UNIVERSITÄT BERN

http://www.math.unibe.ch/continuing_education/cas_applied_data_science/index_eng.html

CAS Applied Data Science - Module 2 – Day 1

Statistical Inference for Data Science

Start working on Notebook 1 (Ilias) !

Dr. Géraldine Conti, PD Dr. Sigve Haug

Bern, 2020-08-25 , Sigve joins 09:15

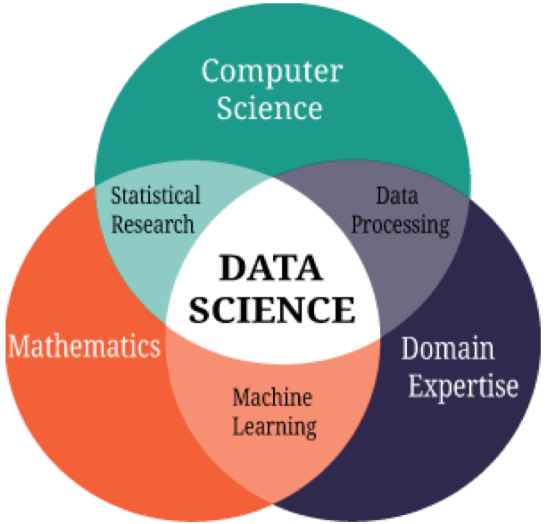
Welcome in the Data Science World !

Data Science uses

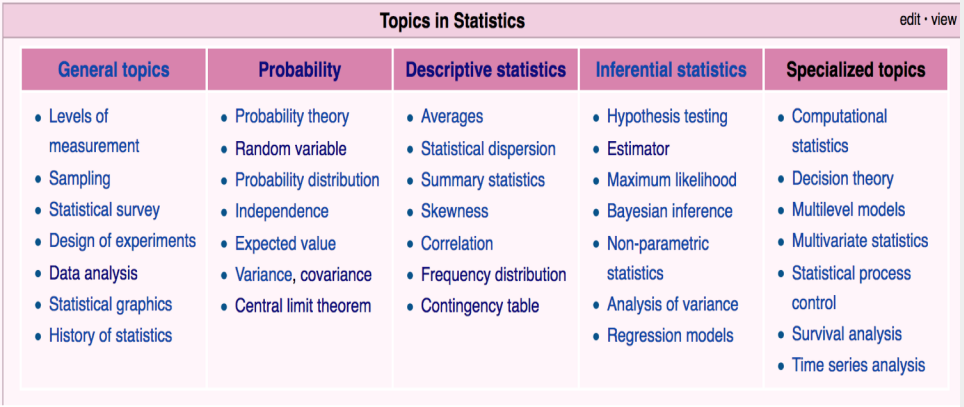
- Mathematics and Statistics
- Computer Science
- Domain expertise

on data to build information and extract knowledge (for decisions and actions)

Very general skills increasingly needed in all empirical research and business



Module 2



Topics in Statistics edit · view

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

Schedule

gconti.epfl@gmail.com
sigve.haug@math.unibe.ch

First day

- Descriptive statistics

Describe the available data

Second day

- Inferential statistics: Parameter Estimation

Draw conclusions from the data

Third day

- Inferential statistics : Hypothesis Testing

Fourth day

- Putting it all together

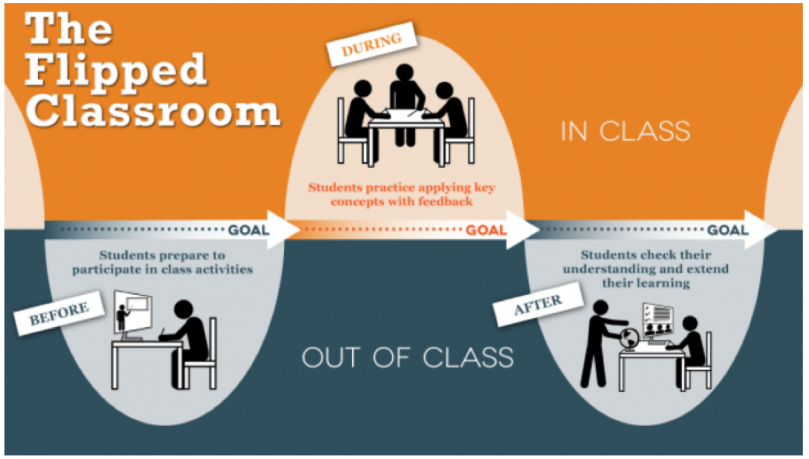
Project

- Presentation session 2020-10-16/19

Teaching method

Inverted classroom based

- Introduction lectures
- Real content you learn yourself with the **notebooks**
- Discussion sessions based on your questions and comments
- **Project** : Poster with poster presentation
- *1-2 questions to post every day on the chat*



