L1 -- Descriptive statistics

Key idea? Most important slide?

Key idea:
Many numerical and visual ways to describe the statistic

The lecture introduces variables and visualization methods to study and interpret data.

Most important slides: Different statistical studies slides 13&14

Observational
Controlled
Simulations

How does the lecture connect elsewhere? What mathematics?

Methods used in probability courses

These concepts were the base for pretty much the whole course

Statistics is used in many different fields from economics to healthcare and medicine. So it applys to many courses and majors.

Where would you use the methods?

the concepts are used pretty much all the time when you do statistical analysis

In prediction and time series analysis when analysing the error terms

In visualizing the data. Can be used as a demonstration in a scientific report.

What was surprising/familiar, easy/difficult? Further questions?

Familiar topics from previous courses were e.g. median, mode, variance and standard deviation.

New/newer topics: kurtosis and skewness

What is the intuition behind statistical moments?

L2 -- Conf. interv. and hypo testing

Key idea? Most important slide?

S21 Basics of hypothesis testing

Most important

S18 p-value

Very important

S20 T1/T2 errors

Important to keep in mind

How does the lecture connect elsewhere? What mathematics?

Statistical Analysis

Descriptive stastistics

Probability calculation

Other lectures

Hypothesis testing is used in most lectures

p-value used everywhere

Where would you use the methods?

Comparing point estimates between groups

science, research, analyzing results

Machine Learning

Election polling

What was surprising/familiar, easy/difficult? Further questions?

t-tests easy to understand

Confidence intervals intuitive

Difficult to control T2 error

Bootstrap is abstract

L3 -- Nonparametric testing

Key idea? Most important slide?

How and when to use sign and rank tests for models whose distribution isn't known (non-parametric)

Signed rank test
i.i.d. sample from continuous, symmetric distribution
-> Calculate differences, rank them in ascending order and sum the values of "one side"

Sign test
i.i.d. sample from continuous distribution
-> calculate number of observations larger than the hypothesis

Best slide - 16

How does the lecture connect elsewhere? What mathematics?

Inside the course tightly linked to t-tests -> different assumptions and ways to calculate the test statistic

Basic principles of probability mathematics were of course needed e.g. in understanding how the test statistic affects p-value etc.

None of our group have previously encountered non-parametric test

Where would you use the methods?

For example in analyzing research findings when nothing really is known about the underlying distribution

In data science, non-parametric can be used to analyze the location of the data

Problems
Not the most powerful methods out there

Machine learning: comparing model outputs

What was surprising/familiar, easy/difficult? Further questions?

Familiar
The test structure overall is familiar from the previous lecture

Surprising
Rank tests can be applied to non-numeric data if it can be ordered.

L4 -- Inference for binary data

Key idea? Most important slide?

Hypothesis testing using one-sample and two-sample tests for Bernoulli Distribution samples

How to apply hypothesis testing to Bernoulli distributions

The relationship between repeated Bernoulli distribution and normal distribution.

slide 12, which has the test statistic for two-sample proportion test

slide13, showcasing contingency table

How does the lecture connect elsewhere? What mathematics?

Mathematics: Classical Probability

Connected to: Lecture 2 (Hypothesis testing, confidence interval)

Confidence intervals relies on normal approximation; estimates of mean and sd. are substituted

Where would you use the methods?

We would be able to use these methods when dealing with binary distributions and want to test their success probability

It is possible to artificially transform data into binary form.

Machine learning sometimes has binary data which can be analyzed with binary methods. (logistic regression...)

Frequency table can be used to illustrate the occurance of binary data

What was surprising/familiar, easy/difficult? Further questions?

Familiar:
- Bernoulli distribution
- Contigency table looks alot like Confusion table in machine learning

Easy: The computations for the sample proportion tests are not difficult to implement in code

Easy: The results for the proportion tests are easy to understand

Further question: Is there formula for the true confidence interval?

L5 -- Distribution tests

Key idea? Most important slide?

Slide 3 is the most important; it outlines why distribution tests are necessary

The key idea is exploring the distribution of data (and its [non]normality)

Normality distribution tests, multiple uses of Chi-squared tests

How does the lecture connect elsewhere? What mathematics?

Relevant to essentially all other lectures, any lecture that assumes some distribution of a sample/population

Mostly computation, but also definition of multinomial distributions

Where would you use the methods?

You could study whether noise from a system is normally distributed.

Economics example: if we use regression to predict future values using past values of a time-series that comes from a distribution that changes over time (e.g. GDP) we will get non-normally distributed t-values.

Tests to determine non-normality of t-values and thus determine appropriate critical values

Tests to determine changes in distributions over time

What was surprising/familiar, easy/difficult? Further questions?

The binomial distribution (familiar from probability) appears in this lecture.

The manual computation feels kind of difficult and clunky, especially with large sets of data

The Q-Q plot is a tangible way to study the distribution of the data

L6 -- Correlation and independence

Key idea? Most important slide?

- linear vs monotonic dependence
- Correlation calculations
- SLIDE: 3, 7, 12

How does the lecture connect elsewhere? What mathematics?

