UNIVERSITÄT BERN

CAS Applied Data Science - Module 2 – Day 2

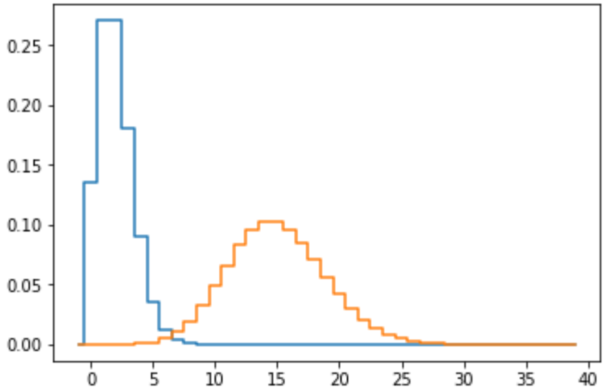
# Statistical Inference for Data Science

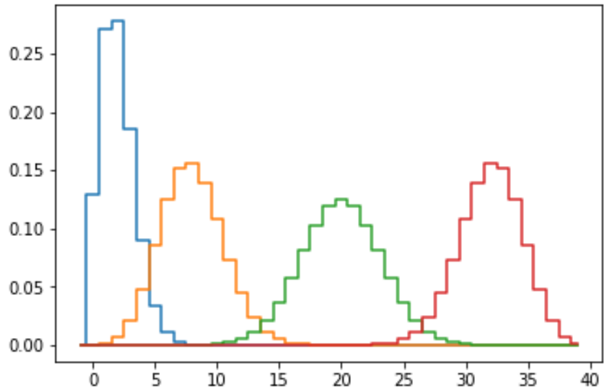
Prof. Dr. Géraldine Conti, PD Dr. Sigve Haug

Bern, 2020-08-26

# Questions from Notebook 2 (see also google form)

- What is the blue distribution ?
- Examples of use ?
- Difference with the other colored curve(s) ?




- Why are measured observables random variables ?
- Which probability distribution of a RV is the most important ?
- Thumb of rule, when is the normal distribution a good approximation ?
- Can you mention 5 descriptive statistical measures ?

# Exercise 2.3.4

## Exercise 2.3.4

It is important to obtain some routine with the computation of probabilities and quantiles.

Let X be binomially distributed with n = 60 and p = 0.4. Compute the following.

- P(X = 24) (PMF), P(X ≤ 24) (CDF)
- Compute the mean and standard deviation of X.

In [28]:

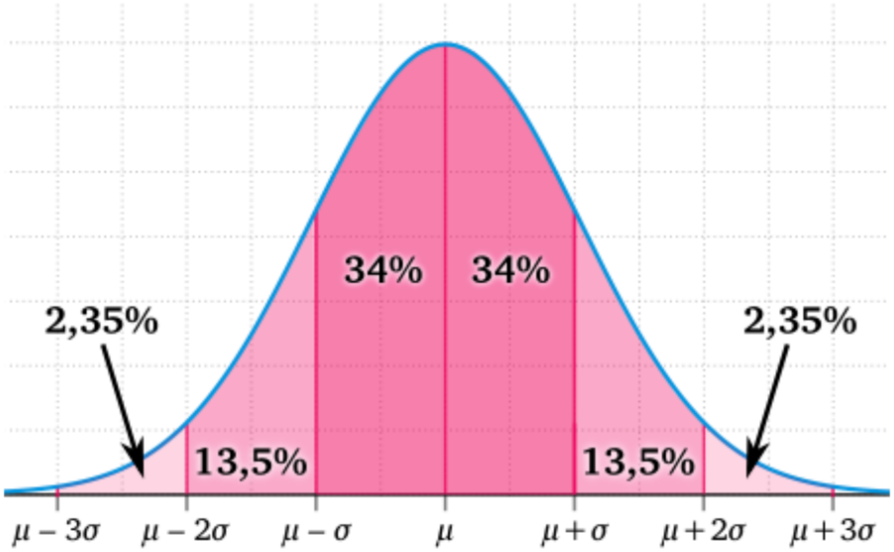
```
p24     = scipy.stats.binom.pmf(24,60,0.4) #
pLEQ24  = scipy.stats.binom.cdf(24,60,0.4) #

print(p24,pLEQ24)
mean, var, skew, kurt = scipy.stats.binom.stats(60, 0.4, moments='mvsk'
print(mean, var)
```

```
0.10466918336534053 0.5557755727497353
24.0 14.399999999999999
```

# Questions from Notebook 2

- Difference between statistical and systematic uncertainties ?
- Interpretation of one standard deviation ?