Why

- Supposed to be better
- More fun
- Learning by doing

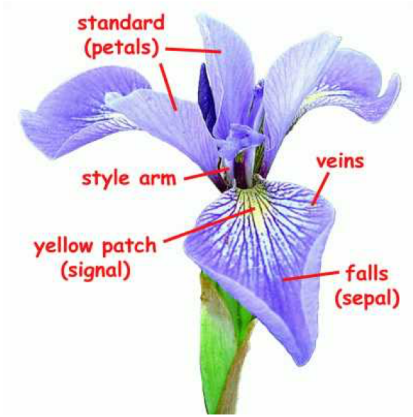
To give back sense to being present (Marcel Lebrun)

Description

The Iris Dataset

- 3 classes : setosa, versicolour, virginica
- 4 observables :
  - petal : length, width
  - sepal : length, width

➤ *How could you use this to characterize your dataset ?*



Foretaste (4th day)

Descriptive statistics, hypothesis testing, …

Some new company recently sequenced the genes of the Iris species Setosa and patented it, apparently in order to preserve this species because it is so beatiful. Due this patent it is not allowed to change the plant.

A big farmer and hater of Iris and with a field where Iris is a disturbing weed, has been using a new product from Sonte Manto for a couple of years. The product is supposed to effectively kill Iris plants.

A big Iris lover collected a sample of Iris plants from the farmer's field and thinks the Iris Setosa setal leaves are bigger than normal. She sent the sample to the company, which in turn came to the conclusion that Setosa must have mutated due to the product from Sonte Manto.

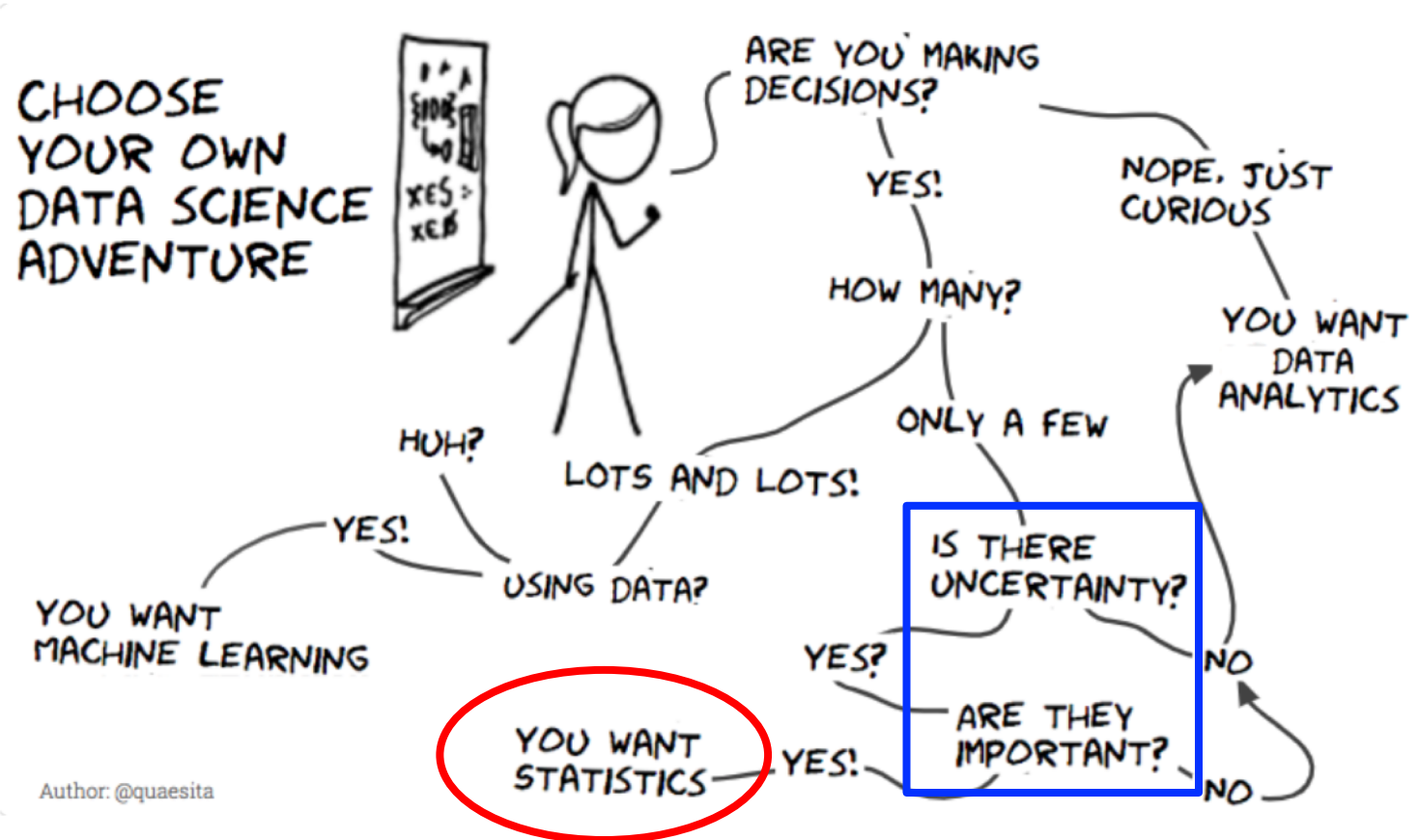
So the company sued Sonte Manto with the claim that they have changed the plant with their product. Sonte Manto may risk to pay a billion dollars.

