Methodology, Statistics and Pitfalls

Data Science and Society Statistics Lecture 1

Matthieu Brinkhuis

October 4, 2018

Utrecht University, Information and Computing Sciences
m.j.s.brinkhuis@uu.nl

Introduction

Topics of today

Topics we will *discuss* today:

- Methodological issues in data analysis
- Some pitfalls

Intended learning outcomes

By the end of this lecture, you'll be able to:

- Recognize that methodology for data science is crucial (1)
- Understand different types of pitfalls (2)
- Apply the principles in your own research (3)
- Analyze potential traps (4,5)

Bloom's Taxonomy:

1. Remember
2. Understand
3. Apply
4. Analyze
5. Evaluate
6. Create

Table of Contents

Structured method for conducting analyses

Structured method helpful in preventing methodological errors.

CRISP-DM (Chapman et al., 2000, p. 12):

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

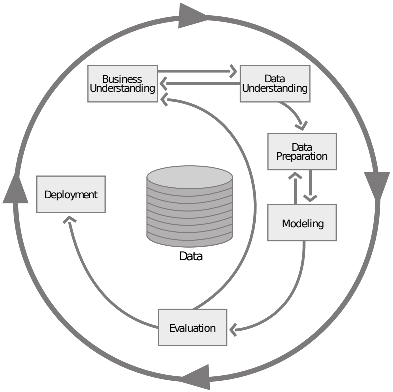

Figure 1: CRISP-DM Process Diagram

Hypothesis testing

Null Hypothesis Significance Testing (NHST)

Traditional steps (I Miller and M Miller, 2014):

1. Formulate H_0 and H_a (e.g., $H_0 : \mu = \mu_0$, $H_a : \mu \neq \mu_0$).
2. Using the sampling distribution of the (appropriate) test statistics, determine critical region of size α.
3. Determine the value of the test statistics from the sample data.
4. Check if it falls in the critical region (reject H_0) our outside (retain H_0).

NHST ii

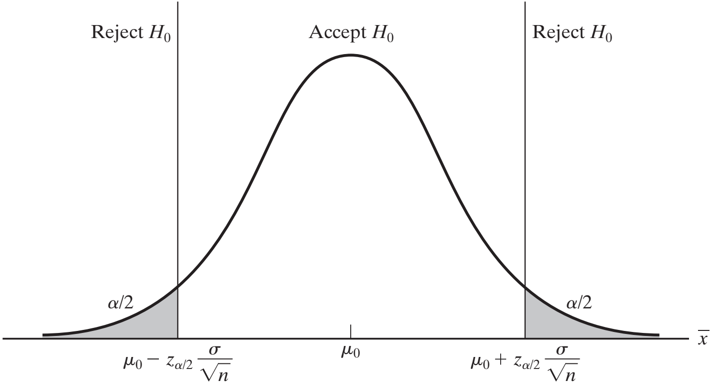

Figure 2: Critical region for two-tailed test (I Miller and M Miller, 2014, p. 360).

What is a p-value?

Depends on who you ask.

- Frequentist: limiting relative frequency, if you could repeat the experiment.
- Bayesian: subjective, degree of belief, personally defined.

Back to the example: if $p \leq \alpha$, the observed data is inconsistent with the null hypothesis, so the null hypothesis must be rejected. Does not prove that the tested hypothesis is true. Guarantees that the Type I error (false positive) rate is at most α .

NHST iv

Problems with $p \leq .05$

Traditionally $\alpha = .05$, but there are calls for change.

“We propose to change the default P-value threshold for statistical significance from 0.05 to 0.005 for claims of new discoveries.” (Benjamin et al., 2018).

Why do you think that is?

Multiple testing

The problem of multiple testing

“When pursuing multiple inferences, researchers tend to select the (statistically) significant ones for emphasis, discussion and support for conclusions. An unguarded use of single-inference procedures results in a greatly increased false positive (significance) rate.” (Benjamini and Hochberg, 1995, p. 289)

How to deal with false discoveries? Be aware! Correction possible, e.g. Bonferroni and Hochberg methods, see R function `p.adjust{stats}`.

