L1 -- Descriptive statistics

Key idea? Most important slide?

Use basic statistical descriptors and visualizations to summarize data

Slide 12 (or 9)

Most helpful visualisations:
- histograms
- scatterplots
- bar charts
- boxplot

Helpful numeric descriptive measures: Mean, median, variance, skewness, correlation

How does the lecture connect elsewhere? What mathematics?

Tools to help decide what kind of inference analysis and methods should one use

Descriptive analysis is the starting point of building more complex models

Statistical descriptors

Where would you use the methods?

Checking key measures before analyzing data in more depth

preparing data for analysis (cleaning, removing missing values, feature engeneering)

Business Intelligence: providing actionable insights based on visual data

When deciding how to plot the data

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

Kurtosis relatively new concept

Concept of giving numerical and comparable values to the properties of the data

Basic measures of location were already quite familiar

Haven't seen box plot many times before

L2 -- Conf. interv. and hypo testing

Key idea? Most important slide?

lecture gives many basic ideas on hypothesis testing

most important slides:
- p-value (s.18)
-boot strapping (s.5-7)

topics:

confidence intervals: adding valuable information about accuracy to point estimates

test statistics, p-values and significance levels

Hypothesis tests: computing a test statistic, which is dependent on the dataset and the null hypothesis, and which has an expected value under the null hypothesis. Major deviation from the expected value proves against the null hypothesis

boot strapping: method for estimating confidence intervals

type I and II errors

types of tests:
- one-sample t-test
- two-sample t-test
- paired t-test (one-sample with difference)
- variance (comparison) test

Where would you use the methods?

Hence the concepts are pretty fundamental, the ideas can be applied to many statistical problems

watch out for assumptions e.g. of normality of sample distribution "If the sample size is large, then the t-test is not very sensitive to moderate deviations from normality."

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

concept of hypothesis testing was familiar (H0, H1)

new: maths behind confidence intervals, different kinds of tests

concepts of boot strapping and p-value hard to grasp, but important

L3 -- Nonparametric testing

atics?

Key idea? Most important slide?

Non-parametric tests: sign tests vs rank tests

Most important slide is 2 because it explains the differences between non-parametric and parametric tests

Second most important slide is 16. It describes use cases for signed and signed ranked tests.

How does the lecture connect elsewhere? What mathematics?

the idea of paramtric/non-parametric came up in later lectures as well

sign tests and rank tests are counterparts of the t-test (previous lecture)

how it relates to other fields of mathematics:
- symmetry (rank tests)
- z statistic (for large sample)
-t statistic (parametric counterpart)

Where would you use the methods?

used to make inferences about population parameters based on statistics
-> works when the population distribution is non-normal

signed rank test -> if symmetry can be assumed; can also be used for categorical values

sign test -> if no assumptions about the distribution can be made

rank tests can be used even when a variable cannot be measured numerically, for an example when ranking bands

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

Test-statistics may be a bit difficult to be interpreted as "large" or "small"

For large sample, the standarised test statistic, can be used for hypothesis testing

easier to understand this lecture after the previous lecture about t-tests
-> topic of testing was more familiar

sign tests can be also used with discrete variables

L4 -- Inference for binary data

Key idea? Most important slide?

Binary observation and Bernoulli distribution

Approximate Confidence interval

One-sample proportion test

Two-sample proportion test

One-sample proportion test used to test whether the success probability of a Bernoulli distribution equals some prespecied value

two-sample proportion test used to test whether two Bernoulli Distribution has the same success probability

mathematics?

Bernoulli and Binomial probability distributions

From the central limit theorem the binomial distribution can be approximated by the normal distribution for large n

Where would you use the methods?

Calculate a confidence interval for the probability of a binary event

one sample proportion

Public health: For example, estimating the proportion of a population that has been vaccinated against a disease to assess the effectiveness of a vaccination campaign.

Medical research: For example, assessing the effectiveness of a new medication in curing a specific disease by comparing the proportion of patients who respond positively to the medication against a known population proportion

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

Frequency tables

id	X	Y
1	0	1
2	0	0
3	0	1
4	1	1

	Y=0	Y=1
X=0	1	2
X=1	0	1

L5 -- Distribution tests

Key idea? Most important slide?

3.

Lecture is about statistical tests to gain information about the distributions of variables.

normality tests, qq-plots, χ2-tests

- In statistics, assumptions on the underlying distribution are done all the time.
- Many statistical methods become ineffective or even give false results if their assumptions do not hold.
- This is why it is very important to test the distributional assumptions separately.

