

Statistical Inference for Data Science

Dr. Anja Mühlemann

30. August 2023





Questions from Day 1

Day 2

Parameter Estimation

Today's Topics

- Point Estiamates
- Least squares / Maximum likelihood
- Confidence intervals
- Regression
- p-values

So far ...

- Descriptive Statistics gives insight on a sample
- However, often we are **not only** interested in the sample itself
- We would like to draw conclusions about the **entire population** from which the sample was drawn

Example:

n volunteers received a vaccine. Now Novartis would like to predict the efficacy of the vaccine. More precisely, Novartis would like to predict the efficacy for the entire population and not only for the volunteers.

Inferential Statistics

Inferential Statistics

With a certain degree of certainty, one would like to draw conclusions from empirical data, even if the data are subject to error or incomplete.

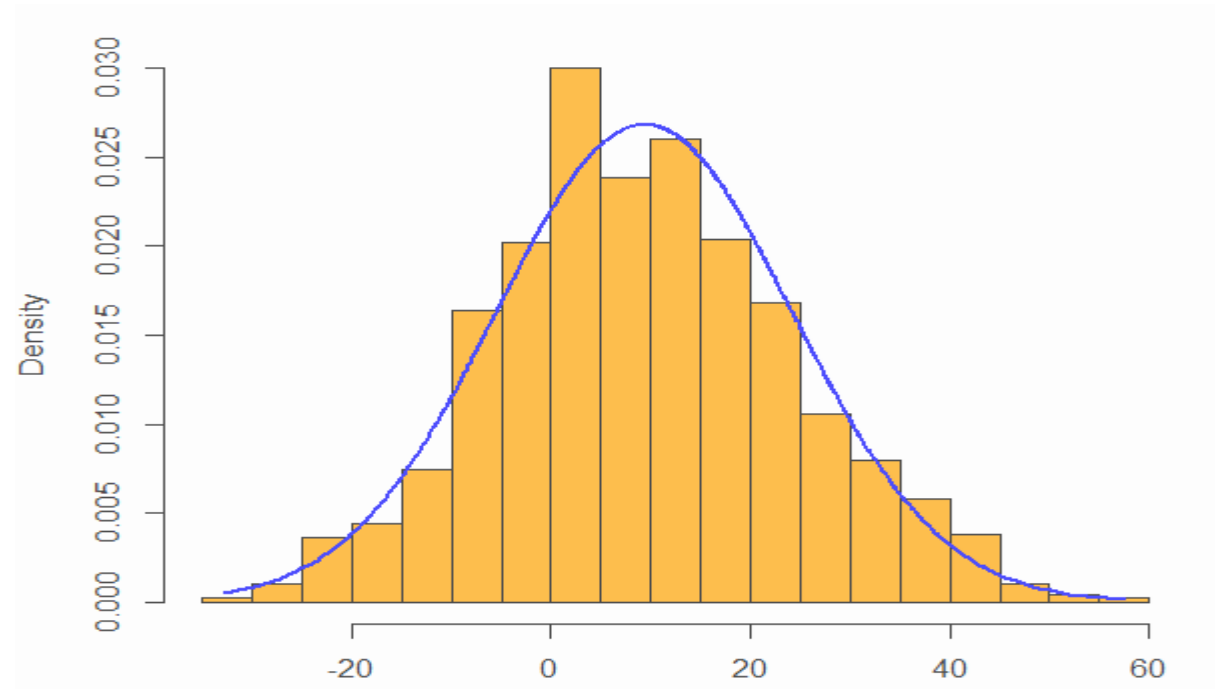
3 main techniques

- **Parameter estimates:** Calculation estimate for unknown parameter of underlying probability distribution
- **Confidence intervals:** Calculation of a region within which unknown parameter should lie with certain degree of certainty
- **Tests:** Tests are intended to prove that a certain effect, e.g. the effect of a vaccine, is indeed present.