Prediction

Prediction vs inference

Statistical learning

$$Y = f(X) + \epsilon \tag{1}$$

where f is unknown function of $X_1, X_2, \ldots, X_p$ and ϵ is a random error term, independent of X (James et al., 2013).

Estimate $\hat{f}$ for:

- prediction (black box)
- inference (interest in associations, what is the type of relationship, etc.)

There is a trade-off between prediction accuracy and model interpretability (Waa et al., 2018; James et al., 2013).

Predicting peculiarities

In prediction (Kaggle?), what do you predict against?

- role of testing data
- how much is there to gain? baseline construction? (burglary)
- from 90% to 100% can be a long way (reducible vs irreducible error)
- how to quantify prediction quality? (model accuracy)
- dynamic prediction (feedback loops, fraud prediction and Netflix challenge)

Training and validating

Validation set approach

Basic idea is to split your data set in two parts (e.g., James et al., 2013, p. 176):

- training set (to fit the model)
- validation set (to evaluate the model)

Only a subset used, test error can be variable. Different cross-validation approaches possible (leave-one-out, k-fold, bootstrap, etc.).

Measuring the quality of fit

- mean squared error $= \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$
- error rate $= \frac{1}{n}\sum_{i=1}^{n}(y_i \neq \hat{y}_i)$

However, there are many more ways, especially in classification.

Classification prediction quality: discrete

		True class: Positive	True class: Negative	Measures
Predicted class	Positive	True positive *TP*	False positive *FP*	Positive predictive value (PPV) $\frac{TP}{TP+FP}$
	Negative	False negative *FN*	True negative *TN*	Negative predictive value (NPV) $\frac{TN}{FN+TN}$
Measures		Sensitivity $\frac{TP}{TP+FN}$	Specificity $\frac{TN}{FP+TN}$	Accuracy $\frac{TP+TN}{TP+FP+FN+TN}$

Figure 3: Confusion matrix (see Wikipedia for more).

Classification prediction quality: discrete

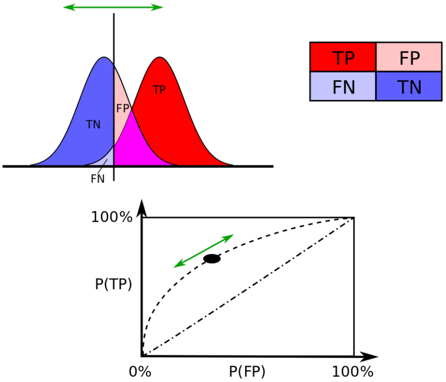

Figure 4: ROC analysis (from Wikipedia ROC).

Inference and Causality

Exploratory vs. confirmatory analysis

Important distinction in statistics. Why would you do exploratory analysis? Why confirmatory analysis?

Using this distinction: what is data mining?

Related concepts:

- Spurious relations
- Coincidence
- Causal relationships (or lack thereof)
- Fishing

Spurious relations

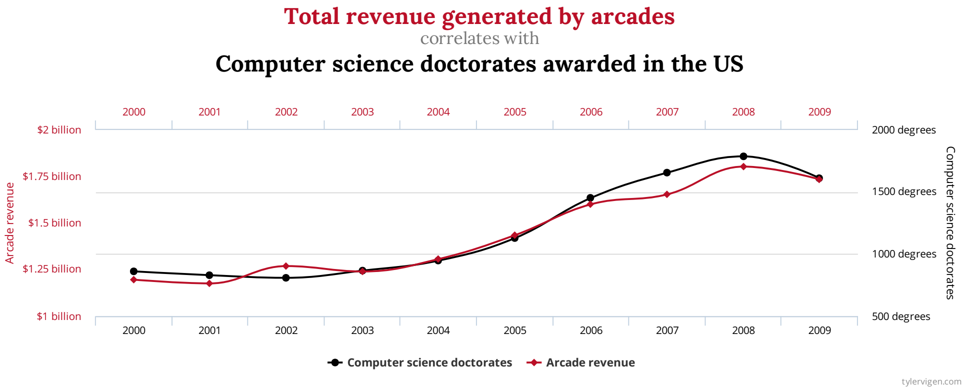

Figure 5: Spurious Correlation example, see tylervigen.com

Related problems

So beware of causal statements, here it seems easy, but if you relate ’obvious’ data, you easily fall for it!

