Introduction to methods for digital humanities

Eetu Mäkelä, D.Sc.
Professor (tenure track) in Humanities–Computing Interaction / University of Helsinki
Docent (Adjunct Professor) in Computer Science / Aalto University
Debriefing of the assignments

• Ready tools:
  • OpenRefine
  • Visualisation tools
• Fundamental concepts of programming
  • Tutorial
  • Libraries
  • Figuring out exercise
• Regular expressions
  • Tutorial
  • Exercise
Reading assignment: The civilizing process in London’s Old Bailey

- Explain, in layman’s terms, what is being measured in Figure 1?
- Explain, in layman’s terms, what is being measured in Figure 2?
- What is the core argument/contribution of the paper?
Reading assignment: The civilizing process in London’s Old Bailey

• What is good about the methods (or how they’re applied) in the article?
• Are there problems in the methods, or how they’re applied?
• What is good about the way the paper contributes to its subject field?
• Are there problems in the way the paper contributes to its subject field?
Fundamental concepts of statistics

Eetu Mäkelä, D.Sc.
Professor (tenure track) in Humanities–Computing Interaction / University of Helsinki
Docent (Adjunct Professor) in Computer Science / Aalto University
Uses for statistics

- Descriptive statistics: summarize a large amount of data into a smaller set of numbers
- Inferential statistics:
  - Evaluate relationships between phenomena
  - Discover underlying models behind a phenomena
  - Estimate the trustworthiness of a claim based on data
Descriptive statistics
Distribution: the set of values you calculate from your data:

### Distribution over people

<table>
<thead>
<tr>
<th>Person</th>
<th>Letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Galileo Galilei</td>
<td>60</td>
</tr>
<tr>
<td>Geri Bocchineri</td>
<td>30</td>
</tr>
</tbody>
</table>

### Distribution over places

<table>
<thead>
<tr>
<th>Place</th>
<th>Letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rome</td>
<td>90</td>
</tr>
<tr>
<td>Florence</td>
<td>30</td>
</tr>
</tbody>
</table>

### Distribution over people + places

<table>
<thead>
<tr>
<th>Person</th>
<th>Place</th>
<th>Letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Galileo Galilei</td>
<td>Florence</td>
<td>30</td>
</tr>
<tr>
<td>Geri Bocchineri</td>
<td>Rome</td>
<td>60</td>
</tr>
<tr>
<td>Galileo Galilei</td>
<td>Rome</td>
<td>30</td>
</tr>
</tbody>
</table>
Summary statistics distil data into fewer descriptive numbers

Distribution over people + places

<table>
<thead>
<tr>
<th>Person</th>
<th>Place</th>
<th>Letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Galileo Galilei</td>
<td>Florence</td>
<td>30</td>
</tr>
<tr>
<td>Geri Bocchineri</td>
<td>Rome</td>
<td>60</td>
</tr>
<tr>
<td>Galileo Galilei</td>
<td>Rome</td>
<td>30</td>
</tr>
</tbody>
</table>

Average letters per origin:
\[
\frac{(30+60+30)}{3} = 40
\]
Summary statistics distil data into fewer descriptive numbers

Distribution over people + places

<table>
<thead>
<tr>
<th>Person</th>
<th>Place</th>
<th>Letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Galileo Galilei</td>
<td>Florence</td>
<td>30</td>
</tr>
<tr>
<td>Geri Bocchineri</td>
<td>Rome</td>
<td>60</td>
</tr>
<tr>
<td>Galileo Galilei</td>
<td>Rome</td>
<td>30</td>
</tr>
</tbody>
</table>

Average letters per origin:
\[
\frac{(30+60+30)}{3} = \frac{2\times30+60}{3} = 40
\]

Distribution over number of letters per people/place

<table>
<thead>
<tr>
<th>Letters</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>60</td>
<td>1</td>
</tr>
</tbody>
</table>
Summary statistics distil data into fewer descriptive numbers

<table>
<thead>
<tr>
<th></th>
<th>Lifetimes</th>
<th>Number of letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>69.12</td>
<td>15.94</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>16.46</td>
<td>30.44</td>
</tr>
</tbody>
</table>

Distribution over number of letters per people/place

<table>
<thead>
<tr>
<th>Letters</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>60</td>
<td>1</td>
</tr>
</tbody>
</table>
**Terminology**

**Average:** a single number that somehow summarizes the typical/ideal/central member in a distribution (e.g. mean, median, mode, …)

**Variability:** a number that somehow summarizes how varied and spread apart the members of a distribution are (e.g. standard deviation, variance, median absolute deviation, …)

**Skewness:** a number that somehow summarizes if more members of the distribution are to the left or to the right of the average
**Terminology**

**Average:** a single number that somehow summarizes the typical/ideal/central member in a distribution (e.g. mean, median, mode, …)

**Variability:** a number that somehow summarizes how varied and spread apart the members of a distribution are (e.g. standard deviation, variance, median absolute deviation, …)

**Skewness:** a number that somehow summarizes if more members of the distribution are to the left or to the right of the average
**Terminology**

**Average**: a single number that somehow summarizes the typical/ideal/central member in a distribution (e.g. mean, median, mode, ...)

