

# 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.

We should be careful also about the number of decimals that we provide, the more decimal, the more confident, ultimately meaning that your model is very good. But then if the number change shortly after it means the estimate was not that good after all, and by changing it we lose in credibility.

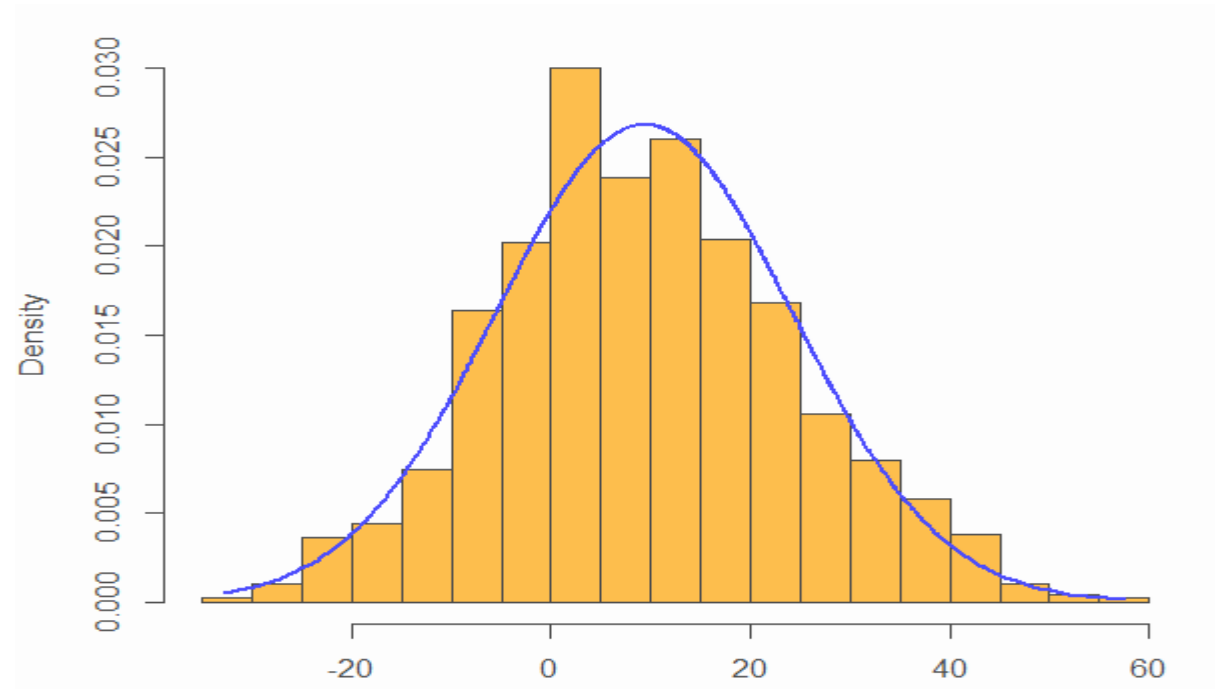
Lots of decimals can be only the result of the math, but that doesn't mean we should use them all.

# Parameter Estimation

If we are not interested in the distribution but we are interested in one value of this distribution the estimates mean and sigma are not really helpful: it is better in this case to provide an interval (the value lies between here and here)

## 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



The larger the sample the more confident we get about our estimate.

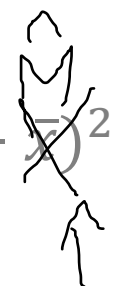
# Normal distribution

It makes sense to use the mean in descriptive statistic as we want to get as close as possible to the "real" value, especially when we are speaking about estimates.

- The normal distribution is uniquely characterized by its mean  $\mu$  and variance  $\sigma^2$  (standard deviation<sup>2</sup>)
- We can estimate those parameters by the sample mean and the sample variance

These two are reasonable parameters to describe a normal distribution.