The court is asking you to give a neutral and scientific advice.

The Project

- Find your own dataset of interest
- Group of 2 people
- 15min presentation, 15min questions
- Half-day presence on October 16th (morning) or 19th (either morning or afternoon)

The context


CHOOSE YOUR OWN DATA SCIENCE ADVENTURE
ARE YOU MAKING DECISIONS?
YES!
NOPE, JUST CURIOUS
HOW MANY?
YOU WANT DATA ANALYTICS
ONLY A FEW
LOTS AND LOTS!
HUH?
YES!
USING DATA?
YOU WANT MACHINE LEARNING
IS THERE UNCERTAINTY?
YES?
NO
ARE THEY IMPORTANT?
YOU WANT STATISTICS
YES!
NO
Author: @quaesita

Describing Data

Why

- Learn about the distributions
- Features, patterns
- Outliers, quality etc
- No inference can be better than what the data gives
- Create trust

Good description is also the basis for good inference. From good description we can easier choose which model we want to use to infer relationships from our data.

Many possibilities

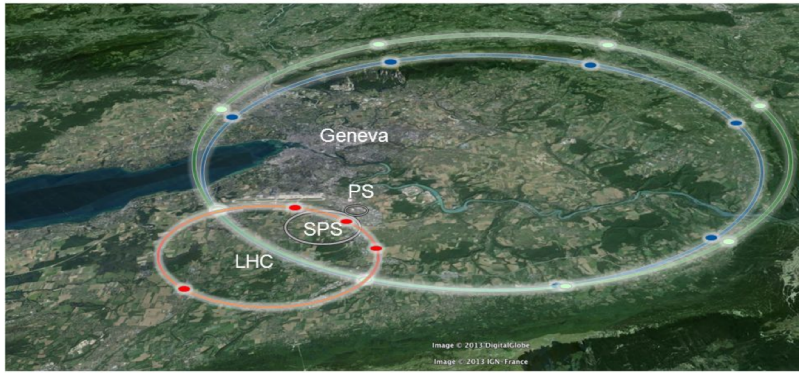
- Listing it all
- Words
- Mathematics (**statistics**)
- Tables
- Graphs (visualisation)
- Animations etc ...

Good description is the basis for good *inference*

- Descriptive statistics helps choosing a good model
- The model can then be used for inference
  - Interpolation
  - Extrapolation
  - Hypothesis testing
  - Discoveries and exclusions

- Basis for good decisions
- Do you want to build this ?