**Variability**: a number that somehow summarizes how varied and spread apart the members of a distribution are (e.g. standard deviation, variance, median absolute deviation, ...)

**Skewness**: a number that somehow summarizes if more members of the distribution are to the left or to the right of the average
## What smells here?

<table>
<thead>
<tr>
<th></th>
<th>Lifetimes</th>
<th>Number of letters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average</strong></td>
<td>69.12</td>
<td>15.94</td>
</tr>
<tr>
<td><strong>Standard deviation</strong></td>
<td>16.46</td>
<td>30.44</td>
</tr>
</tbody>
</table>
What smells here?

<table>
<thead>
<tr>
<th></th>
<th>Lifetimes</th>
<th>Number of letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>69.12</td>
<td>15.94</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>16.46</td>
<td>30.44</td>
</tr>
</tbody>
</table>
What smells here?

<table>
<thead>
<tr>
<th></th>
<th>Lifetimes</th>
<th>Number of letters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average</strong></td>
<td>69.12</td>
<td>15.94</td>
</tr>
<tr>
<td><strong>Standard deviation</strong></td>
<td>16.46</td>
<td>30.44</td>
</tr>
</tbody>
</table>

= assuming a normal distribution, about 70% of sources send between -15 and 45 letters, and about 95% send between -45 and 75.
Summary statistics about number of letters per source city to Galileo Galilei

- **Average** Mean: 15.94, Median: 4, Mode: 1
- **Min**: 1, **Max**: 116
- First quartile: 1, Second quartile 4, Third quartile: 11
- **Standard deviation**: 30.44, Median absolute deviation: 4.45
- **Nonparametric skewness**: 0.39, **Mode skewness**: 0.49
Number of letters per source city to Galileo Galilei
Number of letters per source city to Galileo Galilei

Histogram of letters by city's letters

Frequency

0 2 4 6 8 10 12 14

letters by city's letters

0 20 40 60 80 100 120

Count

0 1 2 3 4 5

number of letters

0 30 60 90 120

n
Number of letters per source city to Galileo Galilei – actual counts
Terminology

- **Unimodal, Bimodal, Multimodal**: how many separate modes (most common values) does the distribution have?
- **Parametric/nonparametric**: can the distribution be described using an equation and some parameters?
Summary statistics work with the distribution they were designed for:

<table>
<thead>
<tr>
<th></th>
<th>Lifetimes</th>
<th>Number of letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>69.12</td>
<td>15.94</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>16.46</td>
<td>30.44</td>
</tr>
</tbody>
</table>

Histogram of \(d2\$Lifetime\):
- Normal, unimodal, parametric

Histogram of \(letters\_by\_city\$letters\):
- Nonparametric, multimodal
Types of variables

- **NOMINAL**: categories, e.g. social classes, cities, countries
- **INTERVAL**: most numbers (technically where it makes sense to measure differences), e.g. speeds, ages, heights, weights, temperatures, numbers of occurrences
  - Not an interval: 1=yes, 2=no, 3=don’t know
- **ORDINAL**: categories that can be ordered, e.g. highly educated, uneducated; bad, okay, good, best
Summary statistics work with the type of variable they were designed for:

<table>
<thead>
<tr>
<th></th>
<th>Lifetimes</th>
<th>Number of letters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>69.12</td>
<td>15.94</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>16.46</td>
<td>30.44</td>
</tr>
</tbody>
</table>

Interval

Nominal

Mode of cities = Florence
Are these scales ordinal or interval? = does the difference between numbers make sense and stay stable? Does an average of the numbers make sense?

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

1. Poor
2. Average
3. Good
4. Very Good

★, ★★, ★★★, ★★★★, ★★★★★, ★★★★★★
Takeaway message

Distributions in the humanities are often weird.

One cannot trust commonly used statistical summaries to work.

