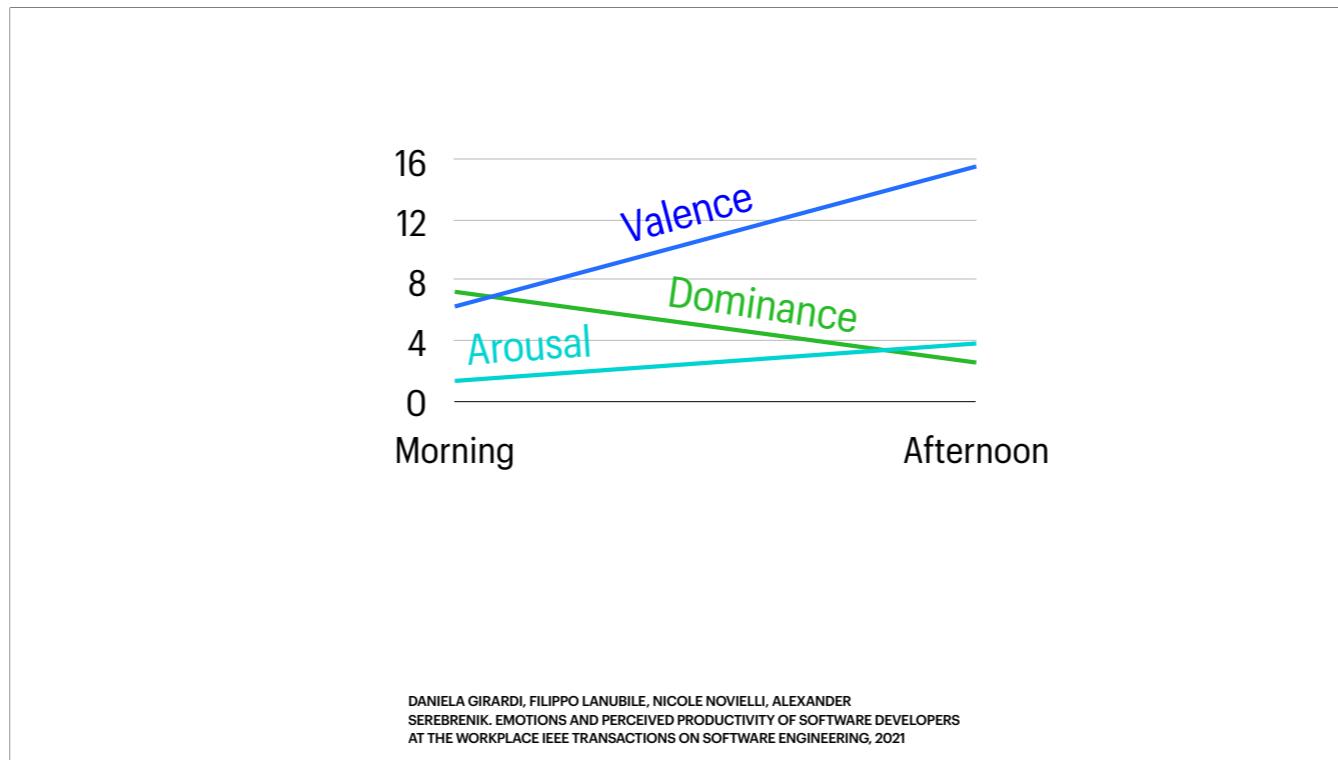
Did you experience anything that might have affected your emotion during the last session?

Done

DANIELA GIRARDI, FILIPPO LANUBILE, NICOLE NOVIELLI, ALEXANDER SEREBRENIK. EMOTIONS AND PERCEIVED PRODUCTIVITY OF SOFTWARE DEVELOPERS AT THE WORKPLACE IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, 2021

lmer(**productivity** ~
+ **valence**
+ **arousal**
+ **dominance**
+ **ampm**
+ **valence:ampm**
+ **arousal:ampm**
+ **dominance:ampm**
+ **(1 | person)**
+ **(1 | company)**,
data=data)

In another study we wanted to understand how self-perceived emotion experienced by developers is related to the self-perceived productivity. In this study we have observed developers at their working space. So we cannot exclude that the perceived productivity can be impacted by time, e.g. due to fatigue. Therefore, time and its interaction with the emotional dimensions are also included in the model as fixed effects. “Interaction” effects mean that the same independent variable (e.g., valence) affects productivity differently in the morning than in the afternoon. We use dark red to indicate interactions.