- Pearson correlation, Spearmans rank
- Bootstraping, confidence intervals, sample-tests
- Hypothesis testing
- Linear regression
- Bivariate normal ditribution

Where would you use the methods?

- Finding correlation
- How elements depend on each other?
- Causality more difficult to find
- Two samples
- Paired data

What was surprising/familiar, easy/difficult? Further questions?

- Linear dependence
- Monotonic dependence
- What is a large enough value to reject H0?

L8 -- Linear regression II

Key idea? Most important slide?

key idea: multiple linear regression (slide 3)

we have a extension on simple linear regression

Slides 26-28 Extensions

Important concepts:
- multicollinearity
- VIF
- R^2
- heteroscedasticity
- diagnostics

How does the lecture connect elsewhere? What mathematics?

linear regression I

Kernel regression

MSE

Not particularly mathematical: mainly test statistics

correlations (multicollinearity)

Where would you use the methods?

In general: When we've identified several independent variables on which the response variable has a linear dependence

attempting to fit a multiple linear regression model to data

Machine learning

Finding outliers with residual plot

What was surprising/familiar, easy/difficult? Further questions?

Comparatively simple to understand

surprising: that one can infer whether assumptions hold by looking at the patterns the residuals exhibit

Certain familiarity due to usage of linear regression

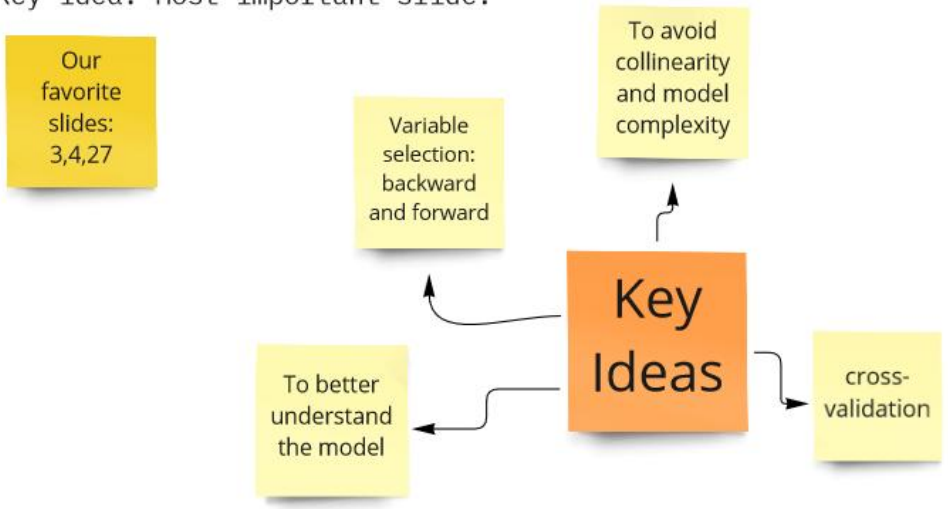
Not too difficult but really important topic

concept of multicollinearity fairly familiar and straightforward

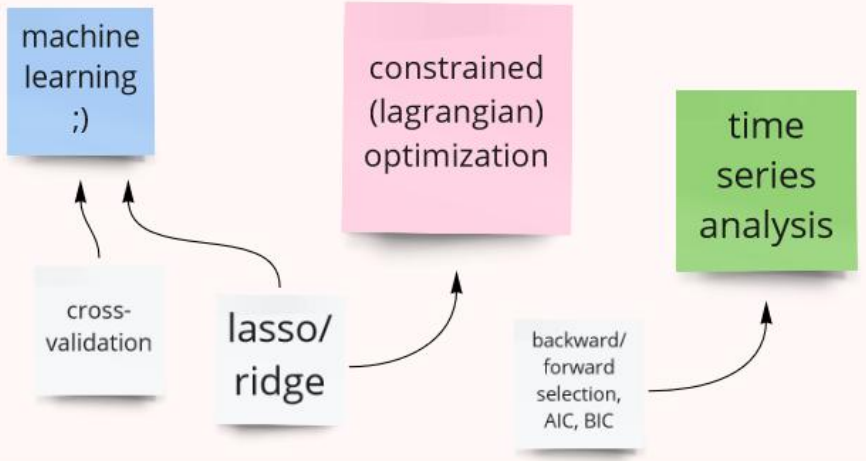
Linear model, SME etc. were already familiar

L9 -- Linear regression III

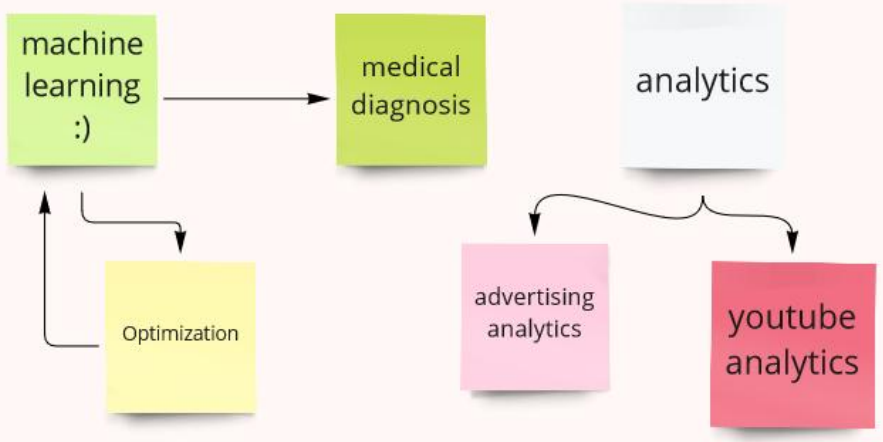
Key idea? Most important slide?