It’s okay (actually the best) to do without them if you can.
Sociolinguistic variation in noun ratio over time: gender

Noun\% vs. Time Periods:

- Women
- Men

Time Periods:
- (1415,1453]
- (1453,1491]
- (1491,1529]
- (1529,1567]
- (1567,1605]
- (1605,1643]
- (1643,1681]
Research process

1. Have data
2. Magic (?)
   • Hedge magic (spreadsheets, Excel graphs)
   • **Common ritual magic (statistics: correlation, regression, ANOVA, PCA)**
     • Relatively simple, commonly understood formulae you could mostly go through with pen and paper if you wanted to
     • Higher ritual magic (SVM, LSA, LDA, SnE)
       • More complex, harder to follow formulae, impossible to work through manually
     • Black magic (most machine learning, neural networks)
       • You feed the machine both an input and a desired output, it derives a mostly unintelligible black box that links the two
   • Flashy magic (proper visualizations)
3. Something interesting shows up
4. Profit!
Uses for statistics

- **Descriptive statistics**: summarize a large amount of data into a smaller set of numbers
- **Inferential statistics**:  
  - Evaluate relationships between phenomena  
  - Discover underlying models behind a phenomena  
  - Estimate the trustworthiness of a claim based on data
Correlation: do values (e.g. social class and eloquence) change together?
About correlations

• Correlation does not imply causation
  • Chance
  • Reverse causation
  • Third cause
  • Bidirectional causation
• http://tylervigen.com/spurious-correlations
Correlation measures a linear relationship

correlation: 0.816
Regression: trying to figure out how exactly a variable depends on another

- Explained visually
Research process

1. Have data
2. Magic (?)
3. Something interesting shows up ← 3 1/2. Evaluate if the interesting something IS REALLY THERE
4. Profit!
Confidence: “Based on the data I have, I’m 95% confident that between 10 and 100 people sent Galileo letters in 1855”

Significance: Given different values (with associated confidences), how likely is it that their difference is only due to chance? (e.g. women don’t curse more than men even when my data says otherwise)

Statistical test: Most often, test of significance
- **Confidence:** “Based on the data I have, I’m 95% confident that between 10 and 100 people sent Galileo letters in 1855”
- **Significance:** is what I see there just by chance?
- If the data follows a *known parametric distribution* and is *randomly sampled*, confidence intervals and significance are easy to calculate.
- In the humanities, neither often holds.
Solution for wonky distributions: bootstrapping

- Confidence is estimated by resampling with replacement the data numerous times and examining the distributions witnessed in the samples.
- The benefit of this is that we do not have to know the exact nature of the distribution in the data to make confidence estimates.
- The downside is that it is highly fragile to biases in the original sample.
Bootstrapping explained visually: proportion of orange books

All books published in the 18th Century

Our collection of books

1 2 3

4 5 6

2/6 = 33%
Bootstrapping explained visually: proportion of orange books

All books published in the 18th Century

Our collection of books

Samples with replacement from our collection

2/6 = 33%

2, 6, 4 = 33%

3, 1, 5, 3 = 33%

3, 5, 1, 2 = 17%

3, 1, 2 = 33%

6, 6, 1, 6 = 66%

1, 5, 5 = 0%

1, 2, 1 = 50%

1, 3, 5

17%
Bootstrapping explained visually: proportion of orange books

All books published in the 18th Century

Our collection of books

Samples with replacement from our collection

With 95% confidence, the proportion of orange books in the 18th Century was between 17% and 50%
Bootstrapping explained visually

- http://www.lock5stat.com/StatKey/bootstrap_1_quant/bootstrap_1_quant.html
- https://www.stat.auckland.ac.nz/~wild/BootAnim/movies/bootstrap2.mp4
- https://www.stat.auckland.ac.nz/~wild/BootAnim/
Bootstrapping
Controlling for false discoveries

Bonferroni Correction: 
\[ p = \frac{p}{n} \]

False Discovery Rate: 
proportion of false discoveries

The probability of at least one false positive in 20 experiments each having a 5% chance of a false positive is 
\[ 1 - 0.95^{20} = 64.15\% \]
Statistics in the humanities:

1. Don’t use statistical summaries when you can just chart the original data
2. Either don’t use fancy inferential statistics, or cooperate with someone with a DEEP understanding of statistics beyond the normal distribution (and tell everyone who tries to question you that doing statistics well in the humanities is a lot harder than they think)
3. (Using bootstrapping to calculate confidence intervals and then just graphing them for evaluation is a nice current general purpose solution that is both easily understandable as well as robust)
Further reading

• **Statistics for the Humanities**
  • Good, easy to read, concise book, **BUT** doesn’t question or highlight reliance on the normal distribution, which is a big problem.

Assignments for next weeks

@ https://jiemakel.gitbook.io/meth4dh/course-instances/helsinki-fall-2018
eetu.makela@helsinki.fi
http://iki.fi/eetu.makela
http://presemo.helsinki.fi/meth4dh