Parameter Estimation

Situation

- We have data
- We have (chosen) a model describing the data
- The model has parameters
- We want to estimate the parameters from the data



Normal distribution

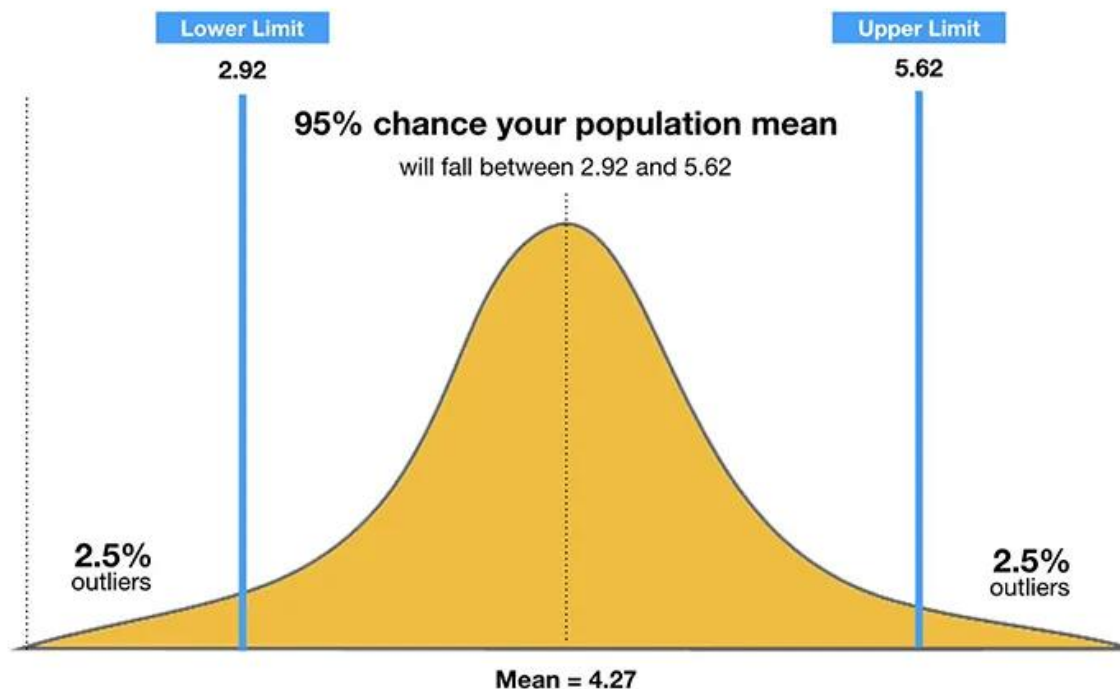
- The normal distribution is uniquely characterized by its mean μ and variance σ^2 (standard deviation²)
- We can estimate those parameters by the sample mean and the sample variance

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Estimators often denoted with 'hat' e.g. $\hat{\theta}$ is an estimate for θ .
- During this CAS you will fit many other model parameters from data.

Confidence Intervals

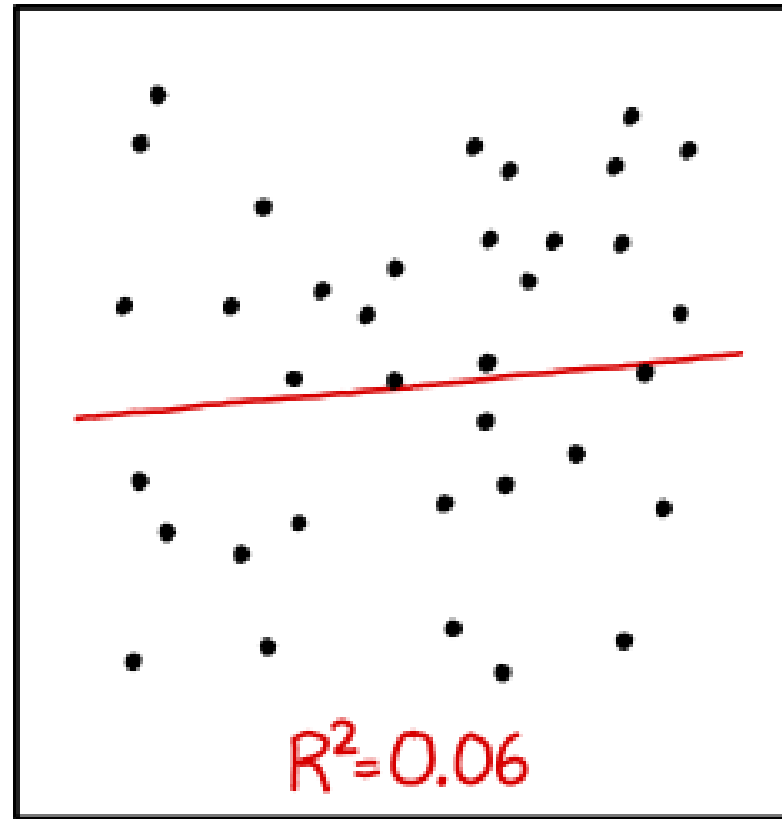
- is an interval that contains a certain parameter with a predetermined certainty (usually 95%)
- In contrast to point estimators, confidence intervals also reveal the uncertainty that arises due to the sample itself and the sample size.