Pitfalls include:

- Simpson’s paradox
- 3rd variable
- Anscombe’s quartet (next)
- Lack of theory
- Lack of experiment

Many things interesting, without making causal statements.

Graphs are essential

“Graphs are essential to good statistical analysis” (Anscombe, 1973)

Four data sets:

- $n = 11$
- $\bar{x} = 9.0$
- $\bar{y} = 7.5$
- $y = 3 + .5x$
- Multiple $R^2 = .667$

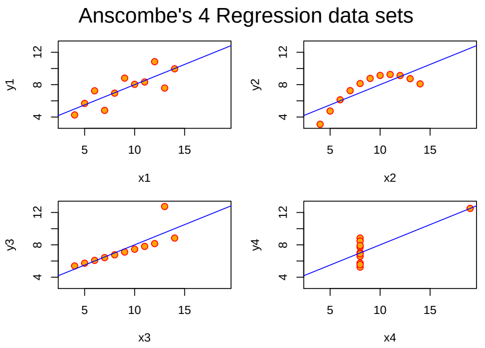

Figure 6: Anscombe’s Quartet, obtained by `example(anscombe)`

Simpson's Paradox i

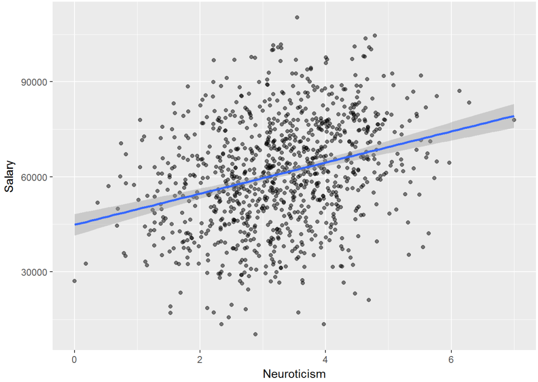

Figure 7: A clear positive relation...

Simpson's Paradox ii

How is this possible?

- Graphs from Paul van der Laken (rpubs.com/lakenp)
- Further reading: Simpson (1951) and Kievit, Frankenhuis, et al. (2013)
- R Package: Simpsons (Kievit and Epskamp, 2012)

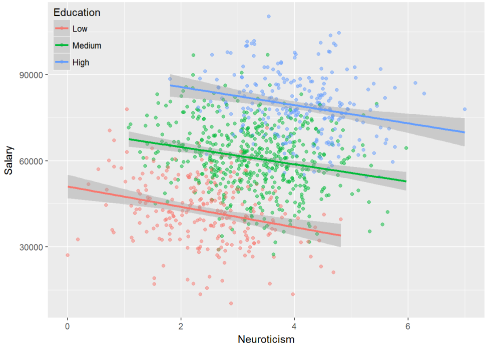

Figure 8: ...until conditioning.

Data quality

Get a grip on the quality of your data

Data quality is a topic by itself, here some thoughts:

- Garbage-in garbage-out principle
- Examples of measures: completeness, validity, accuracy, consistency, availability and timeliness
- Beware of missing data:
  - what deleting or disregarding data can do to your research
  - but what to do then?
  - explicit assumptions
  - model when needed

Sometimes you cannot do without models

Sometimes your data is model generated, example:

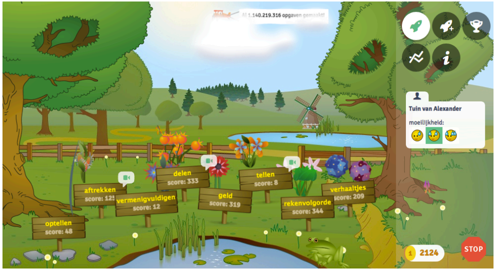

Figure 9: Math garden landing page.

Figure 10: Math garden practice item.

