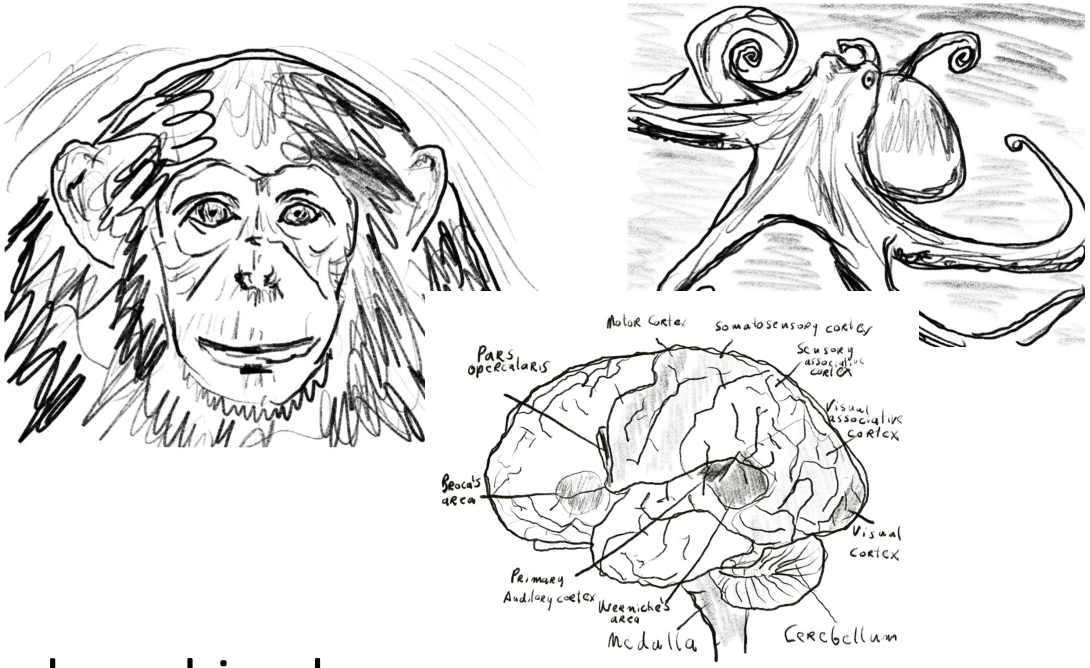
Lecture 1: Learning and Univariate Data Analysis

Andreas Wichert
Department of Computer Science and Engineering
Técnico Lisboa

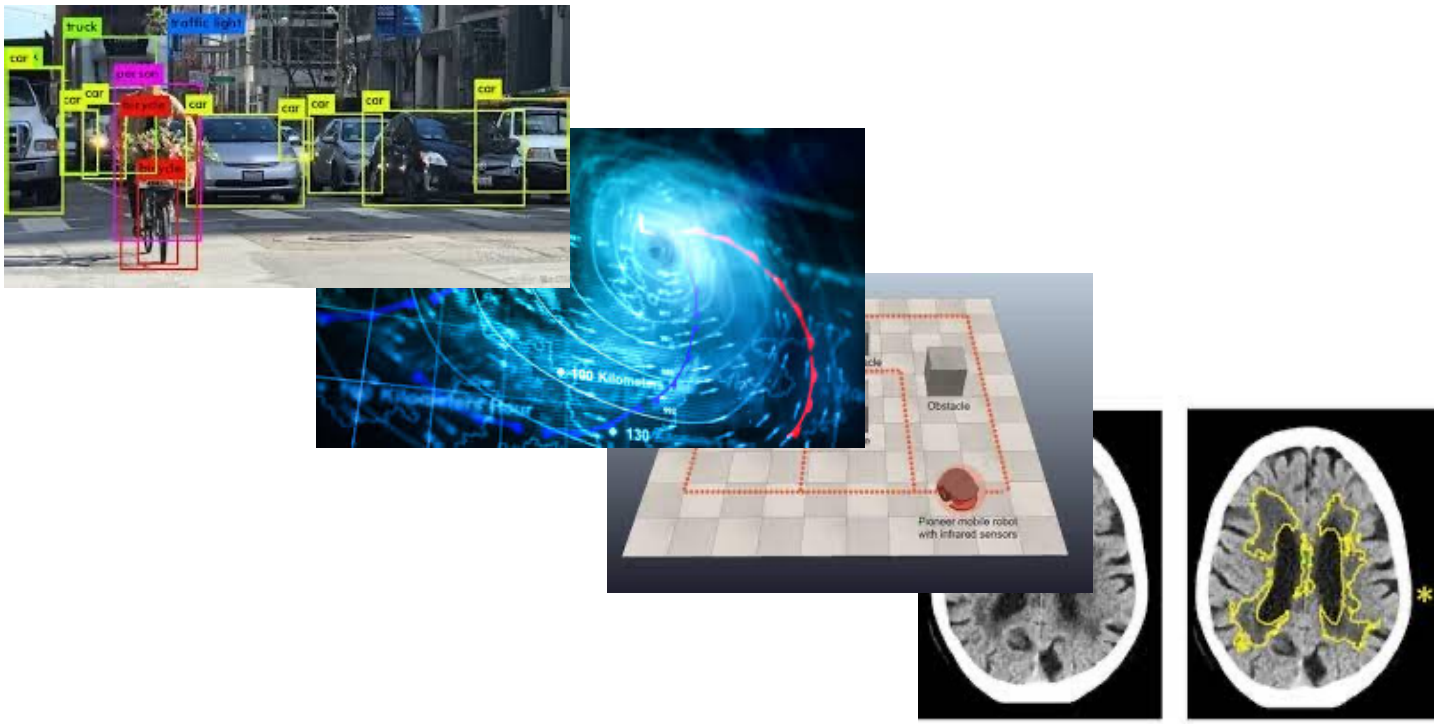
What is machine Learning?



- Parallels between “animals” and machine learnıng
- Many techniques derived from efforts of psychologist / biologists to make more sense “animal” learning through computational models

Statistical Machine Learning

- Changes in the system that perform tasks associated with AI
  - Recognition
  - Prediction
  - Planning
  - Diagnosis



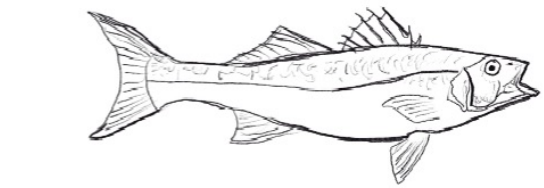
Learning Input output functions

- Supervised
  - With a teacher

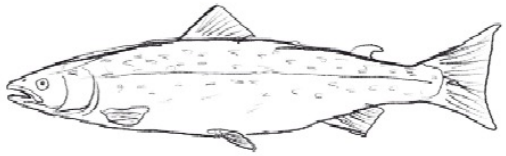
- Unsupervised
  - Without a teacher

- Reinforcemet Learning
  - Actions within & responses from the environment
  - Absence of a designated teacher to give positive and negative examples

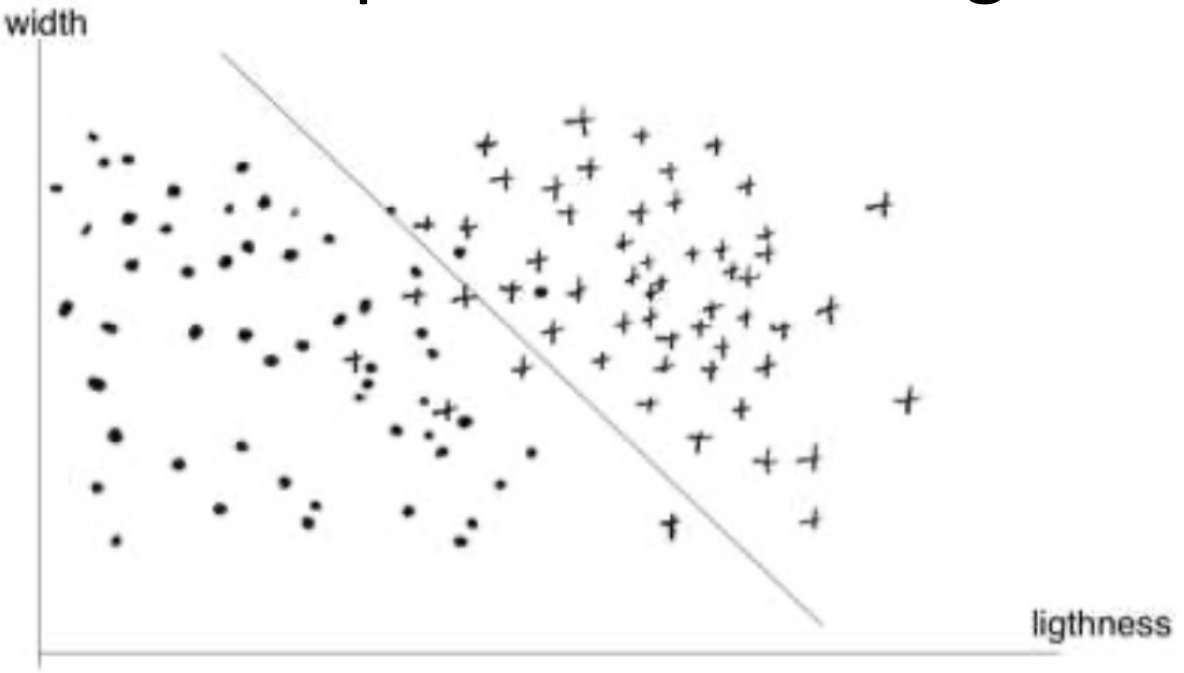
Supervised Learning



(a)



(b)



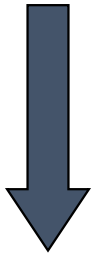
- In a sample, each salmon (b) is represented by a **dot** and each sea bass (a) by a **cross**
- We separate the two classes using a straight line

- Later, during classification, an unknown fish is described by a two-dimensional vector, and depending on which side of the separating line it falls, it is classified as either salmon or sea bass.