And indeed we have observed differences between the percentage of variance of productivity explained by valence, arousal and dominance in the morning vs in the afternoon. What we see is that valence becomes more important in the afternoon and dominance less important. This could be due to fatigue, which is known to impair emotion regulation: in the afternoon people are more tired and their emotions influence perceived productivity more.



A comparison of the same scene photographed with a mobile phone (Nokia 6303i classic) and a DSLR camera (Olympus E-520). Both devices used automatic settings. The resulting photographs have been juxtaposed using GIMP.

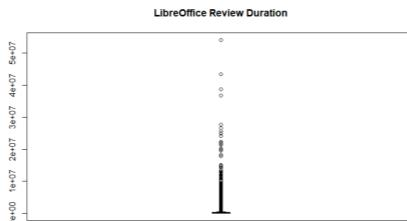


When reporting any kind of linear regression models it is customary to report the coefficient of determination R^2 . In case of mixed effects it is important to report by R^{2m} and R^{2c} . Usually R^{2c} is much higher than R^{2m} since there is indeed difference between the groups that can be attributed to their inherent characteristics and cannot be captured by the fixed effects. If R^2 is very high then the model can be overfitted; if R^2 is very low than it is barely adequate.



Sorry I could not resist

Figure 3: Boxplot for Review Duration of the Libre Office Dataset



Data is very **skewed**: log transform
Zeroes? Add 0.5: $\log(x+0.5)$

```
Residual standard error: 1.082 on 11610 degrees of freedom  
Multiple R-squared:  0.0388,   Adjusted R-squared:  0.03839  
F-statistic: 93.74 on 5 and 11610 DF,  p-value: < 2.2e-16
```

Figure 4: Summary of the model of LibreOffice.

Model should not fail the **F-statistic**
test, but low p-value is not enough

```
> vif(fit)
libreofficeData$linoff_linesins_no_tf  libreofficeData$linoff_linesdel_no_tf  libreofficeData$linoff_linesmod_no_tf
1.346664 1.227526 1.383563
libreofficeData$linoff_abz_no_tf  libreofficeData$linoff_abr_no_tf  libreofficeData$linoff_ra_no_tf
10.547333 5.532017 15.457796
libreofficeData$linoff_rt_no_tf  libreofficeData$linoff_ca_no_tf  libreofficeData$linoff_et_no_tf
33.741779 6.267548 28.653275
```

Figure 16: Results from VIF for Libre Office

Stepwise: VIF > 5, check pairwise correlations to exclude one of the strongly correlated variables, refit a model without the variable and recompute VIFs.

93

Variance inflation factor

In statistics, multicollinearity (also collinearity) is a phenomenon in which one predictor variable in a multiple regression model can be linearly predicted from the others with a substantial degree of accuracy. In this situation, the coefficient estimates of the multiple regression may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set; it only affects calculations regarding individual predictors. That is, a multivariate regression model with collinear predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give valid results about any individual predictor, or about which predictors are redundant with respect to others.

The image consists of three main parts:

- Top Left:** A screenshot of a presentation slide titled "COMPARING MULTIPLE GROUPS". It features four yellow callout boxes with effect sizes and descriptions:
 - "-0.49 (Small)***"
 - "0.70 (Medium)**"
 - "1.55 (Large)***"
- Top Right:** A list of four questions about comparing multiple groups:
 1. NORMALITY TEST
 2. ARE THERE DIFFERENCES SOME GROUPS?
 3. WHERE ARE THE DIFFERENCES (PAIRWISE TESTS)?
 4. HOW BIG ARE THESE DIFFERENCES (EFFECT SIZE)?
- Bottom Left:** A text block describing an aeronautics software development laboratory's goal to identify the best possible programming languages (Pascal, C, PL/M and FORTRAN) in terms of productivity.
- Bottom Right:** A text block about individual programmer productivity differing, which could affect experiment productivity.
- Code Snippet:** A block of R code for a mixed effects model:

```
lmer(log(opened_pr + 0.5) ~ time + intervention + time_after_intervention + age_at_bot + log(total_number_pr_authors) + log(commits) + (1 | name) + (1 | lang), data=data)
```
- Command Line:** A block of R code for a mixed effects model:

```
lmer(productivity ~ + valence + arousal + dominance + ampm + valence:ampm + arousal:ampm + dominance:ampm + (1 | person) + (1 | company), data=data)
```

We have discussed how to compare multiple groups. We have started with a traditional approach to comparison of multiple groups. Then we have discussed different kinds of variables and the ways mixed effects modelling can be carried out in R and what should you be aware of when performing mixed effect modelling