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i,$$

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


More correctly here we should use the estimates mean

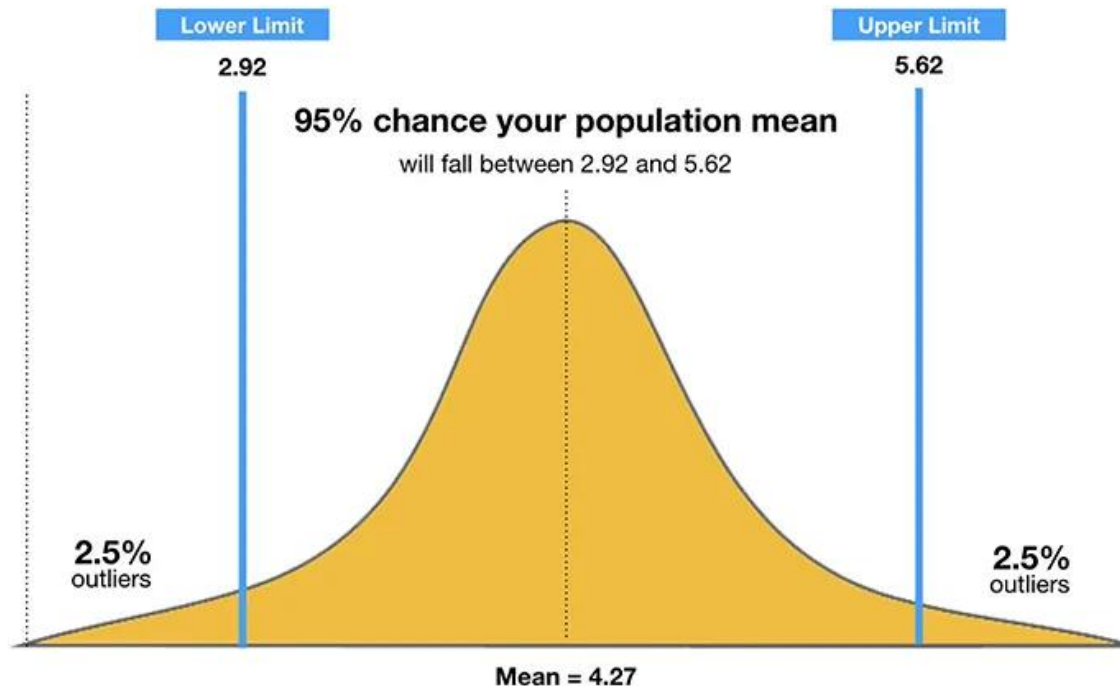
- Estimators often denoted with 'hat' e.g.  $\hat{\theta}$  is an estimate for  $\theta$ .
- During this CAS you will fit many other model parameters from data.

For confidence interval, the assumption is that I have a normal distribution.

# Confidence Intervals

The interval when we provide a single interval should also be reasonable; it should be the value with a specific certainty that the actual value will lie there.

- 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.

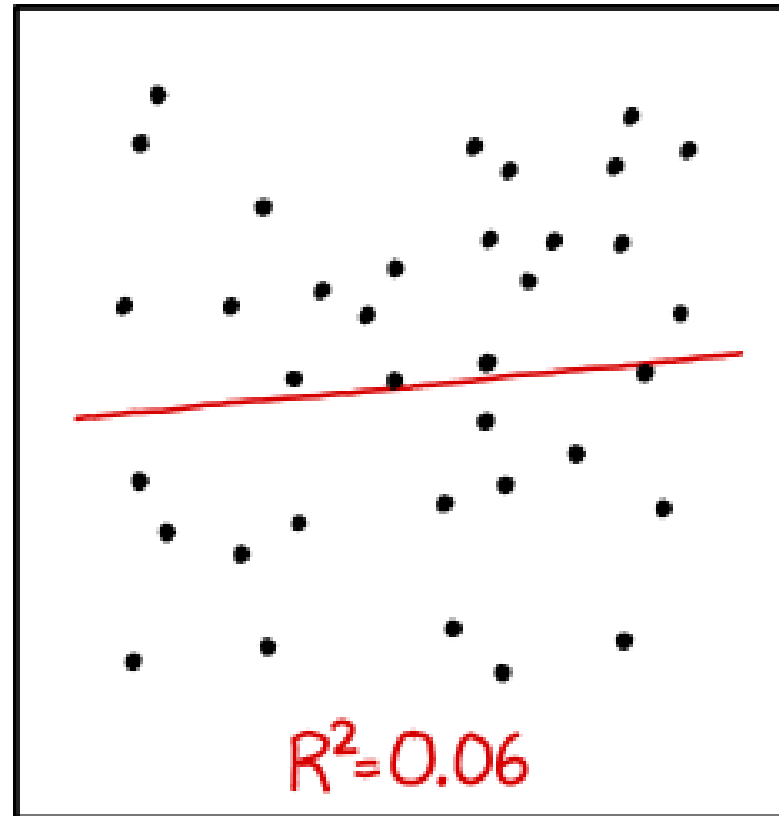


$$\bar{x} \pm t \cdot \frac{s}{\sqrt{n}}$$



# 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.

NOTE: in machine learning the different models are tried out until when a fitting one is found.

# Different types

Non-parametric: we are not sure if it's linear, or if it is not, we don't really know at all what kind of relationship could it be. We need to find more parameters, as for such regression we need more data in order to estimate the structure of the model.