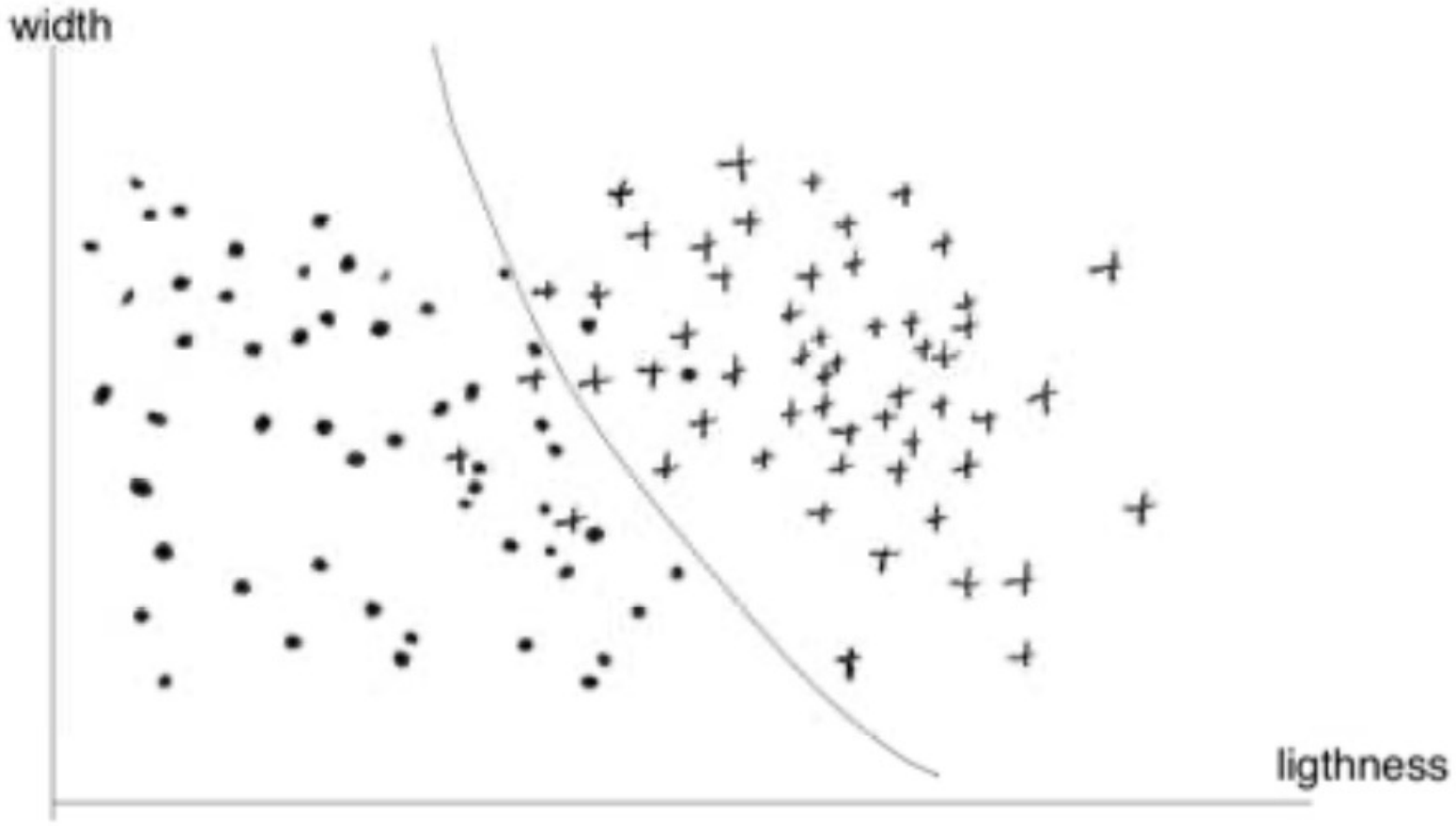
- We might add other features that are not correlated with the ones we already have. A precaution should be taken not to reduce the performance by adding such “noisy features”

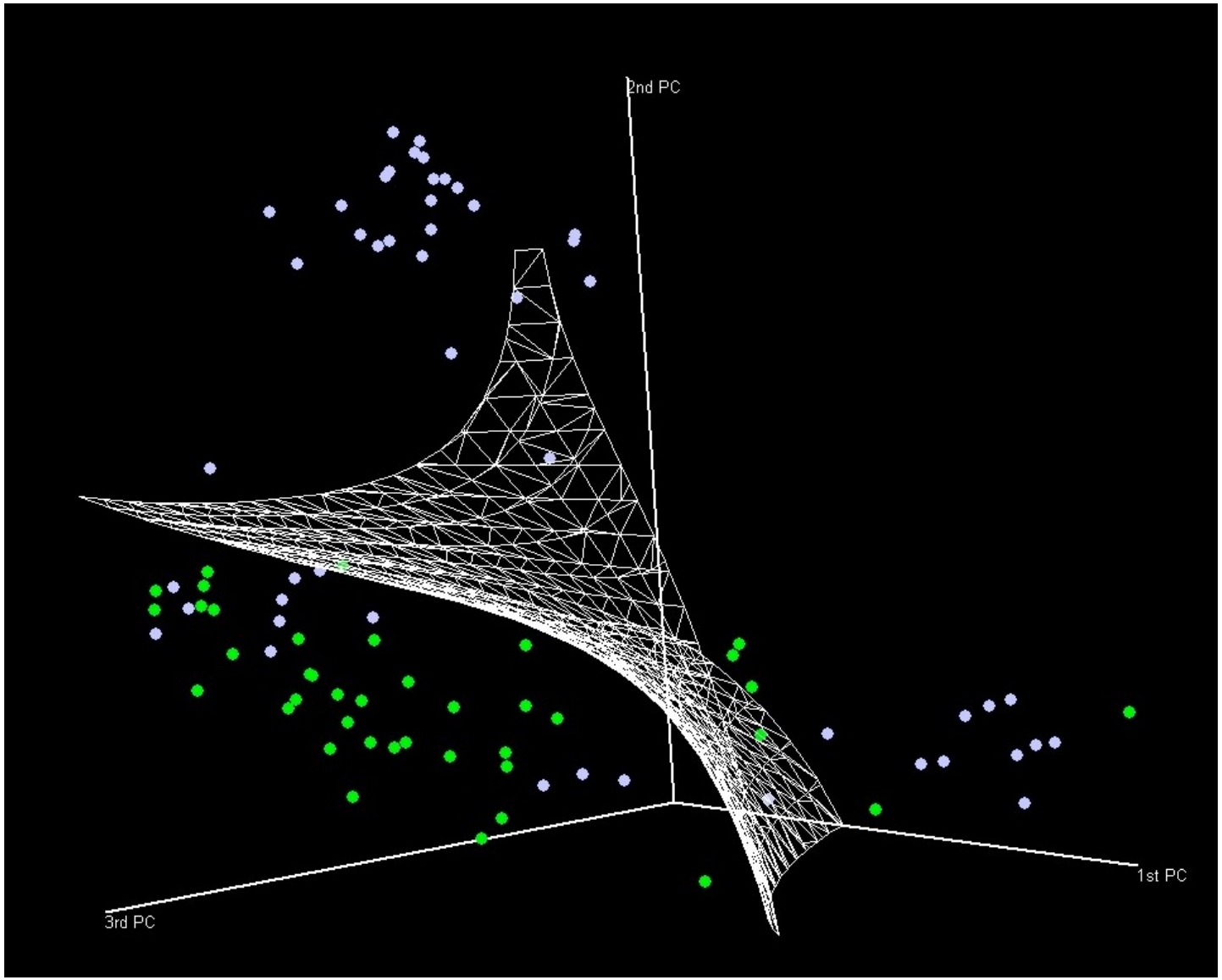
- Ideally, the best decision boundary should be the one which provides an optimal performance such as in the following figure:

- However, our satisfaction is premature because the central aim of designing a classifier is to correctly classify novel input