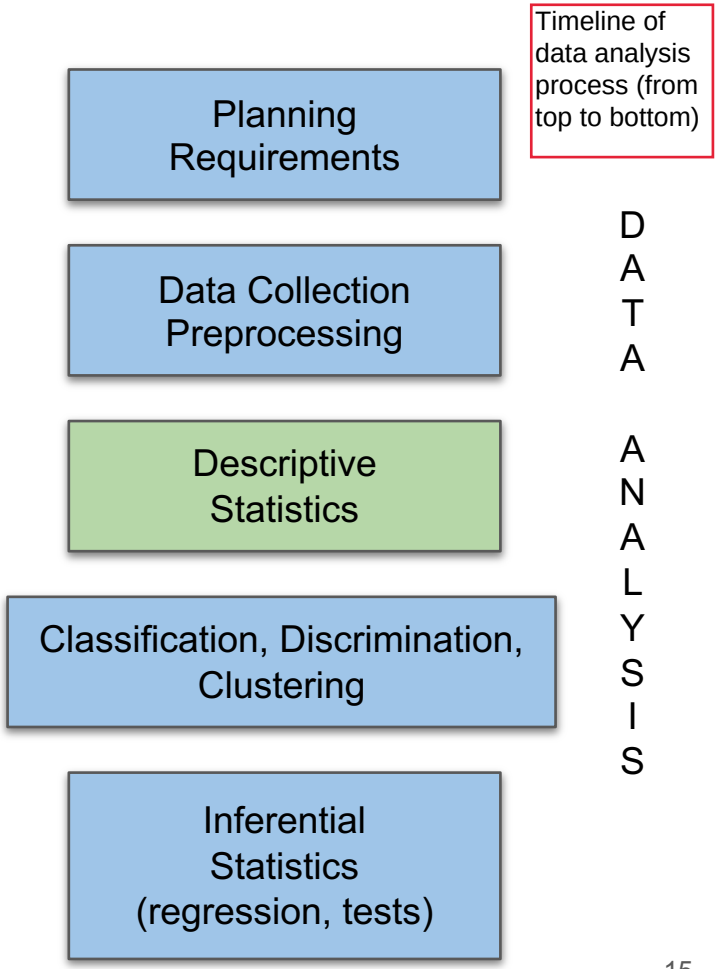
Some possible FCC-hh *geometries*



LHC	HE-LHC	FCC-hh	FCC-hh
27 km, 8.33 T	27 km, **20 T**	80 km, **20 T**	100 km, **16 T**
14 TeV (c.o.m.)	33 TeV (c.o.m.)	100 TeV (c.o.m.)	100 TeV (c.o.m.)

Descriptive statistics : Definition

- Descriptive statistics are used to describe the basic features of the data in a study. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.



Descriptive statistics : Central Topics

1) Probability interpretation

how likely is it for a value to exist/be true?

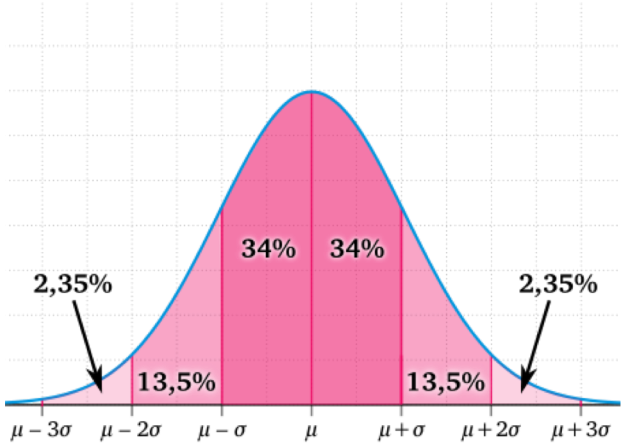
2) Random Variables (RV)

3) Models:
   Probability density functions(p.d.f.) (continuous)
   probability mass functions (p.m.f) (discrete)

4) Moments (center, shape, dispersion)

5) Summary tables and graphs (visualisation)

6) Dependence (correlations,...)