# $2^{nd}$ day: Parameter Estimation

**Two important estimation methods**

- Least squares
- Maximum likelihood

**Regression**

- Linear and non-linear
- Non-parametric

**More concepts**

- p-values
- Confidence Intervals

# Inferential statistics and parameter estimation

## Situation

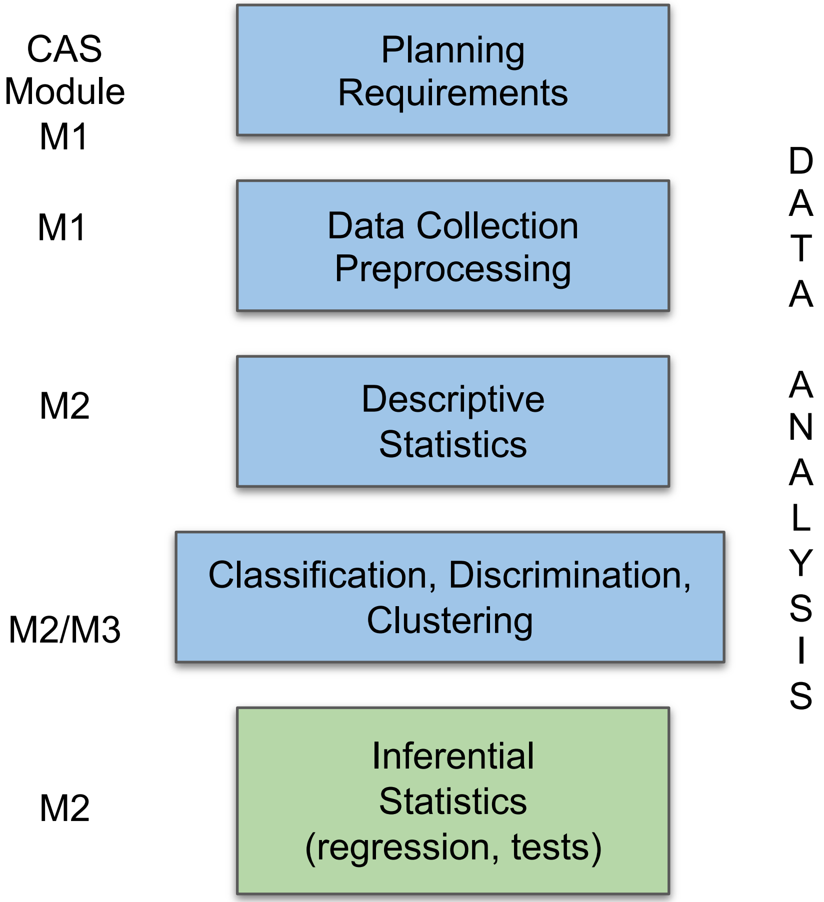
**Topics in Statistics** edit · view

| General topics | Probability | Descriptive statistics | Inferential statistics | Specialized topics |
|---|---|---|---|---|
| • Levels of measurement | • Probability theory | • Averages | • Hypothesis testing | • Computational statistics |
| • Sampling | • Random variable | • Statistical dispersion | • Estimator | • Decision theory |
| • Statistical survey | • Probability distribution | • Summary statistics | • Maximum likelihood | • Multilevel models |
| • Design of experiments | • Independence | • Skewness | • Bayesian inference | • Multivariate statistics |
| • Data analysis | • Expected value | • Correlation | • Non-parametric statistics | • Statistical process control |
| • Statistical graphics | • Variance, covariance | • Frequency distribution | • Analysis of variance | • Survival analysis |
| • History of statistics | • Central limit theorem | • Contingency table | • Regression models | • Time series analysis |

# Inferential statistics

- Statistical inference is the process of using data analysis to deduce properties of an underlying probability distribution.[1]