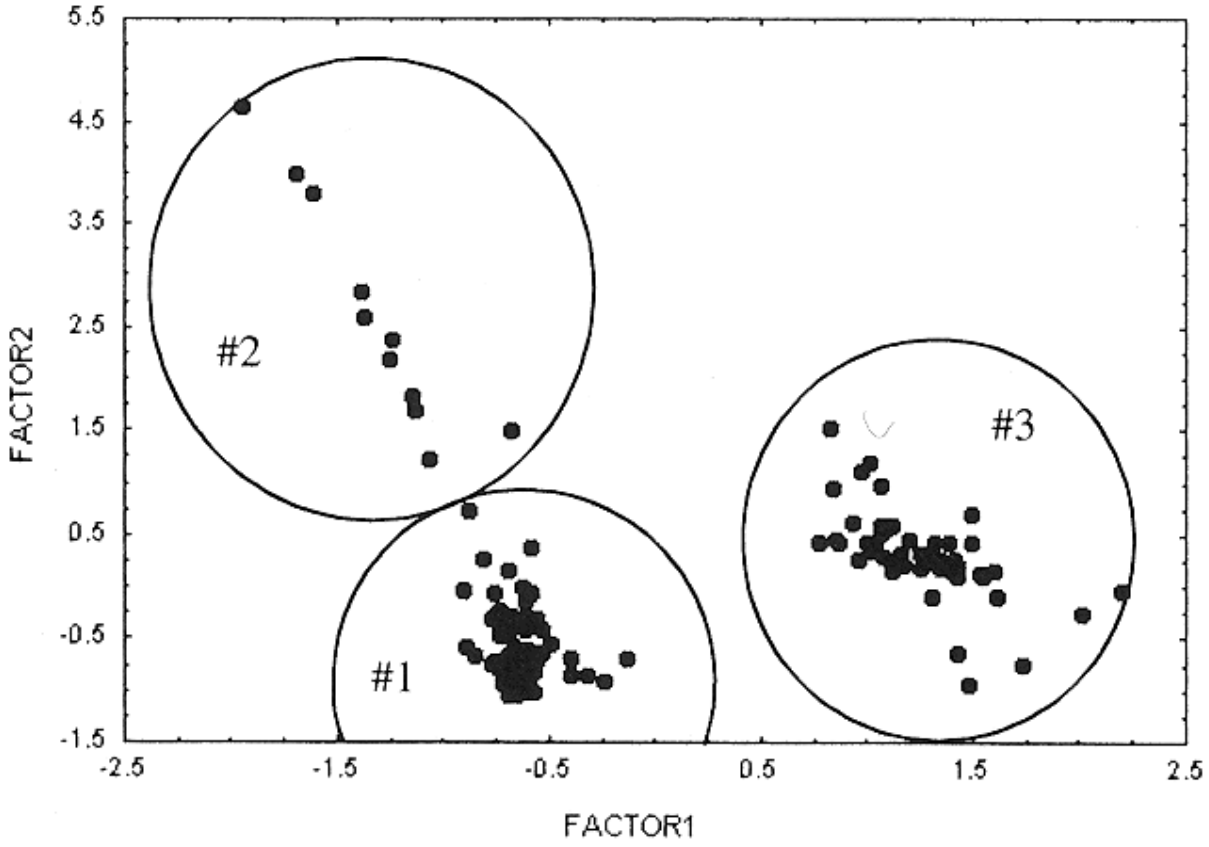


Issue of generalization!


width
ligthness

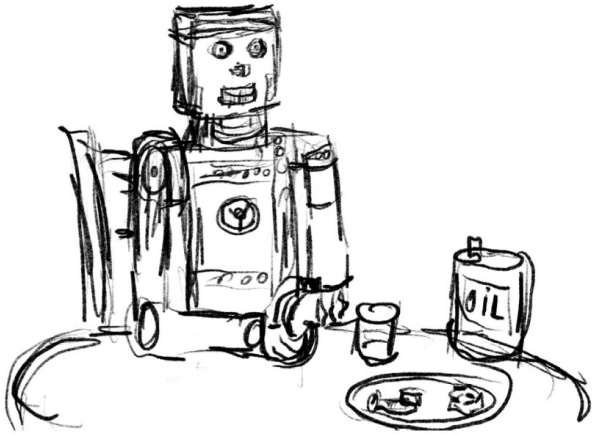

2nd PC
1st PC
3rd PC

Unsupervised Learning



Before *Machine* Learning…..

- Step 1: Select and **Understand your Data**
- Step 2: Preprocess/Prepare your Data
- Step 3: Transform Data

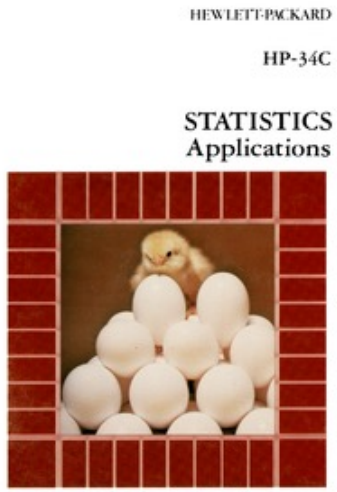


How to Prepare Data for Machine Learning

- Machine learning algorithms *learn from data*
- It is critical that you *feed them the right data* for the problem you want to solve
- Even if you have good data, you need to make sure that it is in a useful scale, format and even that meaningful features are included

1. Select Data and **Understand your Data**

- There is always a strong desire for including all data that is available, that the maxim “more is better” will hold
- This *may or may not* be true
- You need to consider what data you actually need to address the question or problem you are working on
- **What do I want to show?**



Univariate data analysis

- **Random**/aleatory **variable**
  - function $X : \Omega \rightarrow \mathrm{E}$ from a **sample space** Ω to a **measurable space** E
  - height variable is a function which maps a person from a population Ω to her height in $\mathbb{R}^+$
    - the observed height is referred as a **measurement**
  - from now one, we will refer *random variable* simply as *variable*
- **Univariate data**
  - single input variable
  - comprises univariate data statistics and, in the presence of an output variable, **bivariate data statistics**
- **Multivariate data**
  - multiple (input) variables
  - **multivariate order** = number of (input) variables

Statistics for one Variable - Univariate data

- Sample Size
  - The sample size denoted by *N*, is the number of data items in a sample

- Mean *(arithmetic mean)*
  - The arithmetic mean is the average value, the sum of all values in the sample divided by the number of values

$$\mu = \overline{x} = \frac{1}{N} \sum_{k=1}^{N} x_k$$

Harmonic mean

- Harmonic mean can be expressed as the reciprocal of the arithmetic mean

$$H = \frac{N}{\sum_{k=1}^{N} \frac{1}{x_k}}$$

- Harmonic mean of 1, 4, and 4 is

$$\left(\frac{1^{-1} + 4^{-1} + 4^{-1}}{3} \right)^{-1} = \frac{3}{\frac{1}{1} + \frac{1}{4} + \frac{1}{4}} = \frac{3}{1.5} = 2\,.$$

Basic understanding of how spread out numbers in the data

- Maximum
- Minimum

- Range:= Maximum-Minimum
  - Range is the difference between maximum and minimum
  - It is a **very crude** measurement of the spread of data

  - *1, 2, 3, 4, 6, 7, 7, 8.*
  - Max=*8*, Min=*1*, Range=8-1=7

Standard Deviation and Variance

- *Square root* of the variance, which is the sum of squared distances between each value and the mean divided by **population** size (finite population)

  - Example
    - 1,2,15 Mean=6
    - $\frac{(1-6)^2 + (2-6)^2 + (15-6)^2}{3} = 40.66$

$$\sigma = \sqrt{\frac{1}{N} * \sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2}$$

σ=6.37

Sample Standard Deviation and Sample Variance

- *Square root* of the variance, which is the sum of squared distances between each value and the mean divided by **sample** size

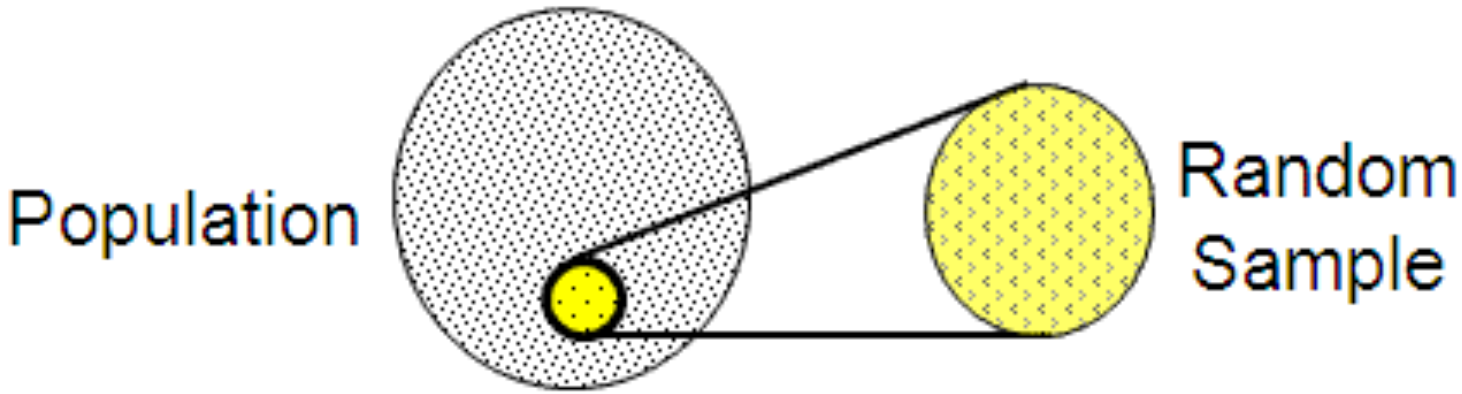
$$s = \sqrt{\frac{1}{N-1} * \sum_{i=1}^{N} \left(x_i - \overline{x}\right)^2}$$

- Example
  - 1,2,15 Mean=6
  - $\frac{(1-6)^2 + (2-6)^2 + (15-6)^2}{3-1} = 61$

s=7.81

- Bessel's correction: Higher estimate (higher variance) because as **we are unable to observe the whole population**