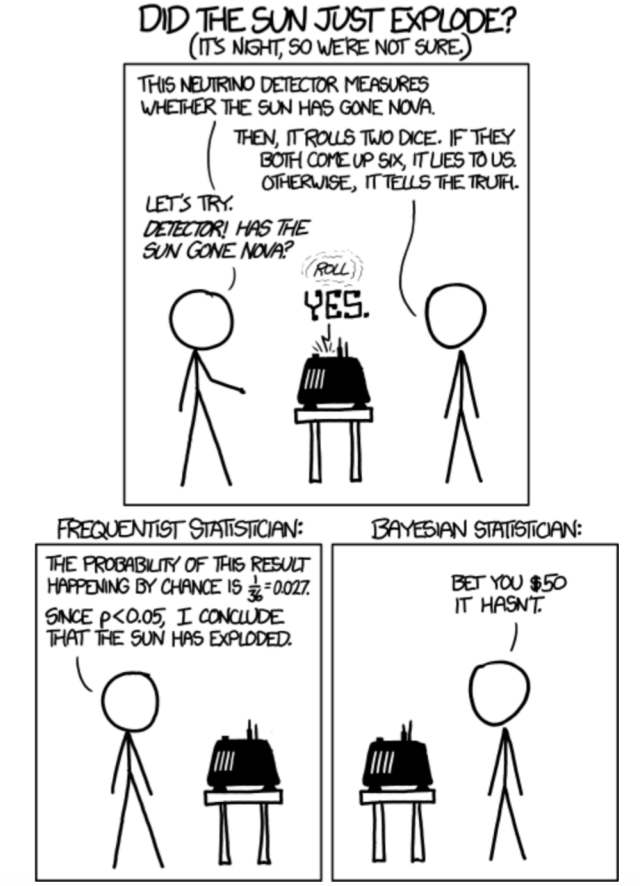


1) Probability interpretation

Objectivist

- Relative frequency
- **Frequentist**
- Frequentist statistics

$$P(A) = \lim_{n \to \infty} \frac{times\ outcome\ is\ A}{n}$$



Subjectivist

- Degree of belief
- **Bayesian**
- Bayesian statistics

$$P(A) = degree\ of\ belief\ that\ A\ is\ true$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

2) Random Variable (RV)

In probability and statistics, a random variable, random quantity, aleatory variable, or stochastic variable is described informally as a variable whose values depend on outcomes of a random phenomenon.The formal mathematical treatment of random variables is a topic in probability theory. In that context, a random variable is understood as a measurable function defined on a probability space that maps from the sample space to the real numbers

Why is it random?

- Based on a sample, not the full population
- Limited resolution
- Quantum mechanics

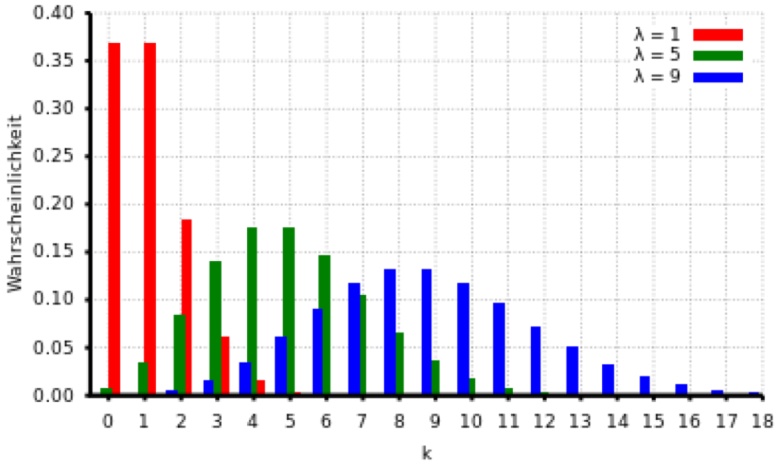
Example

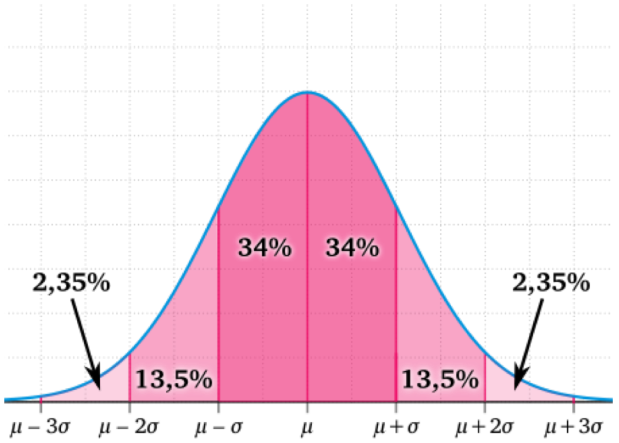
- *RV = amount of CHF on Swiss bank accounts*