How does the lecture connect elsewhere? What mathematics?

Many of the statistical tests used assume a certain distribution

if variables are dependent, result of one helps to predict the other

Where would you use the methods?

Spotting anomalies in account behavior

When doing statistical tests

Checking survey data distribution before further analysis

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

Wide applicability of chi-squared tests

Interpreting Q-Q plots is surprisingly hard

Note the difference between then χ2-homogeneity test and the χ2 test of independence, applied to different sample sets.

L6 -- Correlation and independence

Key idea? Most important slide?

if two random variables are dependent, result of one does help to predict the other

correlation does not imply causation

Slide 11

linear dependence: if random variables x and y satisfy y = ax+b, they are said to be linearly dependent. Can be measured with Pearson correlation coefficient

monotonic dependence: if x and y satisfy y = g(x), g is monotonic, then x and y are monotonically dependent. Can be measured with Spearman correlation coefficient

How does the lecture connect elsewhere? What mathematics?

lecture is good base for topics that include datasets. Many test have assumptions if the variables are dependent or not

calculating correlation coefficients and covariance

computation of significance and hypothesis tests and confidence intervals

linear and monotonic functions

Where would you use the methods?

Always when we want to know if two variables are dependent

For example in computer science we want to know if the amount of bugs in code dependent of the amount of code

Are number of accidents dependent of speed limits in the area?

Are errors in data traffic dependent on each other?

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

correlation and covariance were familiar from previous courses

surprising: if correlation is 0, it does not imply independence

L7 -- Linear regression I

Key idea? Most important slide?

Simple linear regression, assumptions

- Consider n observations (pairs) $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ of (x, y). Assume, for simplicity, that the values x_i are non-random (otherwise we need an assumption of *exogeneity*).
- Assume that the values y_i depend linearly on the value x_i:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n,$$

where the regression coefficients β_0 and β_1 are unknown constants.

Main idea: Linear regression and its use cases. How does it work in practice?

Simple linear regression, assumptions, continued

- The expected value of the errors is $\mathrm{E}[\varepsilon_i] = 0$ for all $i = 1, \dots, n$.
- The errors have the same variance $\mathrm{Var}[\varepsilon_i] = \sigma^2$.
- The errors are uncorrelated i.e. $\rho(\varepsilon_i, \varepsilon_j) = 0, \quad i \neq j$.
- The errors are i.i.d. *(a stronger version of the previous two assumptions).*

Most important slides: 6, 8 and 9

Under the previous assumptions, the random variables y_i have the following properties:

- Expected value: $\mathrm{E}[y_i] = \beta_0 + \beta_1 x_i, \quad i = 1, \dots, n,$
- Variance: $\mathrm{Var}(y_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2$.
- Correlation: $\rho(y_i, y_j) = 0, \quad i \neq j$.
- If we chose to assume that the errors are i.i.d., then y_i are independent of each other.

How does the lecture connect elsewhere? What mathematics?

Applications: Physics (Springs, Hooke's law)

Economics and Finance (Modelling of risk)

Matrix multiplication

How does it compare to the other lectures?

Multiple lectures on Linear Regression

The lecture Linear regression I is the introduction to the concept on which lectures II and III expand on

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

Familiar: General formula for linear regression

The reasons for assumptions of Linear regression are not very well covered on the lecture

What is the reasoning behind the assumptions?

Residual mean square estimation

Under the regression assumptions, an unbiased estimate for the error variance $\mathrm{Var}(\varepsilon_i) = \sigma^2$ is given by

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

L8 -- Linear regression II

Key idea? Most important slide?

multiple linear regression, its parameters and possible problems → the most important slides are: 3, 11, 12

ex: $P=3$

explaning variables

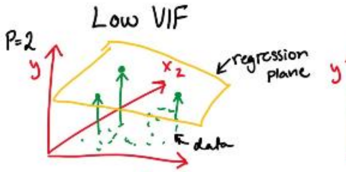
$$X = \begin{array}{c} \begin{matrix} & x_1 & x_2 & x_3 \end{matrix} \\ \begin{bmatrix} 1 & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \cdot & \cdot & \cdot \end{bmatrix} \end{array} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \\ \vdots \end{bmatrix}$$

extra column of 1s.

the reason for extra column of 1s, so that β_0 and x_1 match up and we got the β_0 right!

$$n \times (1+p) \cdot (1+p) \times 1 = n \times 1$$

compatible



now we have data a little from everywhere
→ easy to see how the plane is tilted

now when we have high correlatio with data
→ we cannot see how the plane is tilted!