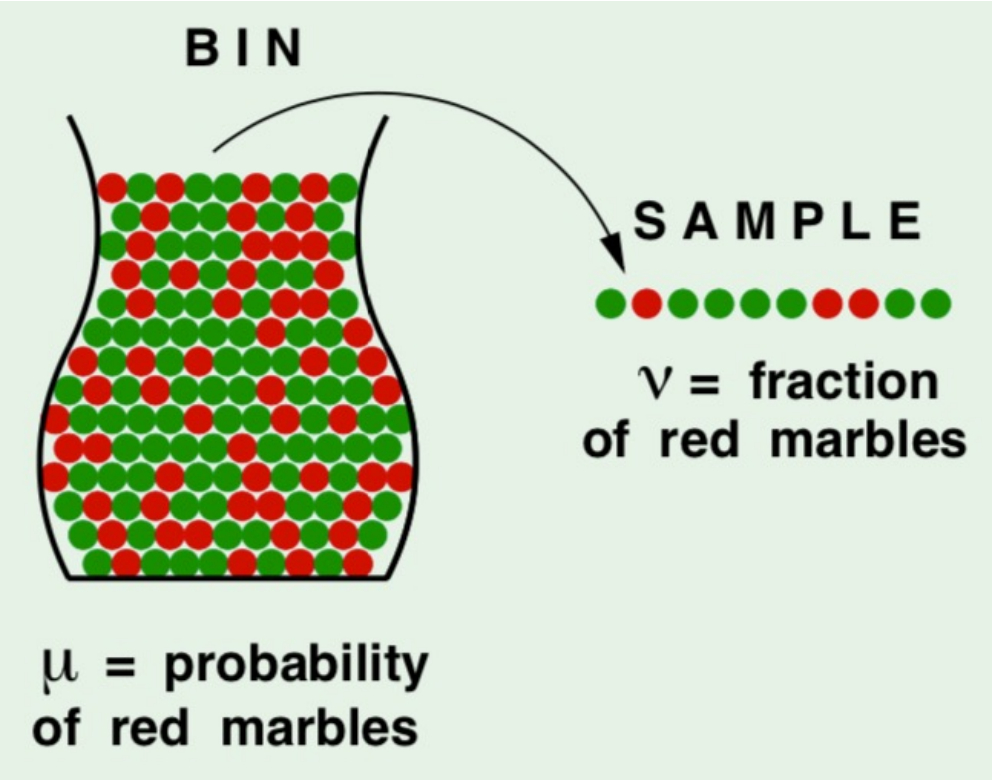
- A population is the entire group that you want to draw conclusions about
- A sample is the specific group that you will collect data from. The size of the sample is always less than the total size of the population



Generalization

- We pick a random sample of *N* independent marbles (with replacement) from this bin, and observe the fraction *v* of red marbles

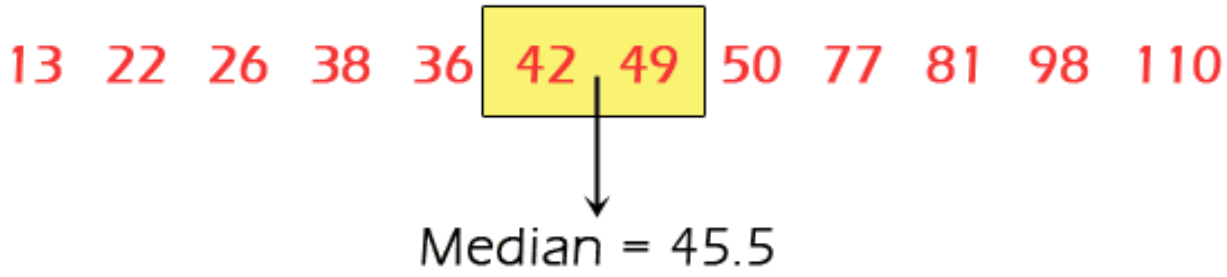
- What does the value of *v* tell us about the value of μ?



Median

- The median is the middle observation in a set of data

- *13, 36, 98, 77, 42, 50, 110, 22, 49, 81, 26, 38*
- Order the data
- *13, 22, 26, 36, 38, 42, 49, 50, 77, 81, 98,110*



Mode

- The mode is the **most common** value in the distribution
- *(1 2 2 3 4 4 4)* the mode is *4*
- If the data are real numbers mode nearly no information
  - Low probability that two or more data will have exactly the same value
- Solution: map into discrete numbers, by rounding or sorting into bins for frequency histograms
- We often speak of a distribution having two or more modes
  - Distributions has two or more values that are common

Visualizing one Variable

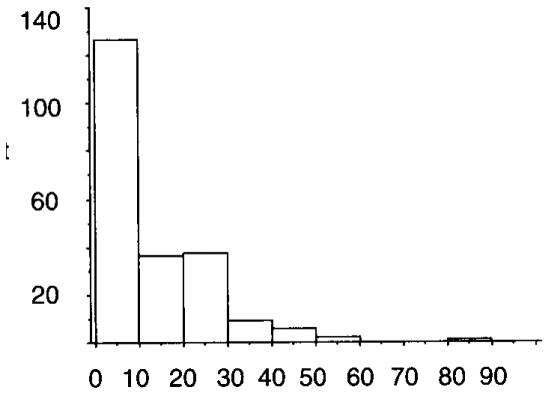
- A good place to start
  - Distribution of individual variables

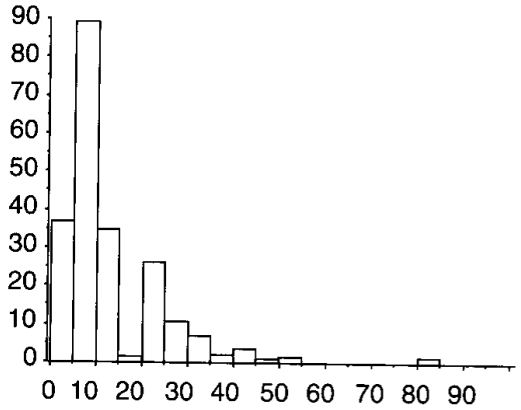
- A common visualization is the frequency histogram
  - It plots the relative frequencies of values in the distribution

To construct a histogram

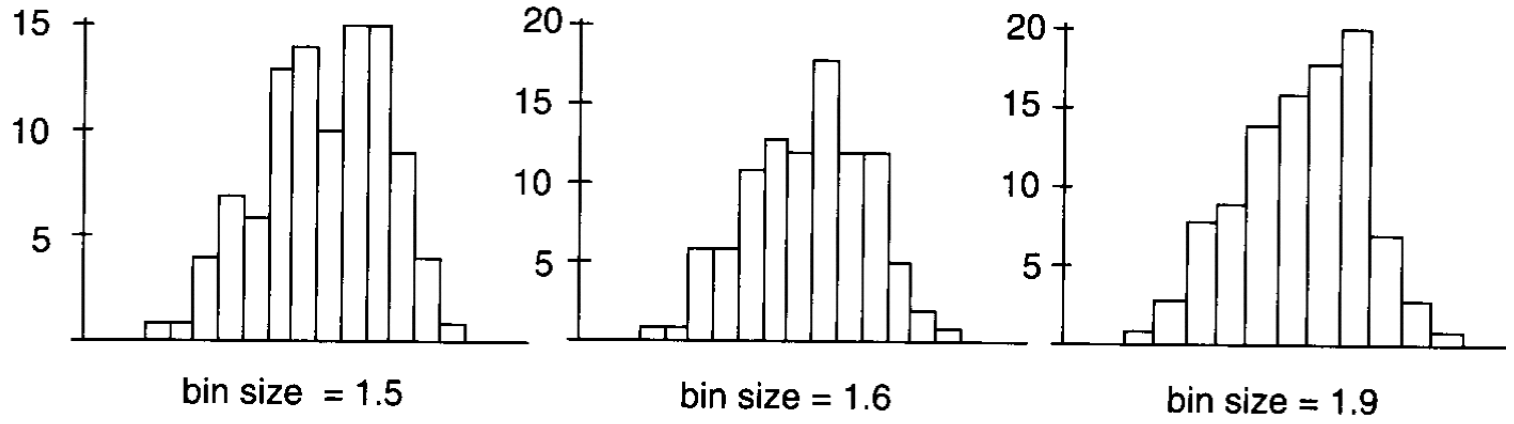
- Divide the range between the highest and lowest values in a distribution into several bins of equal size
- Toss each value in the appropriate bin of equal size
- The height of a rectangle in a frequency histogram represents the number of values in the corresponding bin

- The choice of bin size affects the details we see in the frequency histogram
  - Changing the bin size to a lower number illuminates things that were previously not seen



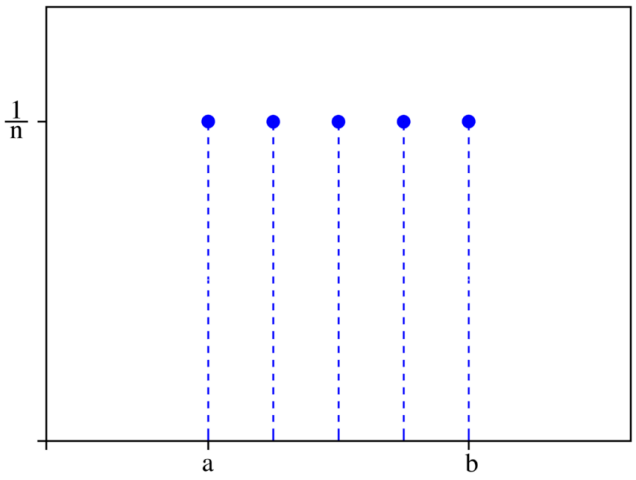


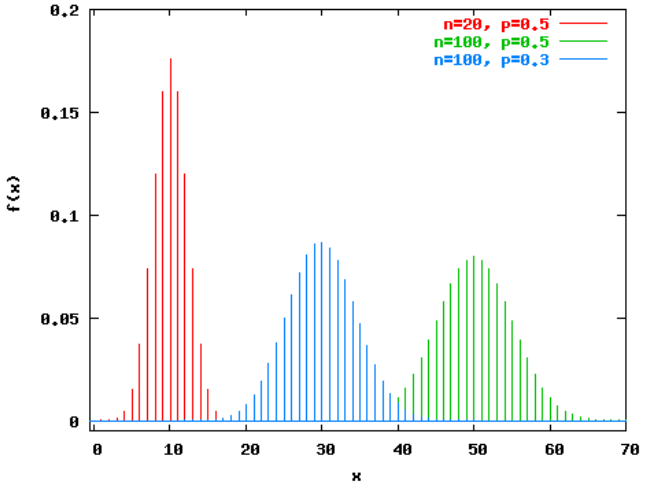
- Bin size affects not only the detail one sees in the histogram but also one‘s perception of the shape of distribution

