EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät
Seminar für Sprachwissenschaft

Data Science for Linguists

Session 13: Statistical Inference

Johannes Dellert

2 February, 2024

Table of Contents

Hypothesis Testing

Resampling Methods

Statistical Modeling and Inference

Testable Hypotheses

- the **science** part of “data science” typically involves forming and testing **hypotheses** about
  - ▷ patterns in our data that are stable against random noise
  - ▷ properties of the processes we assume to have generated our data
- examples of hypotheses about data from the domain of linguistics:
  - ▷ *speakers are more likely to extrapose relative clauses with periphrastic tenses*
  - ▷ *the basic vocabulary of German and Hindi is more similar than either is to that of Georgian*
  - ▷ *the collocations in the works of author A are noticeably more similar to those in works of author B than to those in works by other authors*
- examples of hypotheses about the underlying processes:
  - ▷ *auxiliaries trigger extraposition by contributing to an overall sense of “heaviness”*
  - ▷ *languages whose basic vocabulary show as many similarities as German and Hindi cannot have arisen independently or merely through language contact*
  - ▷ *it is more likely that author A was influenced by author B, than the other way around*
- hypotheses about the data are typically investigated in terms of significance testing
- hypotheses about the underlying processes are typically investigated through statistical modeling and model comparison (which might in turn involve significance testing)

Classical Statistical Tests

- in order to be able to test a hypothesis about our data, we need to be able to translate it into a statement about one or several **statistics** (quantitative summaries of the data)
- if we assume that a statistic is sampled from a known distribution, we can use its observed value to draw conclusions about the plausibility of these assumptions
- more specifically, if we have a well-understood distibution of a statistic given the truth of a competing **null hypothesis**, we can measure how far the statistic differs from the values we would expect under the null hypothesis
- examples of potentially useful statistics and relevant null hypotheses in our example cases:
  - ▷ difference in percentage of extraposed clauses with and without auxiliaries; relative clauses in periphrastic tenses are extraposed just as often as relative clauses without auxiliaries
  - ▷ number of translations in a 100-concept list which exceed a certain phonetic similarity value; the words in German, Hindi, and Georgian are generated independently
  - ▷ Jaccard coefficients between the results of applying a collocation extraction algorithm on the works of both authors; the two authors are randomly selected (i.e. their works are as similar as between an average pair of authors)

Significance Tests

- in a classical significance test, we **reject** the null hypothesis in case the probability of seeing a value which is at least as extreme as the one we actually observed (the **p-value**) is lower than a certain pre-defined threshold (the **significance level**)
- the significance level at which we consider the null hypothesis as rejected can be interpreted as our willingness to make a type I error (where we reject the null hypothesis even if is true); conventional choices include 1%, 2%, and 5%
- it is generally good practice to report the p-values, and let the reader decide whether they consider them sufficient evidence for rejecting the null hypothesis
- not all null hypotheses and assumptions about distributions are equally good, they can differ in their statistical **power** (the probability of not making a type II error, i.e. avoiding the situation where we fail to reject the null hypothesis although it is false)
- very important caveat in classical significance testing: we need to always ensure that the assumptions are compatible with the actual distributions of the data (for instance, many classical tests assume that the relevant variables are normally distributed!)

Multiple Tests

- deciding on a given significance threshold means that even if there are no patterns at all in the data, in a certain percentage of tests we will erroneously reject the null hypothesis
- this implies that if we test different things long enough (**p-hacking**), some tests is bound to come back as significant, even on completely random data!
- principles of good science, all with the goal of reducing the risk of p-hacking:
  - ▷ determine a single hypothesis you want to test before even looking at the data (or only after looking a small subset of the data, think "development set" in machine learning)
  - ▷ clean the data without your hypothesis in mind (maybe even let a different person do it)
  - ▷ if you need to test several hypotheses on the same data, correct for multiple testing by lowering the significance threshold (e.g. **Bonferroni correction**, dividing through the number of tests, though this does not correct the false discovery rate)
- note that this implies that exploring several hypotheses on the same dataset, as when reusing a dataset for several studies, is always very problematic (although it cannot be avoided in practice - for instance, there is a limited number of documented languages!)