this isn't the same for every account (random)--> so we calculate the mean (which is also random)

- Sample by checking 1000 accounts
- Get a (normal?) distribution
- A new sample will not yield exactly the same distribution (statistical uncertainty)

3) Probability density/mass distributions





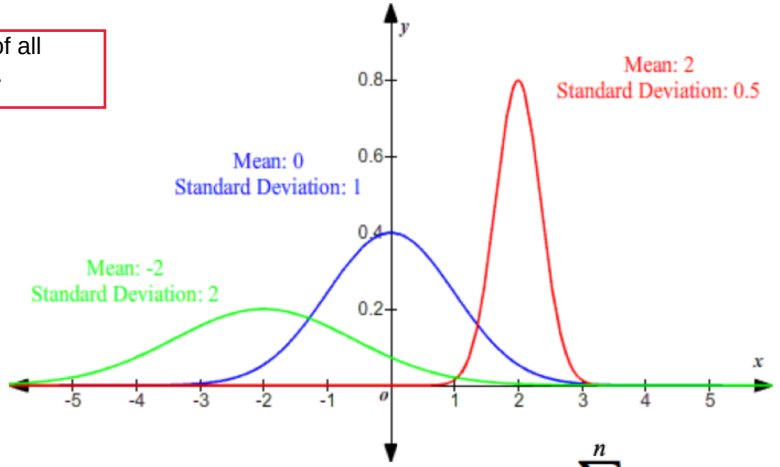
➤ *Which is which (pdf, pmf) ?*

4) Moments

$$\alpha_n \equiv E[x^n] = \int_{-\infty}^{\infty} x^n f(x)dx$$

Definition of all Moments...

the 0th moment is just the total probability (=1)



The moments are used to describe the data :

- 1. Moment (n=1) : Mean

...so for the first moment you have x^1

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\mu \equiv \alpha_1$$

- 2. Moment (n=2) : Variance σ^2
  - Standard deviation σ
- 3. Moment (n=3) : Skew (symmetry)
- 4. Moment (n=4) : Kurtosis (tails)

$$\sigma^2 \equiv V[x] \equiv \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \ldots = \alpha_2 - \mu^2$$

Standard error : $s_{\bar{x}} = \frac{s}{\sqrt{n}}$

Standard error is not equal to standard deviation. The standard error gets smaller and smaller the more measurements you have (the closer you get to the "real" values)

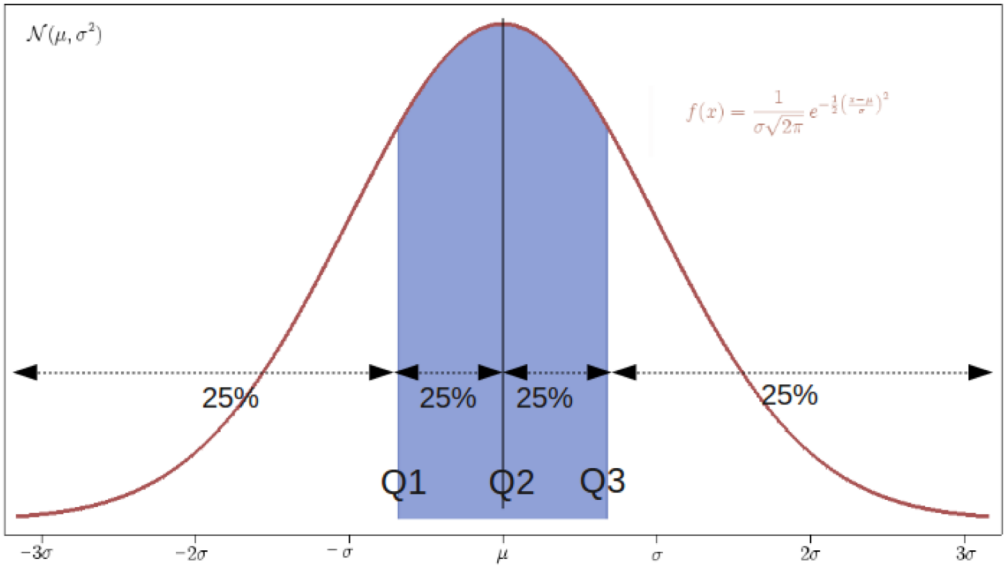
$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