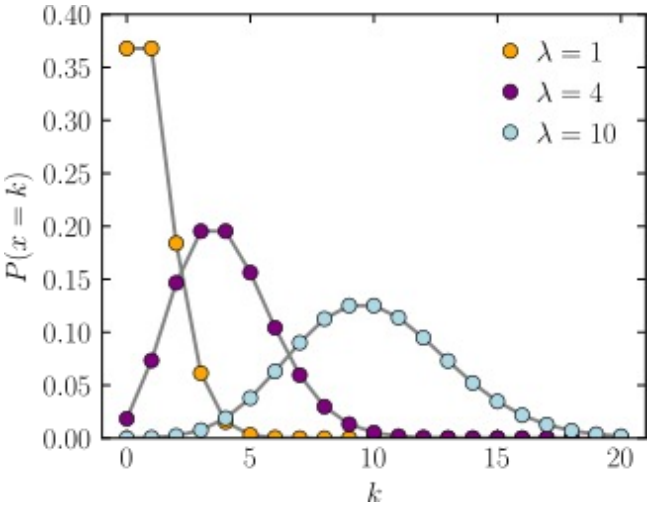


Discrete distributions

Difficult to handle…..







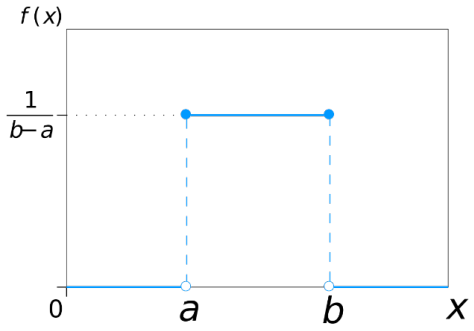
- Uniform

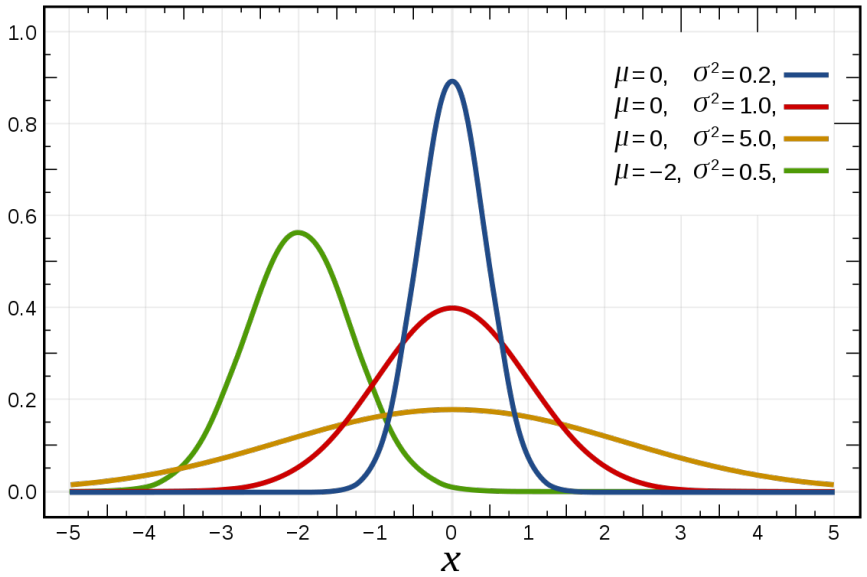
Binomial

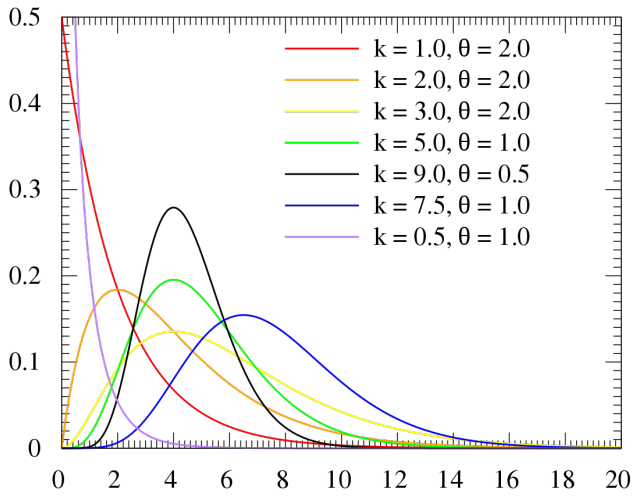
Poisson

Continuous distributions

"Easy" to handle







- Uniform

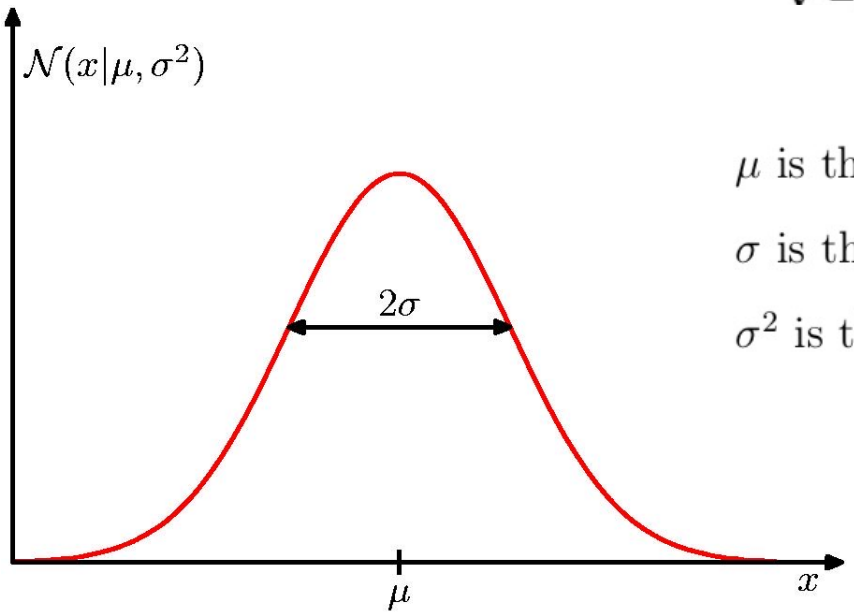
Gaussian

Gamma

Gaussian Distribution

- Gaussian distribution or normal is defined by the probability

$$p(x|\mu,\sigma^2) = \mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\cdot\pi}\cdot\sigma}\cdot\exp\left(-\frac{1}{2\cdot\sigma^2}\cdot(x-\mu)^2\right)$$



μ is the mean

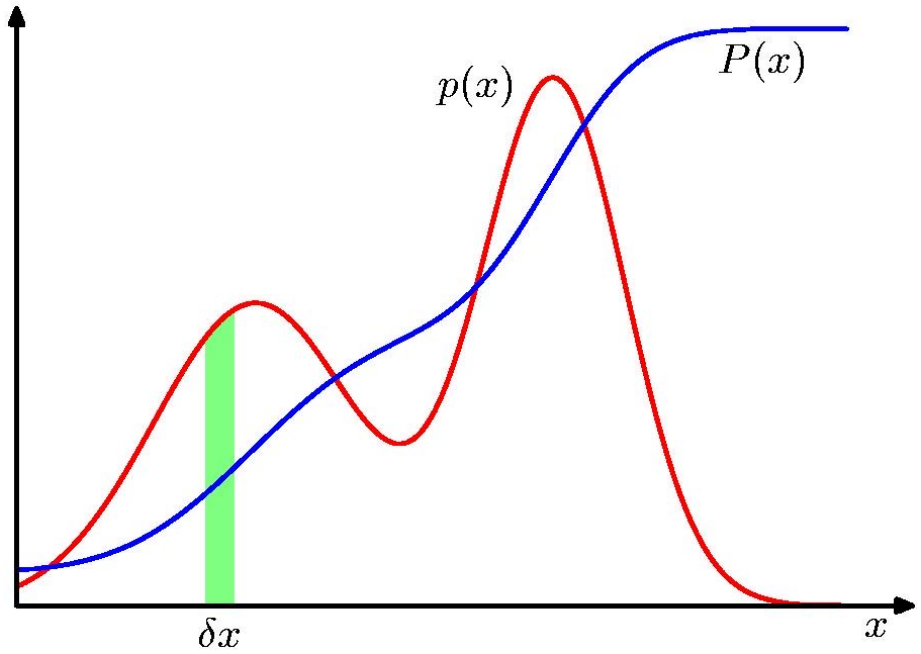
σ is the standard deviation

σ^2 is the variance

$$\int_{-\infty}^{\infty}\mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

Probability Density Function (PDF)



$$p(x \in (a, b)) = \int_a^b p(x)\,\mathrm{d}x$$

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

Cumulative distribution function (CDF)