$$\bar{x} \pm t \cdot \frac{s}{\sqrt{n}}$$

Regression

- Estimate the relationship between variables



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Different types

Linear

- Linear refers to the relationships between the x_i and y

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

- e.g. straight line $y = \beta_0 + \beta_1 x_1$
- Inter- and extrapolation allows prediction

Non-linear

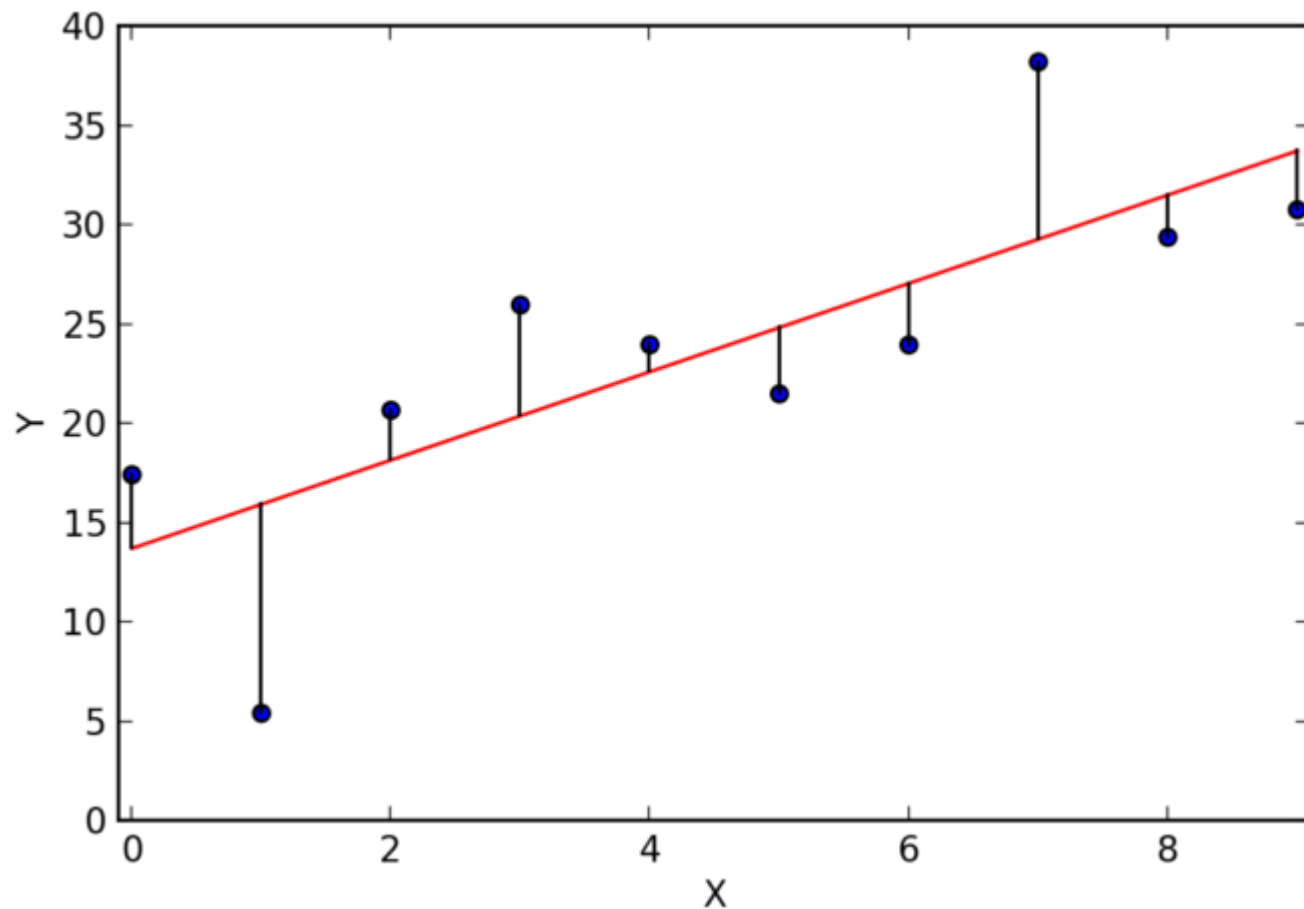
- When dependent variable (y) not linear in the parameters

