CAS Applied Data Science – Module 2 – Day 2

Statistical Inference for Data Science

Dr. Anja Mühlemann

28. August 2024

UNIVERSITÄT
BERN

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.

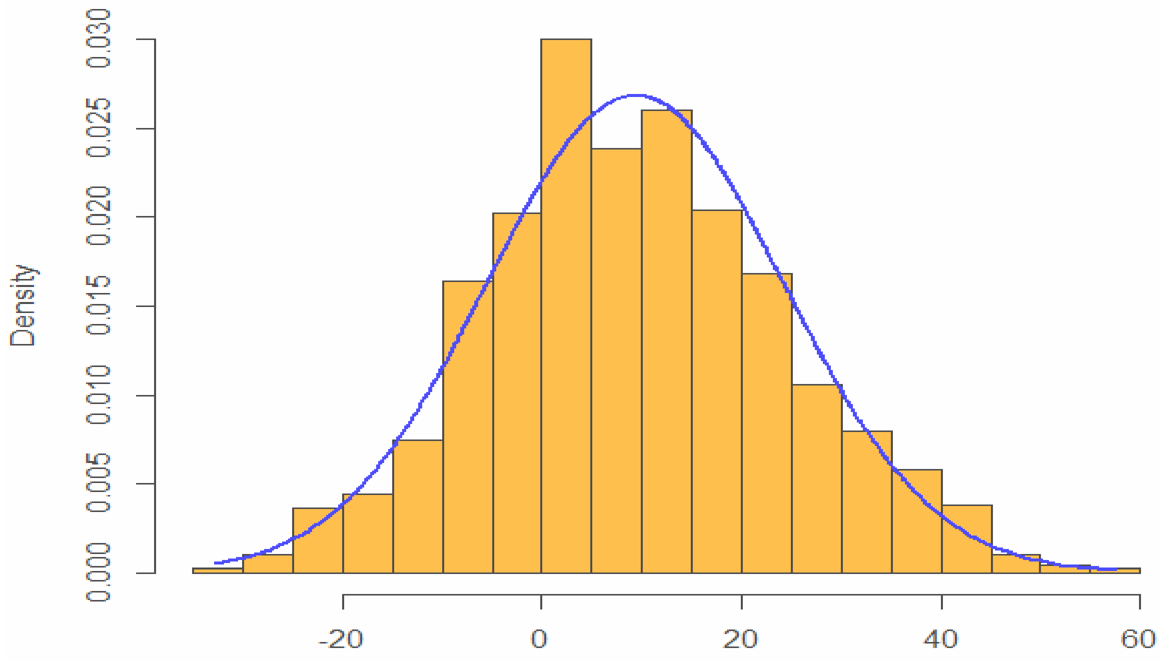
3 main techniques

- **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 $(\text{standard deviation}^2)$

- We can estimate those parameters by the sample mean and the sample variance

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

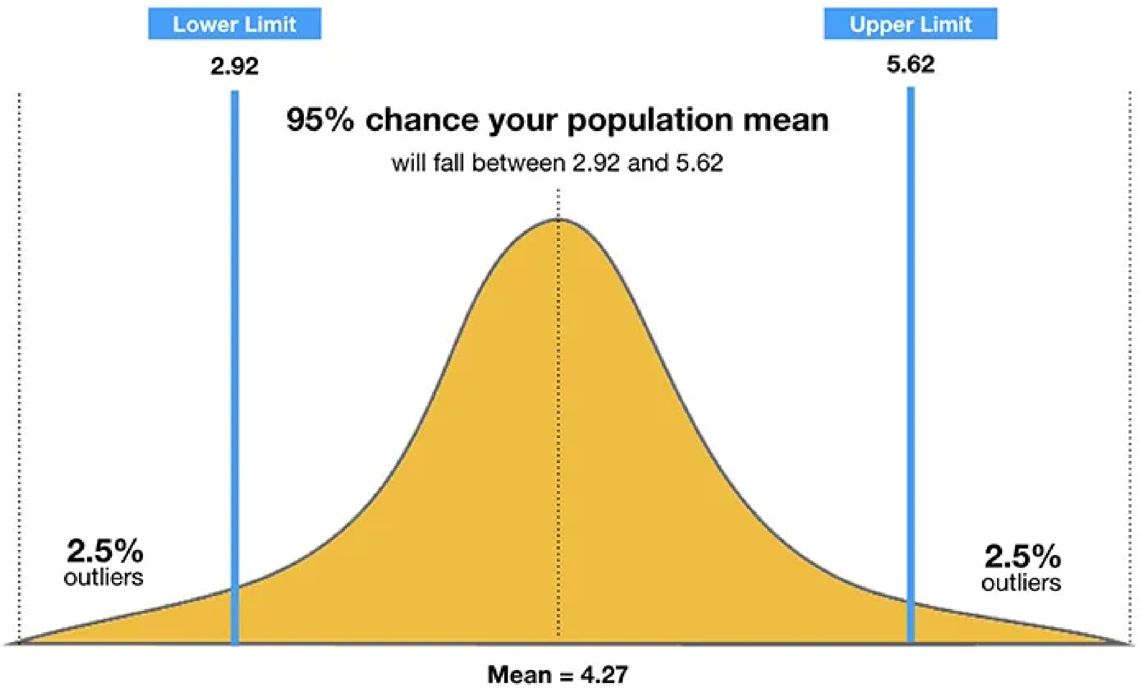
- Estimators often denoted with 'hat' e.g. $\hat{\theta}$ is an estimate for θ.