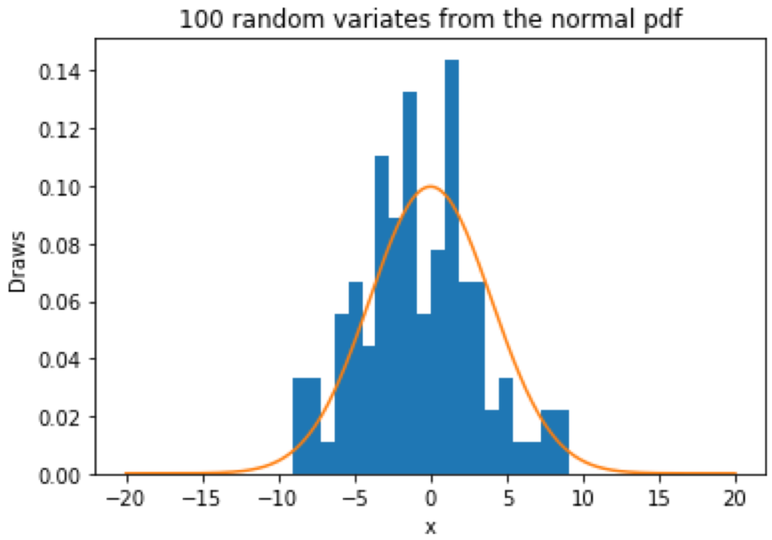
- Inferential statistical analysis infers properties of a population, for example by testing hypotheses and deriving estimates. It is assumed that the observed data set is sampled from a larger population.



# Parameter estimation

## Situation

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




Example: data in blue histogram, model in orange line graph. What is the mean and the sigma of the model?

# Common point estimators (from data)

## Mean

- Estimator for the mean of n measured x values

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

## Variance (and standard dev.)

- Another estimator

$$\widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu})^2$$

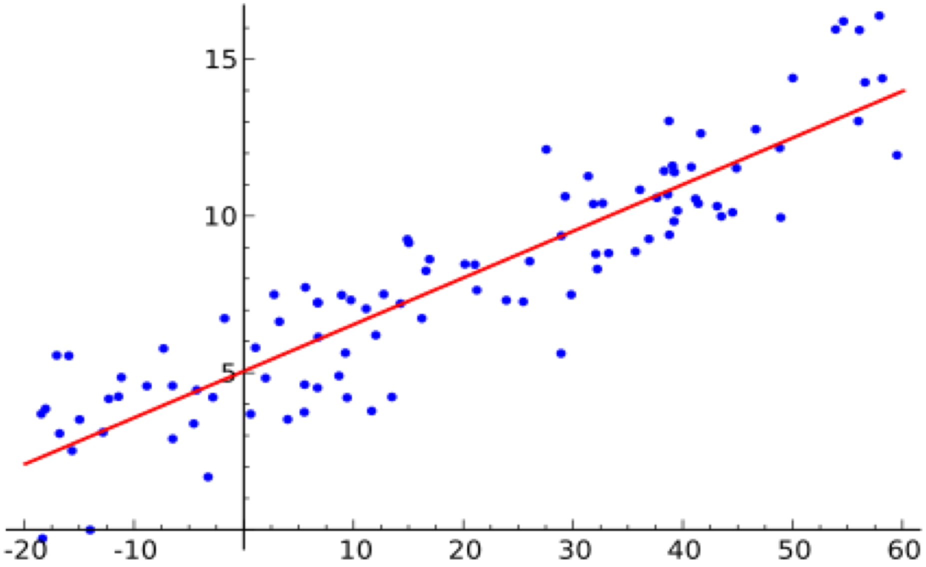
- Estimators often marked with *hat.* With Machine Learning techniques we fit many other model parameters from data.

# Two important estimation methods

- Least Squares
- Maximum likelihood

# Least Squares (LS) Method

- Minimise distance (residual) between data point and parameter
- Under certain conditions same as Maximum Likelihood Estimator



# Maximum Likelihood (ML) Method

- The probability $P(x|\theta)$ is the probability to have gotten the data x actually obtained, given the theory (a set of parameters $\theta$)
- The likelihood function $L(\theta|x)$ uses instead the data as input
    - Notation : $L(\theta) = L(\theta|x)$
    - It can be a simple function or something very complicated
- The ML estimator of $\theta$ maximises the likelihood function L
- For computational reasons, mostly the *–(log likelihood)* is used

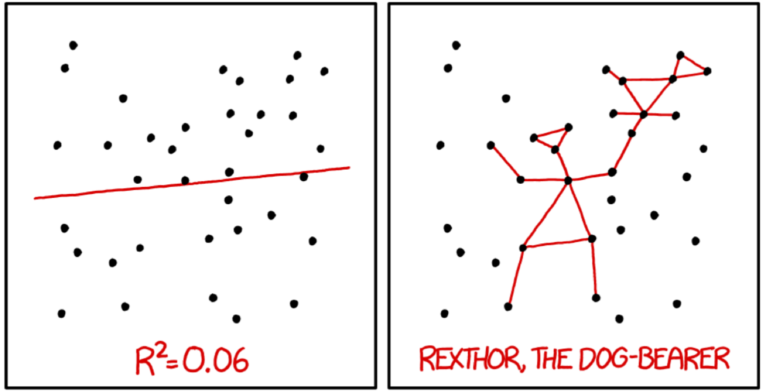
$$\boxed{-\frac{\partial \ln L}{\partial \theta_i} = 0\,, \qquad i = 1, \ldots, N}$$