Non-parametric

- Parametric models have parametric form
- If we have no clue about the exact relationship, we may use non-parametric estimation (e.g. isotone regression)
- These normally need more data, because also the model structure must be somehow estimated

Least Squares

- Minimize distance between data points and the regression line
- Under certain conditions same as Maximum Likelihood Estimator



Maximum Likelihood

- Estimate parameters of probability distribution, so that under the assumed statistical model the observed data is most probable.

- Family of parametric models (pdf)

$$\{f(x; \theta): \theta \in \Theta\}$$

- Find $\theta \in \Theta$ that maximizes the Likelihood function

$$L(x; \theta) = \prod_{i=1}^n f(x_i; \theta)$$

- For computational reasons, mostly the log-likelihood

$$\log(L(x; \theta)) = \sum_{i=1}^n \log(f(x_i; \theta))$$

Typical ML- Methods

- **Linear regression** (with logistic regression for classification)
- Decision trees (and random forest)
- Principal Component Analysis (dimension reduction)
- Nearest neighbor methods (k-means)
- **Neural Networks**
- In this CAS you will practice linear regression and neural networks (Module 3)

Typical ML- Methods

Up to here check the old downloaded one!!

- Most methods typically use either Least Squares or Maximum Likelihood to fit the optimal parameters.
- When the model is fitted, it can be used for hypothesis testing or classification.
- The simplest model is fitting a straight line to some data points. This model has 2 parameters.
- According to some sources, it is true that GPT-4 has 1.7 trillion parameters



Linear Regression

Leverage indicate how much the different values influence the model.

Leverage plot: Cook's distance. Gives an idea of how much the points influence the plot. Points should be inside the cook distance.