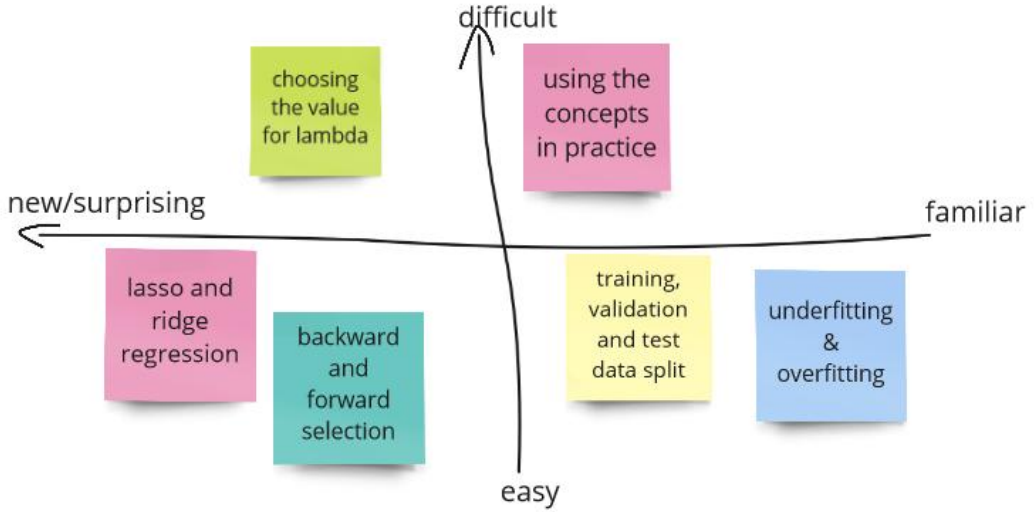
How does the lecture connect elsewhere? What mathematics?



Where would you use the methods?



What was surprising/familiar, easy/difficult? Further questions?



L10 -- Analysis of variance

Key idea? Most important slide?

There are 2 ways to test the mean/ median of multiple population (>2 groups):
- **ANOVA:** requires all to be normal distribution
- **Kruskal Wallis:** requires all to have a same histogram shape

If all groups dont have equal mean/ median we have to use t-test to test which group is not equal, but there is risk of Type I (need to reduce this risk)

Most important slides: 4, 5, 17
Good example: 10-12 & 16

How does the lecture connect elsewhere? What mathematics?

ANOVA is closely related to multiple linear regression. It is equivalent to regressing the x-variable on the indicator variables of the groups

The type I error probability increase during the t-test (pair test) to check which group's mean is different

ANOVA often being used in manufacturing, when making decision to choose which supplier to move on, based on data (compare test samples among different suppliers)

Where would you use the methods?

ANOVA
Purpose: to test whether the means of many groups are equal (number of group > 2).
Requirement:
- All groups follow normal distribution
- Variance each group are equal

Kruskal Wallis
Purpose: to test whether the median of many groups are equal (number of group > 2).
Requirement:
- All groups don't have to be normal distribution
- Histogram must have a same shape

Boniferroni correction
Purpose: to avoid increasing probability of type I error when checking which group's mean is different
Method: by adjusting significant value for accepting H0
(New significant = alpha/ C)

Barlett's Test
Purpose: To verifythe requirement of ANOVA: variances of all groups are equal

Implementations: ANOVA tests the equality of the expected values of the groups. Example: quantified behaviour of the several groups of people

What was surprising/familiar, easy/difficult? Further questions?

Familiar:
-Hypothesis testing
-Multiple linear regression
- t-test

The number of observation in each group of Kruskal Wallis test is not required to be the same

ANOVA separates variance into 2 group: variance between group mean and variance in each group

L11 -- Kernel regression

Key idea? Most important slide?

Slide 5: General idea

Flexible, if you want to fit a model close to data in weird shape

draw a curve close to the points

Kernel regression

Local model, surrounding points define the model

Finding a non-linear relation between random variables

(Slide 6: outlining the mathematical definition)

How does the lecture connect elsewhere? What mathematics?

First contact with kernels

Alternative for linear (h-o polynomial) regression

Where would you use the methods?

When out of ideas or previous methods fail

When the phenomenon is not explainable by a polynomial → Motorcycle helmet crash data

In studies no datasets, where kernel would be the obvious choice

Google: Used in predicton of home sale value based on home area in square feet

What was surprising/familiar, easy/difficult? Further questions?

The general idea is easy to understand: draw the lines using the nearby points

How does the kernel selection affect the regression

↓

Gaussian / normal or triangle