## 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

The more far away we go with the extrapolation the more the confidence becomes less

## Non-linear

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

## Non-parametric

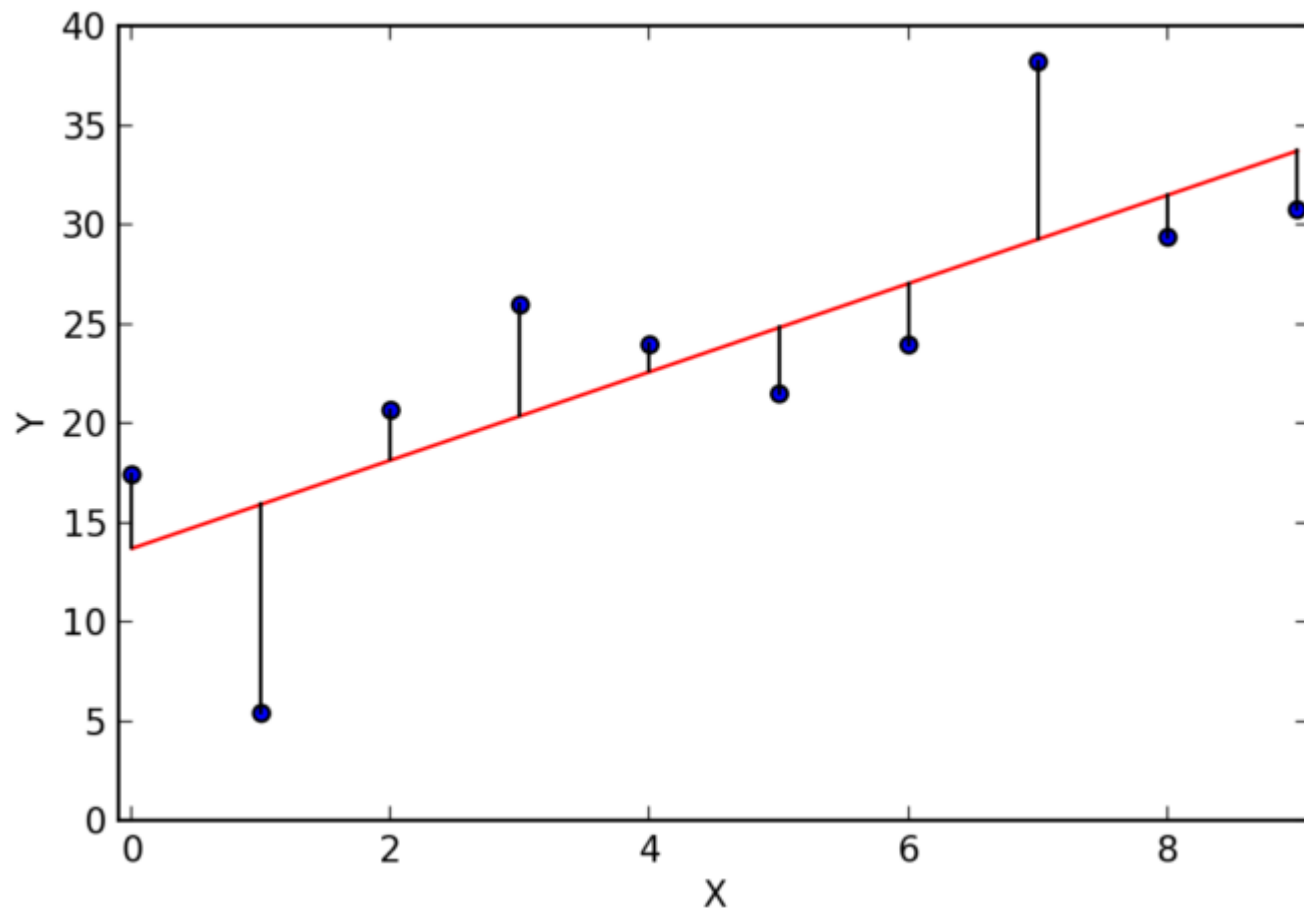
- Parametric models have parametric form
- If we have no clue about the exact relationship, we may use non-parametric estimation (e.g. isotone regression)
- These normally need more data, because also the model structure must be somehow estimated

Least squares could be potentially used also for curves others than only lines, but it becomes more complicated.

# Least Squares

Target is to minimize the distance (quadratical) of the points from the estimated line.

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



NOTE: All these models are parametric models.

# Maximum Likelihood

More difficult: we don't look at the differences of the data

- 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)  
Output is 0/1 values (logistic regression)
- 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