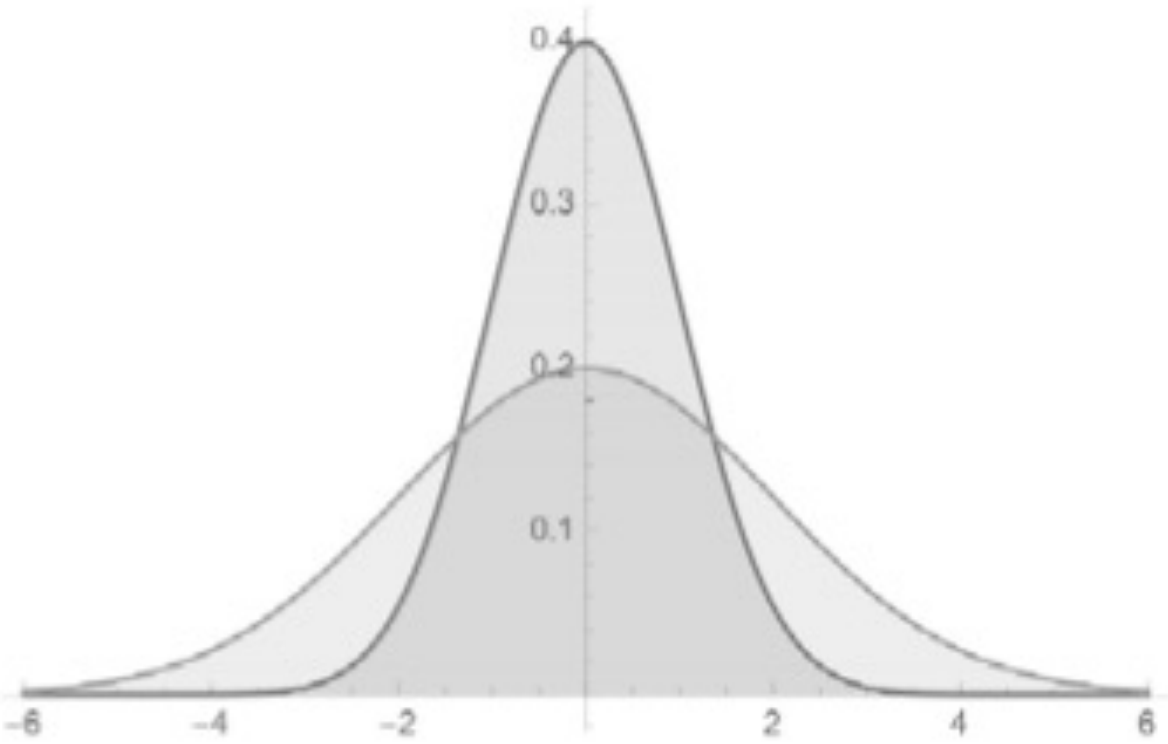
$$p(x) \geqslant 0 \qquad \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Relative Probability

- Gaussian distribution is a type of continuous probability distribution for a real-valued random variable.
- The Gaussian distribution or normal distribution is defined as PDF (Probability Density Function) that reflects the **relative** probability.
- The **PDF may give a value greater than one** (small standard deviation).
- It is the area under the curve that represents the probability. However, the PDF reflects the relative probability.

    - Does a continuous probability distribution exist in the real world?

- Johann Carl Friedrich Gauss (30 April 1777 - 23 February 1855) was a German mathematician, he was the first to suggest the normal distribution law.



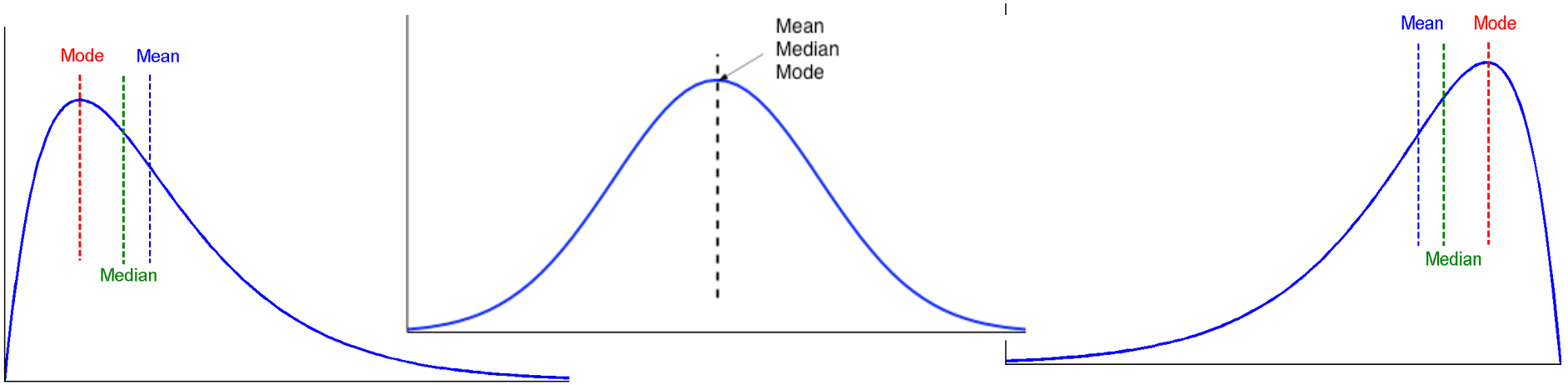
- Two Gaussian (normal) distribution with $\mu = 0$ $\sigma = 1$ and $\mu = 0$ $\sigma = 2$. μ describes the centre of the distribution and σ the width, the bigger σ the more flat the distribution.

Skew

- In a skewed distribution the bulk of the data are at one end of the distribution
  - If the bulk of the distribution is on the right, so the tail is on the left, then the distribution is called left skewed or negatively skewed
  - If the bulk of the distribution is on the left, so the tail is on the right, then the distribution is called right skewed or positively skewed
- The median is to be robust, because its value is not distorted by outliers
  - Outliers: values that are very large or small and very uncommon

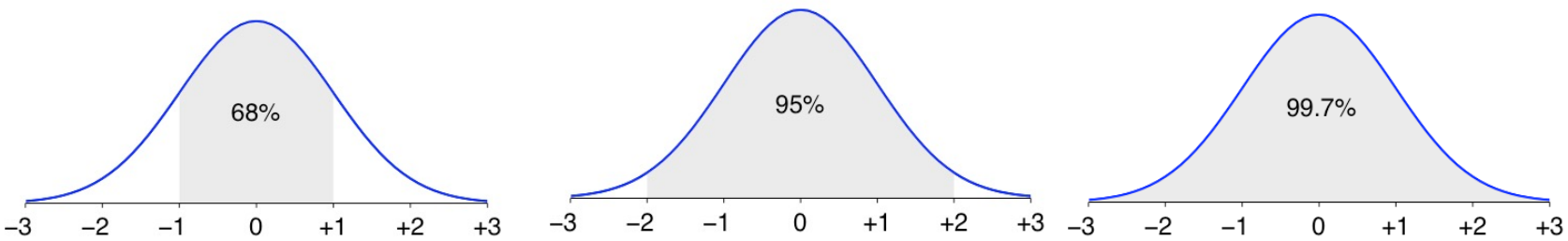
Univariate data statistics: skew

- In a skewed distribution the bulk of the data are at one end of the distribution
  - If the bulk of the distribution is on the right (tail is on the left): **left skewed** or negatively skewed distribution
  - If the bulk of the distribution is on the left (tail is on the right): **right skewed** or positively skewed distribution
- **Symmetric** distributions are not skewed