- 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%)

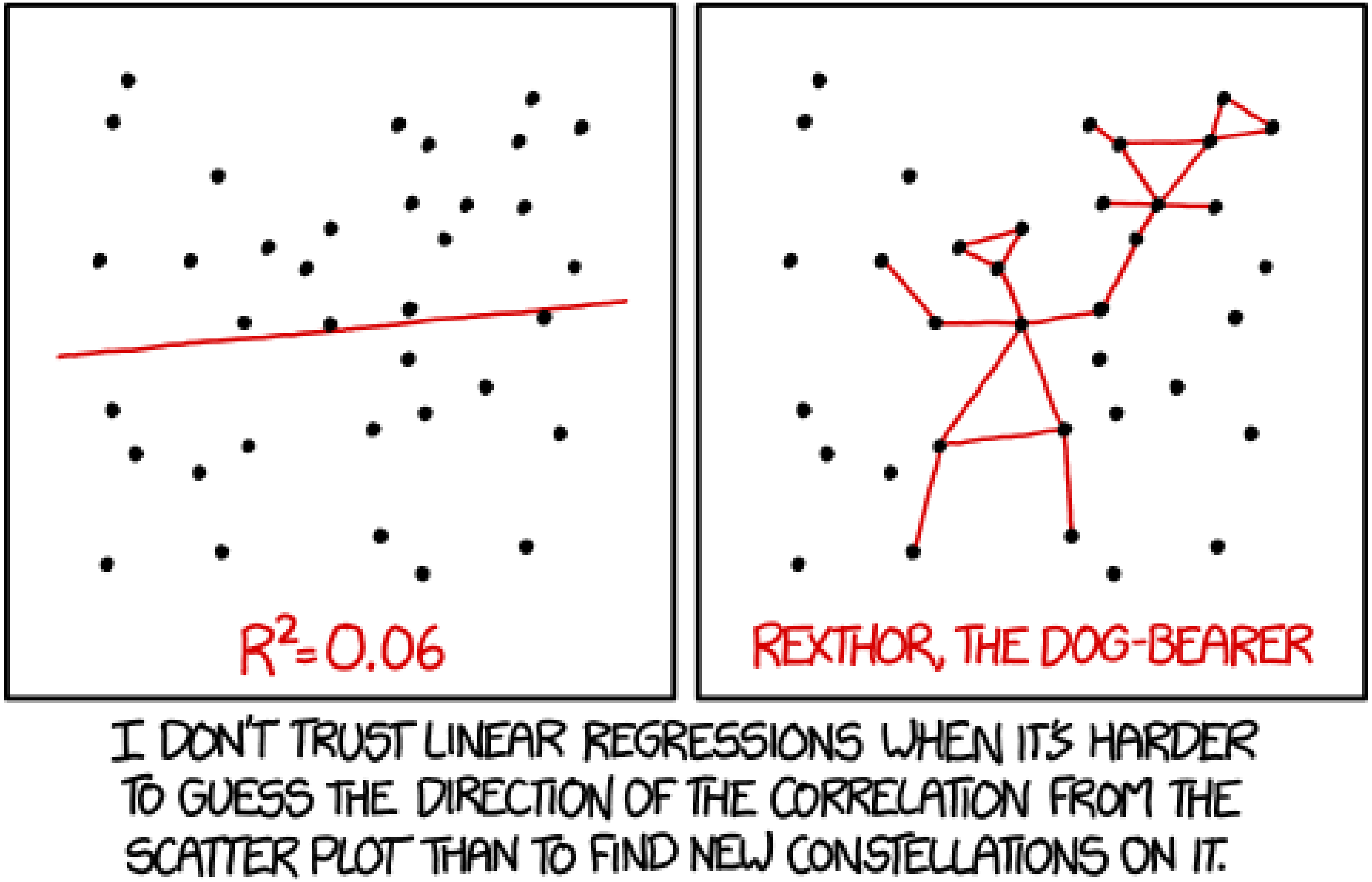
- In contrast to point estimators, confidence intervals also reveal the uncertainty that arises due to the sample itself and the sample size.