Quantiles

The value below which x% of the data is located.

Values indicating certain surface fractions

- Percentiles
- Median (50%)
- Quartiles



Are you able to ”see or imagine” this distribution as a box plot?

5) Summary tables

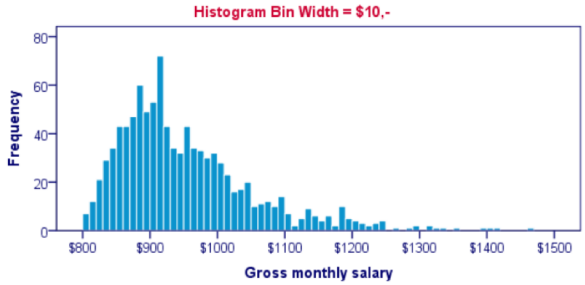
Grouped data

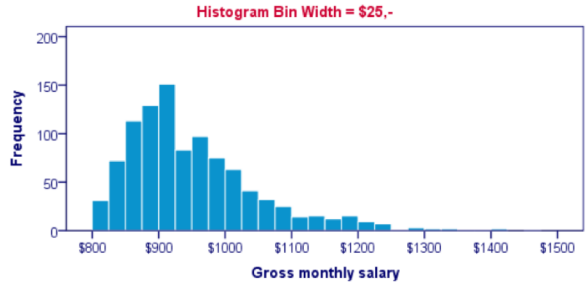
- Frequency of data in defined bins/ groups

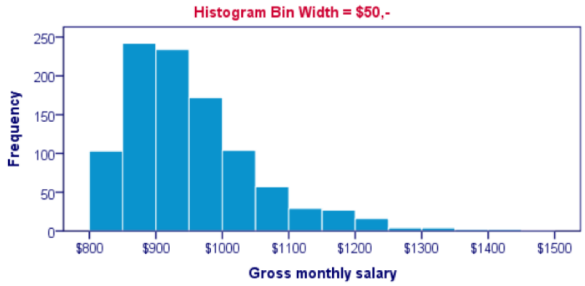
- The graph to the table would be a histogram

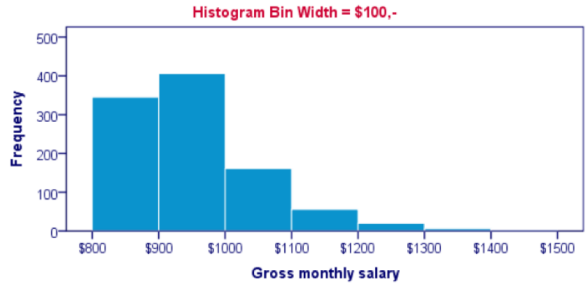
20	25	24	33	13
26	8	19	31	11
16	21	17	11	34
14	15	21	18	17

Time taken (in seconds)	Frequency
$5 \leq t < 10$	1
$10 \leq t < 15$	4
$15 \leq t < 20$	6
$20 \leq t < 25$	4
$25 \leq t < 30$	2
$30 \leq t < 35$	3