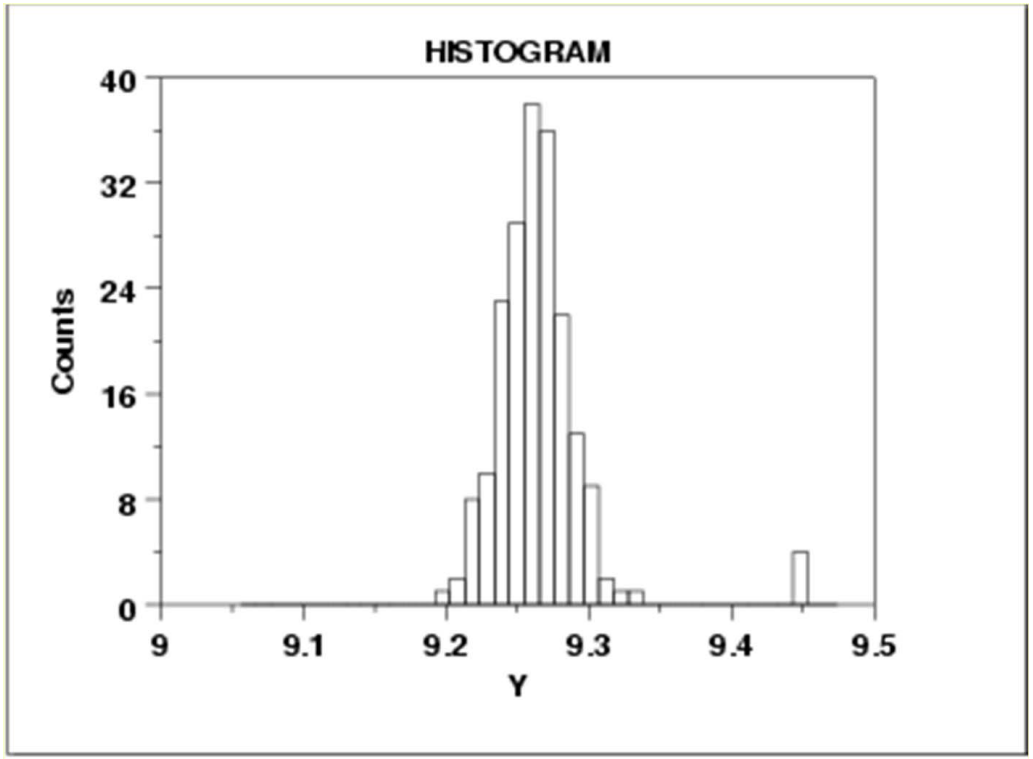
Properties of Normal distribution

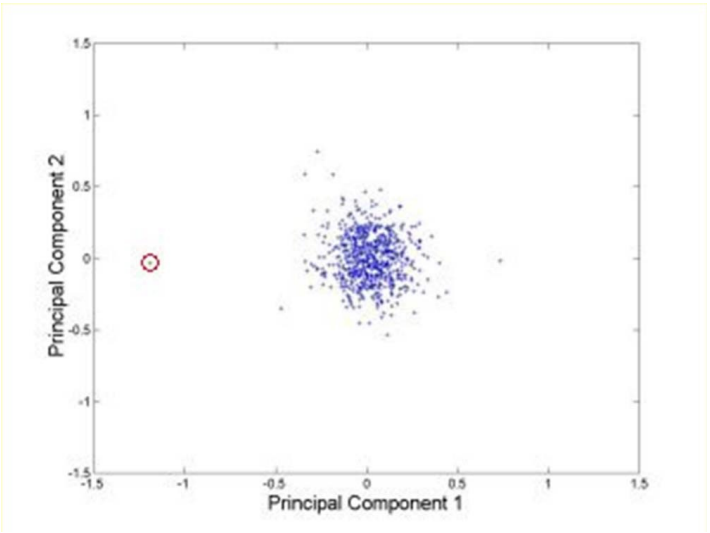
- Many real-world variables are well-approximated to a Gaussian curve
- How to check if one variable satisfies the Gaussian assumption?
  - Statistical test, Z-test, t-test
- Interesting properties of the Normal curve:
  - from μ–σ to μ+σ: contains about 68% of the measurements (μ: mean, σ: standard deviation)
  - from μ–2σ to μ+2σ: contains about 95% of it
  - from μ–3σ to μ+3σ: contains about 99.7% of it



Outliers

- **Outlier** values = uncommon values
  - unexpected measurements against a variable distribution





Outliers

- **Outlier** values = uncommon values
  - unexpected measurements against a variable distribution
- *One cannot do much about outliers expect find them, and sometimes, remove them*
- Mean and the variance are based on averages, hence sensitive to outliers
- Outliers can cause strong effects that can wreck our interpretation of data
  - for example, the presence of a single outlier can render some statistical comparisons insignificant
- Removing requires judgment and depend on one‘s purpose

Trimmed mean

- Another robust alternative to the mean is the trimmed mean
  - Lop off a fraction of the upper and lower ends of the distribution, and take the mean of the rest
    - *0,0,1,2,5,8,12,17,18,18,19,19,20,26,86,116*
  - Lop off two smallest and two larges values and take the mean of the rest
    - Trimmed mean is *13.75*
    - The arithmetic mean *22.75*

Interquartile Range

- Interquartile range is found by dividing a **sorted** distribution into four containing parts, each containing the same number
- Each part is called quartile
- The difference between the highest value in the third quartile and the lowest value in the second quartile is the interquartile range

Interquartile Range (Even number)

- *48, 52, **57**, 64, 72, 76, 77, **81**, 85, 88*
  - *Q1=57, Q3=81*
  - *IQR=Q3-Q1=81-57=24*
  - *Q2=Median=(72+76)/2=74*

Interquartile Range (Odd Number)

- *48, 52, **57**, 61, 64, 72, 76, 77, **81**, 85, 88*
  - *Q1=57, Q3=81*
  - *IQR=Q3-Q1=81-57=24*
  - *Q2=Median=72*

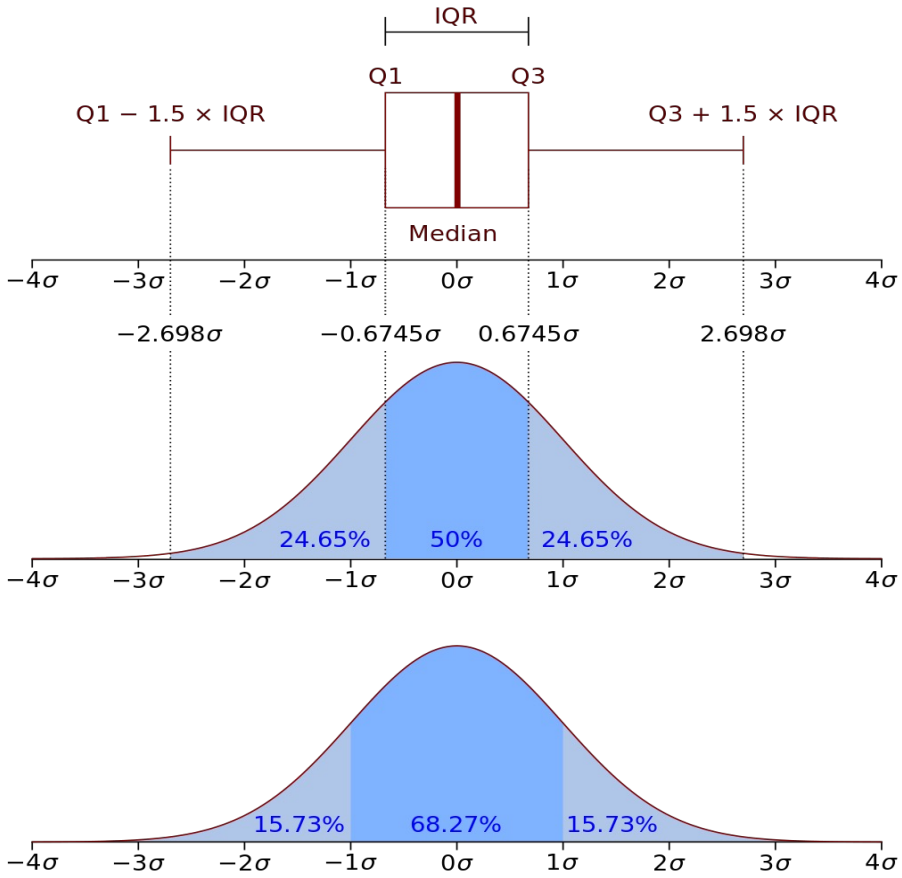
Exclusive Method

Interquartile Range (Odd Number)

- *48, 52, **57, 61**, 64, 72, 72; 76, **77, 81**, 85, 88*
  - *Q1=(57+61)/2=59, Q3=(77+81)/2=79*
  - *IQR=Q3-Q1=79-59=20*
  - *Q2=Median=72*

Inclusive Method

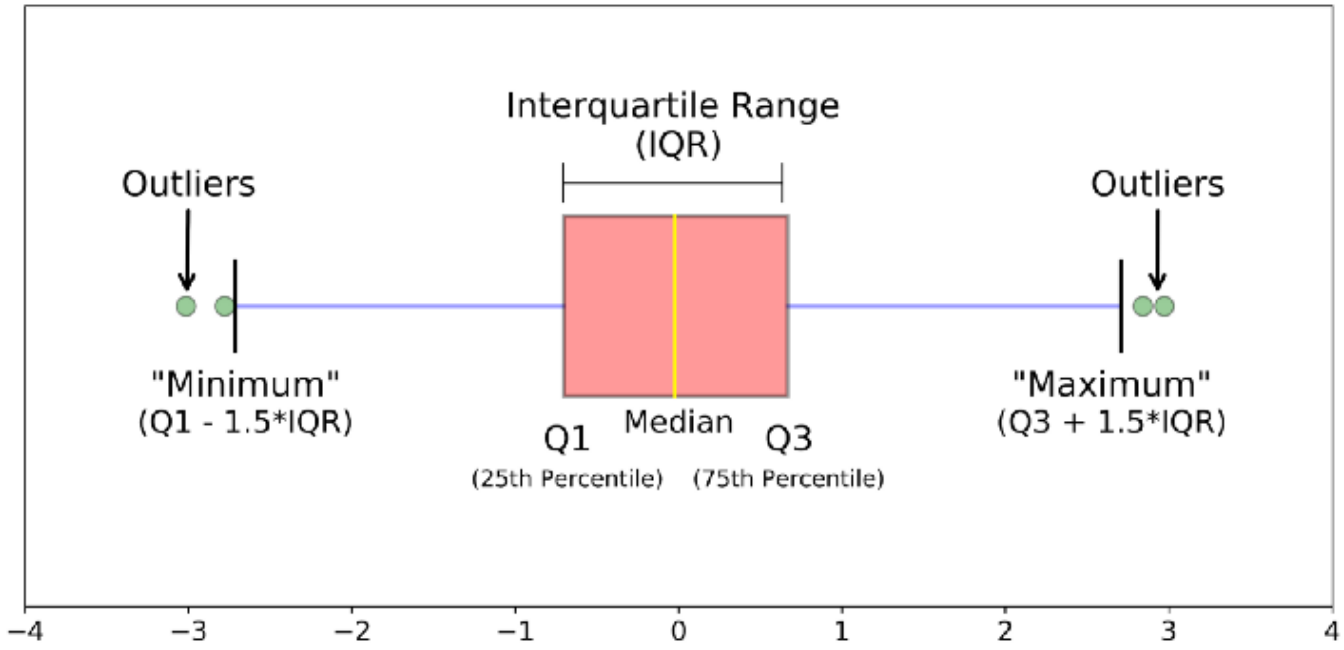
Boxplot



[Q1-1.5*IQR, Q3+1.5*IQR]