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

Regression

- Estimate the relationship between variables



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

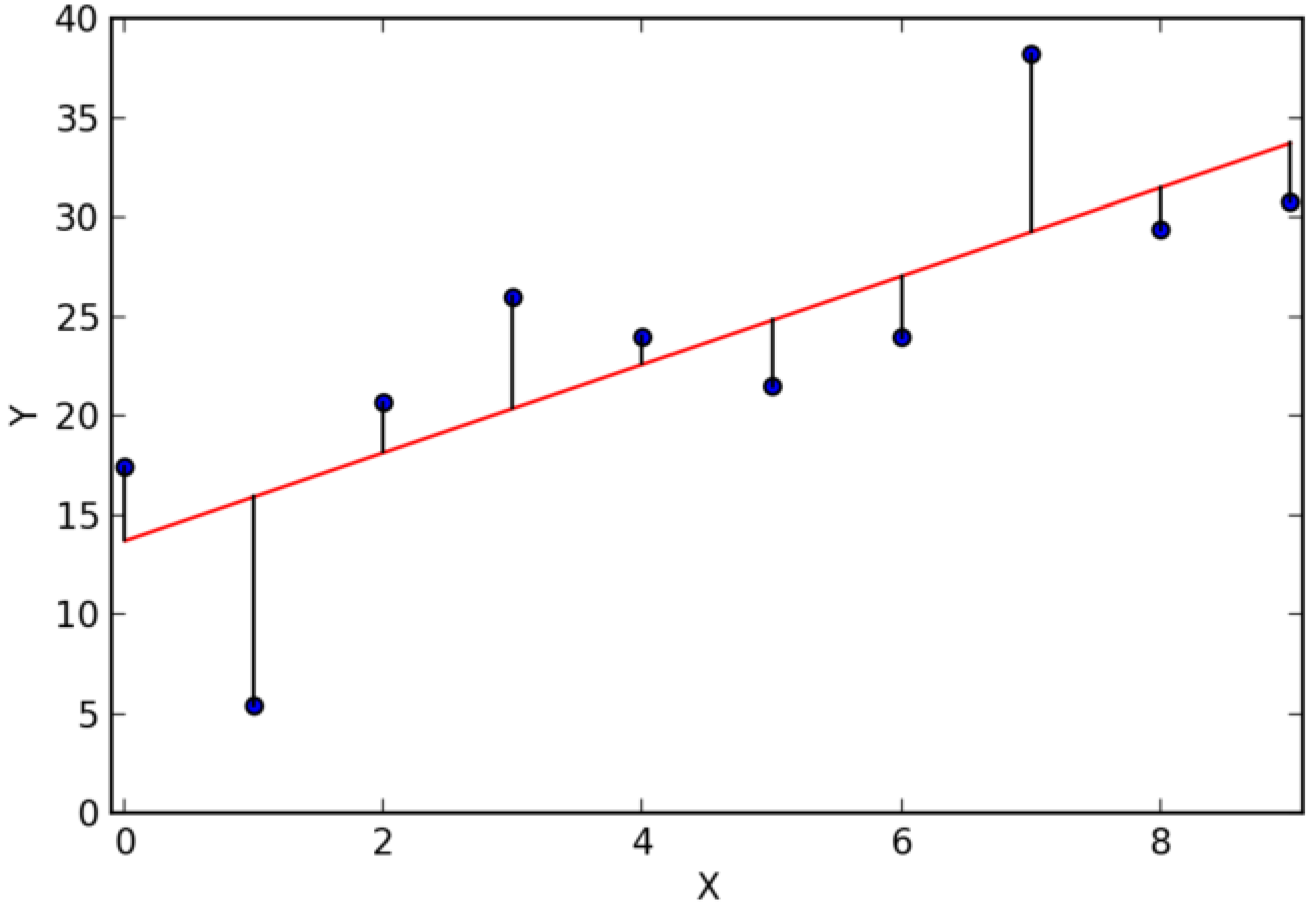
- When dependent variable (y) not linear in the parameters