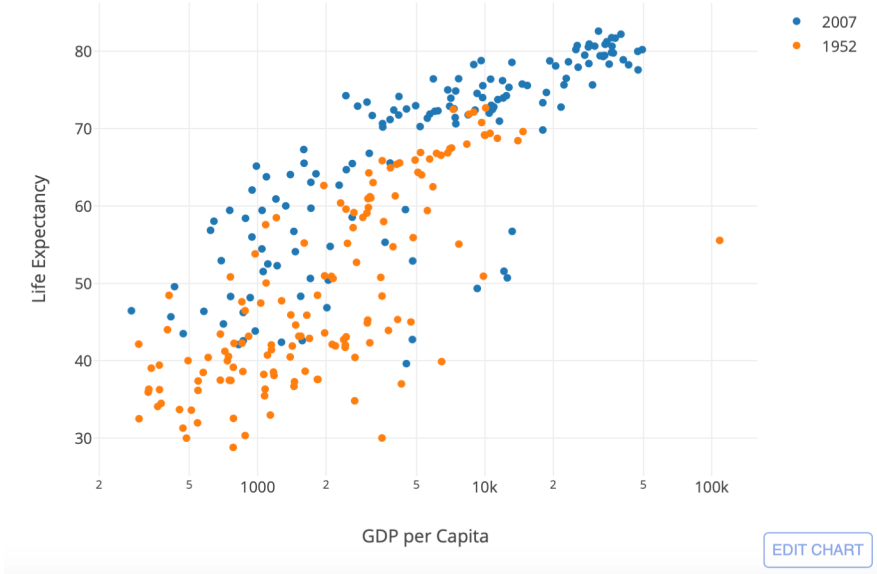


5) Summary tables

Contingency tables

- Frequency of multivariate data in a matrix format
- Here two random variables (sex and handedness)
- The graph to the table would be a scatter plot

Sex \ Handedness	Right handed	Left handed	Total
Male	43	9	52
Female	44	4	48
Total	87	13	100

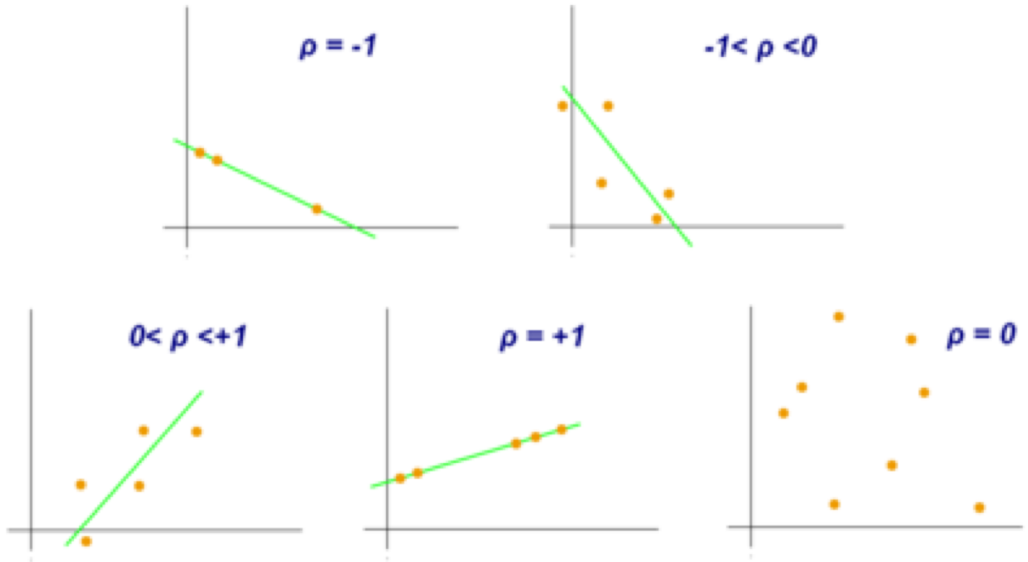


$$\operatorname{cov}(X,Y) = \operatorname{E}\left[(X - \operatorname{E}[X])(Y - \operatorname{E}[Y])\right]$$

6) Dependence

Population Covariance Formula	**Cov(X,Y)=** $\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{N}$
Sample Covariance Formula	**Cov(X,Y)=** $\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{N-1}$

- Two random variables with covariance= 0 are said to be uncorrelated
  - The normalised covariance is the correlation ρ
  - Two random variables with positive covariance are said to be correlated
  - negative is anti-correlated
  - If the variables are the same, we get the variance (Cov(**X**,**X**))



The Iris Dataset

- 3 classes : setosa, versicolour, virginica
- 4 observables :
  - petal : length, width
  - sepal : length, width

➤ *How could you use this to characterize your dataset ?*