PhD Notebook

Patrick Li

2022-01-14

Contents

Chapter 1

Welcome

I am Patrick Li.

Chapter 2

Introduction

This note consists of:

1. records of weekly meetings
2. literature review
3. TO-DO list
4. milestones
5. links to resources

Chapter 3

Literature - Visual Inference

3.1 Graphical Inference for Infovis

BibTex:

```
@article{wickham2010graphical,
  title={Graphical inference for infovis},
  author={Wickham, Hadley and Cook, Dianne and Hofmann, Heike and Buja, Andreas},
  journal={IEEE Transactions on Visualization and Computer Graphics},
  volume={16},
  number={6},
  pages={973--979},
  year={2010},
  publisher={IEEE}
}
```

3.1.1 Keywords

Statistics, visual testing, permutation tests, null hypotheses, data plots.

3.1.2 Introduction

Infovis focuses on uncovering new relationships by **tools of curiosity**, but most statistical methods focuses on examining relationships by **tools of skepticism**. Neither extreme is good. Hence, graphical inference try to fill the gap between them. It claims that this kind of inference can provide a tool for skepticism that can be applied in a curiosity-driven context.

3.1.3 What is inference and why do we need it?

Inference is about drawing conclusions about the population from the sample. There are two components of statistical inference, testing and estimation. For graphical inference, the focus is to test whether what we see in a plot of the sample is an accurate reflection of the entire population or not. The test statistic in visual inference is a plot of the data. A **null dataset** is a sample from the null distribution, and a **null plot** is a plot of a null dataset. The benefit of visual inference is that it can be used in complex data analysis settings that do not have corresponding numerical tests.

3.1.4 Protocols of graphical inference

3.1.4.1 Rorschach

Rorschach protocol is used to calibrate our vision to the natural variability in plots in which the data is generated from scenarios consistent with the null hypothesis.

3.1.4.2 Line-up

The line up is consisted of $n-1$ decoys and 1 plot of the true data. If we set $n = 19$, then under the null hypothesis, there is only 5% chance to pick the plot of the true data. If we recruit K observers, then under the null hypothesis, the p-value is $P(B(K, 0.05) \geq k)$.

3.1.5 Examples

To use the line-up protocol, we need to: 1. Identify the question the plot is trying to answer 2. Characterize the null-hypothesis 3. Figure out how to generate null datasets

There are two techniques that can be applied in many caeses:

1. Resampling. Permutation and bootstrapping.
2. Simulated data from a assumed model.

3.1.5.1 Tag clouds

A tag cloud can be used to visualize frequency of words in a document. Words are arranged in various ways, often alphabetically, with size proportional to their frequency. The null hypothesis for a comparison tag cloud is that the two documents are equivalent, the frequency of words is the same in each document. Null data can be generated by randomly permute the one of the column.

3.1.5.2 Scatterplot

A scatterplot displays the relationship between x and y. A strong null hypothesis is that there is no relationship between x and y variables.

3.1.6 Power

The power of a statistical test is the probability of correctly convicting a guilty data set. The capacity to detect specific structure in plots can depend on the perceptual properties.

3.1.7 Use

An R package: nullabor

3.2 Statistical inference for exploratory data analysis and model diagnostics

Chapter 4

Literature - Computer vision

Chapter 5

Milestones