Table of Contents

Hypothesis Testing

Resampling Methods

Statistical Modeling and Inference

Resampling Methods

- a **resampling** method involves repeatedly drawing samples from a dataset, and repeating some computation on many such datasets, in order to get an impression of the variability of results against variations of the input data
- the two most commonly used resampling methods are **cross-validation** (known as the standard technique for avoidance of overfitting in machine learning) and the bootstrap
- the **bootstrap** is most commonly used to provide a measure of accuracy of parameter estimates (quantifying the uncertainty arising from the structure and the size of the sample)
  - ▷ basic idea: emulate obtaining new samples by sampling with replacement from the original sample as many observations as were contained in the original sample
  - ▷ with increased sample size, the resamples will be more similar to the original dataset, whereas for smaller sample sizes, resamples will vary a lot (and contain many duplicates)
  - ▷ widely applicable and extremely powerful statistical tool, as it allows to derive a measure of variability for the outputs of any complex algorithm (example: phylogenetic tree inference)

Table of Contents

Hypothesis Testing

Resampling Methods

Statistical Modeling and Inference

Statistical Modeling and Inference

- a statistical **model** specifies an assumed mathematical relationship between one or more random variables and one or several non-random variables (parameters)
- statistical **inference** is the estimation of model parameters which provide the best fit to data (a definition which actually includes many approaches to machine learning)
- especially in a Bayesian setting (see next slide), we do not derive **point estimates** (as we generally do when training parameters in machine learning), but **compatibility intervals** for parameter values, which can provide a less reductionist approach to investigating hypotheses
- in order to investigate a hypothesis about a data-generating process, we build models which differ in the structural property that we want to investigate, and then perform **model selection** by measuring which model can fit the data best
- model selection is typically based on some **information criterion**, a measure which balances out goodness of fit against model complexity (again to avoid overfitting)
  - ▷ the **Akaike Information Criterion (AIC)** is defined in terms of the number k of model parameters and the maximum likelihood $\hat{L}$ of the data given the fitted model: $2k - 2\ln(\hat{L})$
  - ▷ the **Bayes factor** is the quotient of the marginal likelihoods $p(D|M_1)$ and $p(D|M_2)$ for the two models M_1 and M_2 (each integrated over prior probabilities of the parameters)

Bayesian Statistics

- basic idea: define prior beliefs for the parameters, use Bayes' theorem to compute belief updates based on the data in order to derive **posterior distributions** for the parameters:

$$p(M|D) = \frac{p(D|M) \cdot p(M)}{p(D)}$$

- Bayes' theorem involves the explicit statement of the **likelihood** $p(D|M)$, the probability of the data given the parameter values, which can involve very complex computations that are often best understood in terms of a **data-generating process**
- in practice, the posterior distributions will almost never be directly computable as a function, but will instead be approximated by a collection of samples from the posterior distribution (which can be summarised in various ways to draw conclusions)
- while Bayesian inference provides more systematic approaches to many of the mentioned problems of classical statistics (especially uncertainty), implementation can be very tricky, and running Bayesian inference on larger models requires substantial computing resources

Preliminary Course Plan

1	**27/10**	**IPython and Jupyter**
2	**03/11**	**Introduction to NumPy**
3	**10/11**	**Pandas and Data Frames**
4	**17/11**	**Data Cleaning and Preparation**
5	**24/11**	**Linguistic Preprocessing**
6	**01/12**	**Data Wrangling**
7	**08/12**	**Data Aggregation and Grouping**
8	**15/12**	**Visualisation with Seaborn**
9	**22/12**	**Modeling and Prediction**
10	**12/01**	**Classification**
11	**19/01**	**Clustering**
12	**26/01**	**Pattern Extraction and Density Estimation**
13	**02/02**	**Statistical Inference**
14	09/02	Data Science Projects

Questions

Questions?

Comments?

Suggestions?