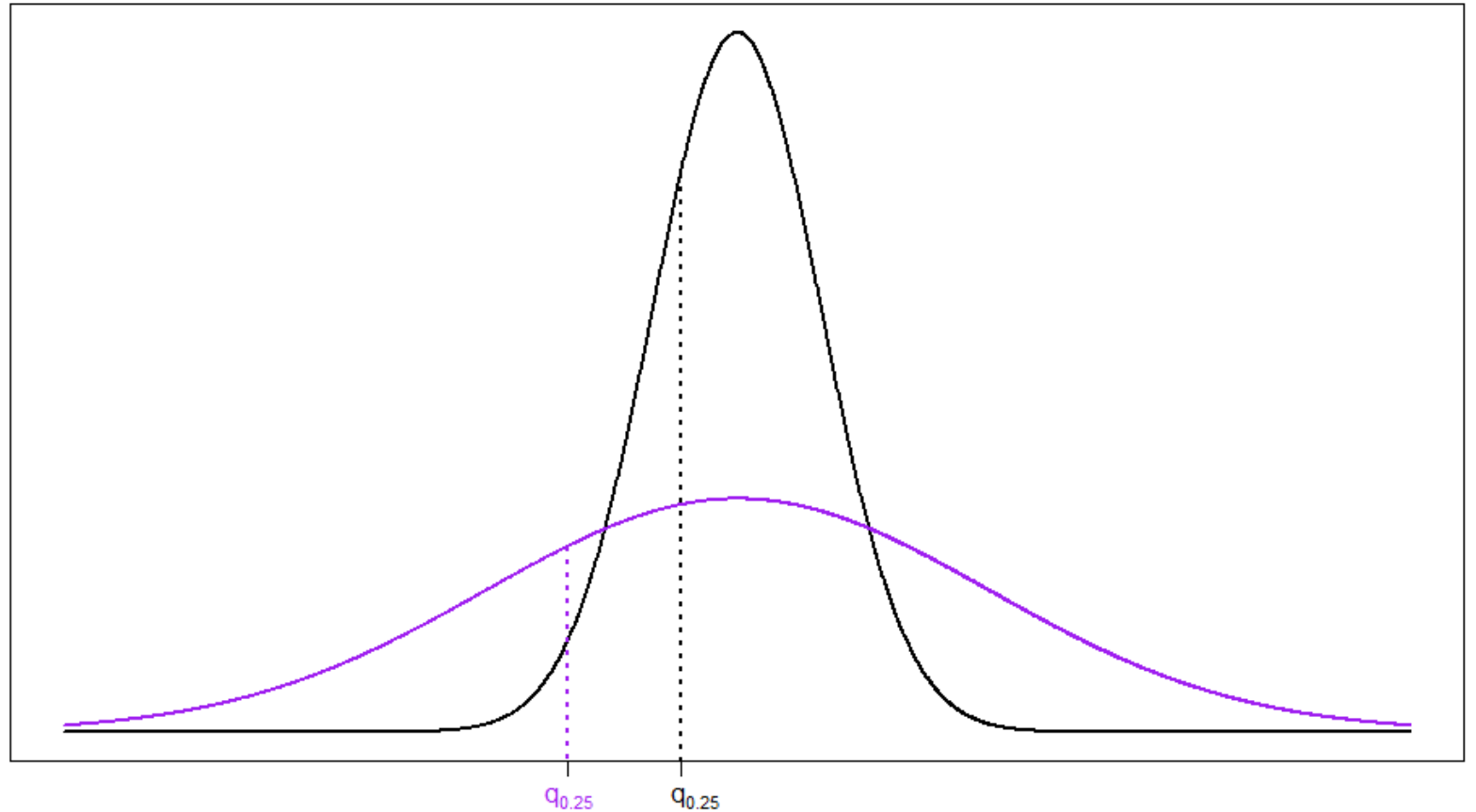
- (Least Squares) Linear regression is very widely used but we should not neglect the assumptions on which it relies
 - Normality This means basically that we can control our error.
 - Linearity
 - Homoscedasticity (errors have equal variances) This means that for small or big x we expect similar error (not very small for small values and very big for big x).

In the notebook there are plot to help you decide on all three assumptions

Normality assumption (Q-Q-Plot)

Diagnostic Tool: Q-Q plot
(Quantile - quantile plot)

We take the quantile of a normal distribution and we compare it with the quantile of our distribution.



If there is a linear correlation I expect that plotting the two quantiles for different points I obtain a line. This idea is general, we

NOTE: in the notebook we'll have a function for the Cook's curve, w

Red is always the standard normal distribution Purple is our distribution

If we have a normal distribution we obtain a line
as our curve is just a linear transformation
of the normal distribution

Normality assumption (Q-Q-Plot)

We see a typical belly

We see here a curved shape as the two
quantiles changes a lot moving in the two curves

What to do: we select a group of data of the same size of our samples,
we analyze it with a normal distribution, we do a Q-Q plot between our
distribution and the freshly obtained normal distribution