Closing

Next meeting

Practical remarks:

- Read article *and* the response letters on the Google Flu (Lazer et al., 2014a; Broniatowski, Paul, and Dredze, 2014; Lazer et al., 2014b)
- Read article on p-values (Benjamin et al., 2018)
- Tutorial on data analysis in R and Spark

Intended learning outcomes recap

By the end of this lecture, you'll be able to:

- Recognize that methodology for data science is crucial (1)
- Understand different types of pitfalls (2)
- Apply the principles in your own research (3)
- Next lecture: *Analyze potential traps* (4,5)

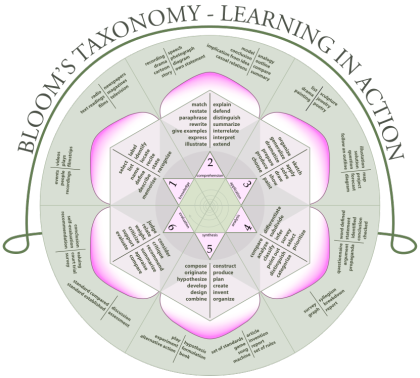

Figure 11: Bloom's Taxonomy (image from Wikipedia).

Thank you.

References i

- Anscombe, FJ (1973). “Graphs in Statistical Analysis”. In: *The American Statistician* 27.1, pp. 17–21. DOI: 10.2307/2682899. JSTOR: 2682899 (cit. on p. 24).
- Benjamin, DJ et al. (2018). “Redefine statistical significance”. In: *Nature Human Behaviour* 2.1, p. 6. DOI: 10.1038/s41562-017-0189-z (cit. on pp. 11, 31).
- Benjamini, Y et al. (1995). “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. English. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1, pp. 289–300. URL: http://www.jstor.org/stable/2346101 (cit. on p. 12).

References ii

- Broniatowski, DA et al. (2014). “Twitter: Big data opportunities”. In: *Science* 345.6193, p. 148. DOI: 10.1126/science.345.6193.148-a. eprint: http://www.sciencemag.org/content/345/6193/148.1.full.pdf (cit. on p. 31).
- Chapman, P et al. (2000). *CRISP-DM 1.0. Step-by-step data mining guide*. Tech. rep. The CRISP-DM consortium. URL: ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf (cit. on p. 6).
- James, G et al. (2013). *An introduction to statistical learning*. Springer Texts in Statistics. Springer. DOI: 10.1007/978-1-4614-7138-7 (cit. on pp. 14, 16).

References iii

- Kievit, RA, WE Frankenhuis, et al. (2013). “Simpson’s paradox in psychological science: a practical guide”. In: *Frontiers in Psychology* 4. DOI: `10.3389/fpsyg.2013.00513` (cit. on p. 26).
- Kievit, RA et al. (2012). *Simpsons: Detecting Simpson’s Paradox*. R package version 0.1.0. URL: `https://CRAN.R-project.org/package=Simpsons` (cit. on p. 26).
- Lazer, D et al. (2014a). “The Parable of Google Flu: Traps in Big Data Analysis”. In: *Science* 343.6176, pp. 1203–1205. DOI: `10.1126/science.1248506` (cit. on p. 31).

- Lazer, D et al. (2014b). "Twitter: Big data opportunities - Response". In: *Science* 345.6193, pp. 148–149. DOI: 10.1126/science.345.6193.148-b. eprint: http://www.sciencemag.org/content/345/6193/148.2.full.pdf (cit. on p. 31).
- Miller, I et al. (2014). *John E. Freund's Mathematical Statistics with Applications*. 8th ed. Pearson Education Limited (cit. on pp. 8, 9).
- Simpson, EH (1951). "The Interpretation of Interaction in Contingency Tables". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 13.2, pp. 238–241. JSTOR: 2984065 (cit. on p. 26).

References v

Waa, J van der et al. (2018). “Contrastive Explanations with Local Foil Trees”. In: *2018 Workshop on Human Interpretability in Machine Learning (WHI)*. (July 14, 2018). Stockholm, Sweden. arXiv: `http://arxiv.org/abs/1806.07470v1 [stat.ML]` (cit. on p. 14).