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)

$$\{f(x;\theta)\colon\ \theta \epsilon \Theta\}$$

- Find $\theta \epsilon \Theta$ 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

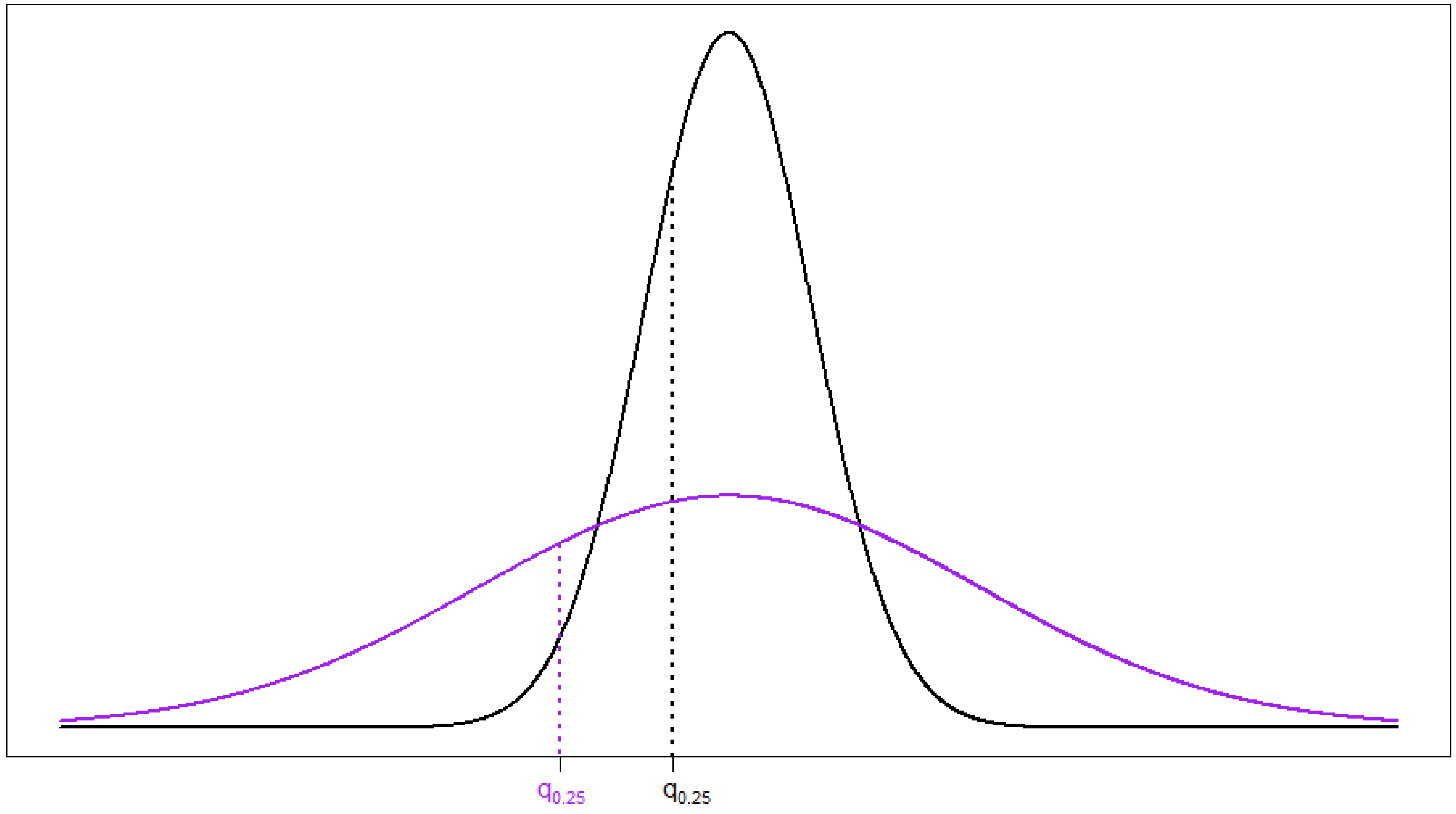
- Most methods typically use either Least Squares or Maximum Likelihood to fit the optimal parameters.
- When the model is fitted, it can be used for hypothesis testing or classification.
- The simplest model is fitting a straight line to some data points. This model has 2 parameters.
- According to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

Linear Regression

- (Least Squares) Linear regression is very widely used but we should not neglect the assumptions on which it relies
  - Normality
  - Linearity
  - Homoscedasticity (errors have equal variances)

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

Normality assumption (Q-Q-Plot)



Normality assumption (Q-Q-Plot)