# Example : Poisson distribution

- Probability Mass Function : $$\mathrm{P}(k|\mu) = \exp(-\mu)\frac{\mu^k}{k!}$$
- Example :
  - We fix $\mu = 20$ : probability of getting 30 is

    P(30|20) = exp(-20) x $20^{30}$ /30! = 0.07
  - We measure k=20 : with ML, we compute $\mu = 20$ !

# Regression




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.

- Linear and non-linear
- Non-parametric

*parameters from data - not vice versa*

# Different types

## Linear

- Example straight line y = ax + b
    - Fit a and b with LS or ML
- You can then predict the future (extrapolation and interpolation)
- *Linear* refers to the parameters

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$$

## Non-linear

- When dependent variable ($y_i$) not linear in the parameters, it is called non-linear regression (obviously)

## Non-parametric

- Parametric models have a shape assumed
- If we have no clue about the shape, we may use non-parametric estimations
- These normally need more data, because also the shape must be somehow estimated

# Typical Machine Learning Methods 1

When using data to fit the model (pdf) parameters, we call it machine learning. We can choose among an infinite number of models/methods. Popular ones are

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

# Typical Machine Learning Methods 2

Most (all?) methods typically use either Least Squares or Maximum Likelihood for finding the optimal parameters.

When the model is fitted, it can be used for hypothesis testing, i.e. future predictions or classification.

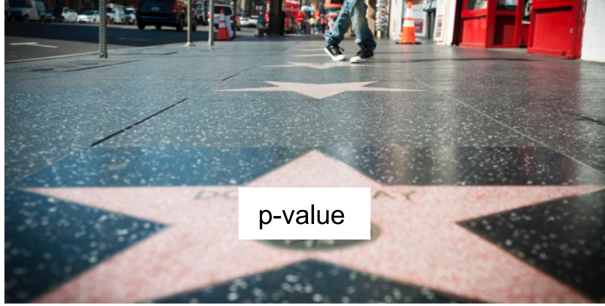
The simplest ML is fitting a straight line to some data points. The model has 2 parameters.

GPT-3 is a neural network with the capacity of 175 billion parameters (the bigges ever?)

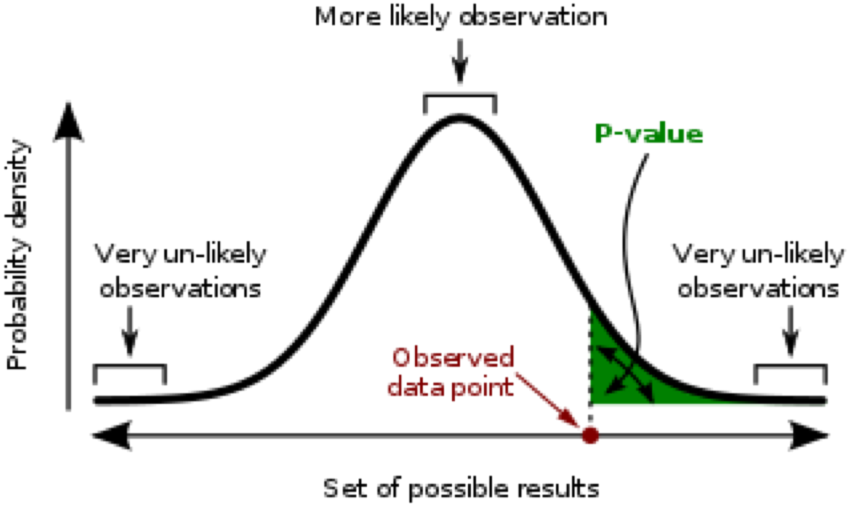
# More concepts

- p-value
- Confidence Interval

# p-value




- p-value is fraction of the surface above a certain data value
- Exercise (in pairs)
    - Assume a normal distribution of the age in the class.
    - Use the participants' ages to calculate the p-value of your age




A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

# Confidence Intervals (CI)

- One sided and two sided

- In the normal case we have the nice relation between standard deviations and confidence levels: 2SD = CL of 95 %