How does the lecture connect elsewhere? What mathematics?

AI, Machine Learning, 3D Space. The goal is to build model between explanatory (independent) variables and response (dependent) variables

In terms of mathematics, multiple linear regression involves finding the line of best fit through the data points by minimizing the sum of squared differences between the observations and the fitted values. This is done using **matrix algebra** and **calculus to derive** a formula for finding the regression coefficients that minimize the sum of squared errors

Where would you use the methods?

When you have multiple explaining variables for one data point for example in datascience for predicting weather (the explaining variables would be temperature, humidity and precipitation)

VIF helps choose variables we should keep in the dataset for better results and models. It examines the relationships - multicollinearity

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

The new concept of multiple variables was easy to grasp, because it was very similar to the previous lecture, but with just added explaining variables and regression coefficients

VIF function was familiar and easy because I used it yesterday.

easy to understand, because you can always think of the concept first in three dimentions where it is simple geometry

L9 -- Linear regression III

Key idea? Most important slide?

Variable selection and shrinkage methods / "continuous" variable selection (Ridge, LASSO)

Methods to choose a model that fits the data well (big R^2) but does not have too many variables

There is also a trade-off between overfitting and underfitting a model: cross-validation and penalizing of complex models can be used to find the most suitable model

3, 15, 17, 20

Cross-validation: keep some of the

overfitting

Choose the model that minimizes avg.

"seen"

How does the lecture connect elsewhere? What mathematics?

Mathematics involved: Computation

The size of the estimated coefficients: Vectors --> Norms

Measuring the model's goodness of fit and Bayesian information criterion: Logarithms, matrix algebra

Machine

Model variable selection

Where would you use the methods?

Variable selection can be used when...

...wanting to interpret the model better (understand the phenomenon underlying the data)

...predictors are not related to the response add extra noise in prediction

... wanting to avoid collinearity

...you have cost issues, and it might be cheaper to observe only a subset of the variables

LASSO models are usually used to get a quick idea of which features are important for predicting the outcome variable. Can be used in finance e.g. in portfolio management.

Ridge regression is most suitable when a data set contains a higher number of predictor variables than the number of observations.

Forward and backward selection can be used to identify the most relevant variables

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

Somewhat familiar with LASSO

The content itself was not that hard to internalize

Familiarity from L7 and L8

L10 -- Analysis of variance

- ANOVA = statistical method for comparing the averages of 2+ groups and determining whether the observed differences between the groups are greater than would be excepted
- Kruskal-Walls = Compares the differences between groups in organized observations (non-parametric-ANOVA)

ANOVA, the basic idea

In analysis of variance, the total variance is divided into two parts. The first part measures the variation between the group means and the second part measures variation within the groups.

If the first part is much larger than the second part, there is evidence against the null hypothesis and we reject it. Otherwise, it is plausible that the group means are equal.

Hence, the test of the equality of the expected values is based on the comparison of between-groups variance and within-groups variance (giving the name of the method).

How does the lecture connect elsewhere? What mathematics?

-Connects to other lectures discussing about analysis of sample properties (e.g. median and expected value).
-The main mathematics here is calculation of sums instead of propabilistic methods.

Where would you use the methods?

We would use them anywhere we would require knowledge if there is some difference between various groups regarding some unknown variable

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

Found it relatively easy, since connects to previous topics, where population properties are compared.

L11 -- Kernel regression

Key idea? Most important slide?

Key Idea: Trying to fit something locally instead of globally, the general idea of how Kernel regression works

Most Important slide: 5th slide since it tells abbout the general idea of local regression

How does the lecture connect elsewhere? What mathematics?

This lecture discusses how to fit data locally instead of globally like in the other lectures of this course, Linear Regression is used in local linear regression

Mathematics of the lecture were computational

Where would you use the methods?

Great tool to see trends in the data, When using data where the distribution is unknown the kernel distribution is useful

Could be used for example in ML problems

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

It was surprising that Epanechnikov kernel was found to be generally the best choice
Method was unfamiliar but it felt natural that there would be a statistical method that would fit locally, In theory Kernel regression is simple but the implimentation is more difficult