Boxplot

- Outliers outside the interval [Q1-1.5*IQR, Q3+1.5*IQR]

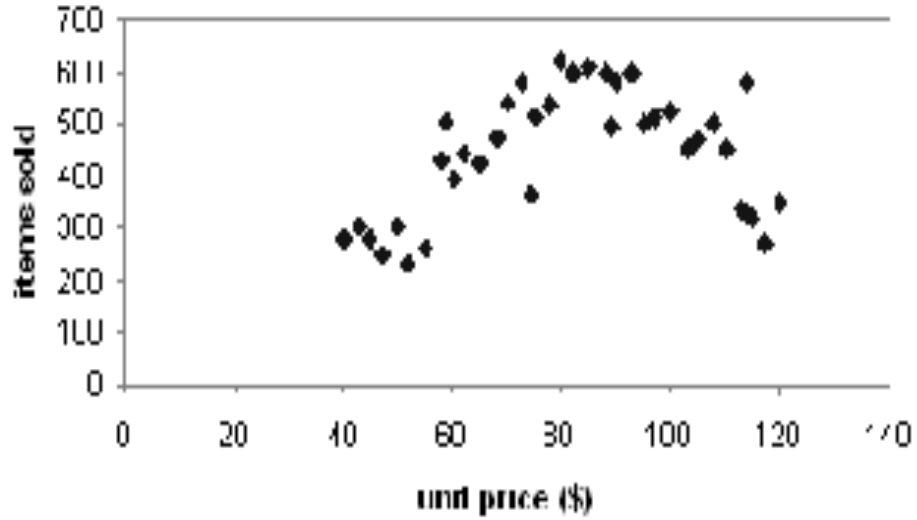


Bivariate data statistics

- Considering **pairwise** input variables:
  - check whether the two distributions are correlated
  - if highly correlated, variables may be redundant
  - select the one with the highest variability

Scatter plot

- Provides a first look at bivariate data to see clusters of points, outliers, etc
- Each pair of values is treated as a pair of coordinates and plotted as points in the plane



Covariance

- Covariance is a measure of how much two random variables change together in a **linear way**
- *In a linear relationship, either the high values of one variable are paired with the high values of another variable or the high values of one variable are paired with the low values of another variable*

$$cov(X,Y) = \frac{\sum_{k=1}^{N}(x_k - \overline{x}) \cdot (y_k - \overline{y})}{N-1}$$

- The sample covariance has N − 1 in the denominator rather than N due to Bessel's correction which allows us to avoid underestimating the real covariance.

Example

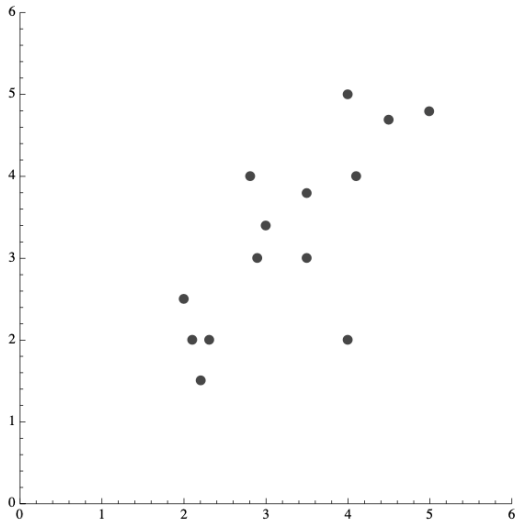
For a two variable *(X,Y)* dataset,

{(2.1, 2), (2.3, 2), (2.9, 3), (4.1, 4), (5, 4.8), (2, 2.5), (2.2, 1.5), (4, 5), (4, 2), (2.8, 4), (3, 3.4), (3.5, 3.8), (4.5, 4.7), (3.5, 3)}

the sample covariance of the data set is *0.82456*

Ordering the list by *X*, we notice that the ascending X values are matched by ascending *Y* values

$$cov(X,Y) = \frac{\sum_{k=1}^{N}(x_k - \overline{x}) \cdot (y_k - \overline{y})}{N-1}$$



Scatter plot

Covariance for Population

- If we have the whole population, the covariance is

$$cov(X,Y) = \frac{\sum_{k=1}^{N}(x_k - \overline{x}) \cdot (y_k - \overline{y})}{N}$$

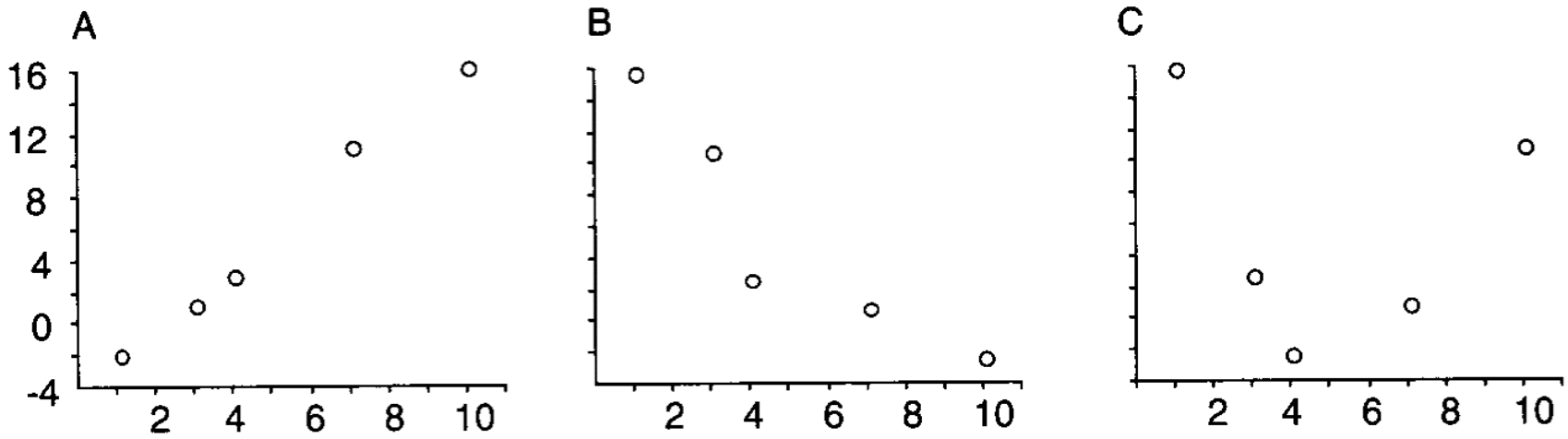
- In Machine Learning, we never have access to the whole population, so, we will use the sample covariance without further distinction

Correlation (Pearson correlation)

- Correlation coefficient
  - also called Pearson's product moment coefficient

$$r_{xy} = \frac{\sum_{k=1}^{N}(x_k - \bar{x})(y_k - \bar{y})}{(n-1)s_x s_y} = \frac{cov(X,Y)}{s_x s_y}$$

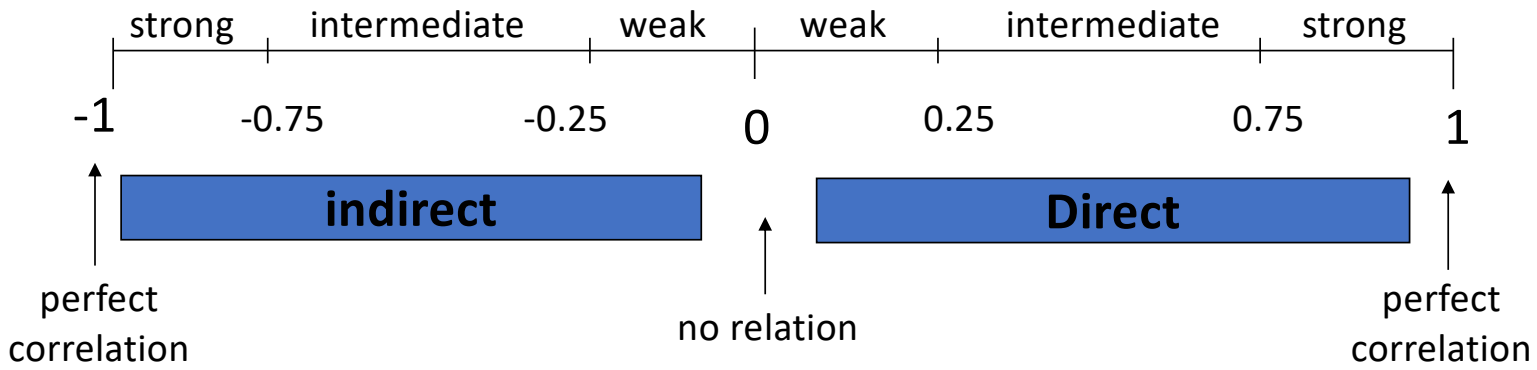
  - If $r_{X,Y} > 0$, *X* and *Y* are positively correlated (*X*'s values increase as *Y*'s).
    - *The higher, the stronger correlation.*
  - $r_{X,Y} < 0$: negatively correlated
  - $r_{X,Y} = 0$: independent;



- *A (with $r_{X,Y} > 0$) is* positively correlated
  - (*X*'s values increase as *Y*'s).
- *B (with $r_{X,Y} < 0$) is* negatively correlated
  - (*X*'s values increase, *Y*'s values decrease).
- *C (with $r_{X,Y} = 0$)* independent (or nonlinear);

Correlation...

- Relationship between two quantitative attributes
  - correlation: degree to which two attributes are related (in [-1,1])
    - the *sign*: nature of association (>0 direct; <0 inverse)
    - the absolute *value* of r: strength of association
    - unable to infer causal relationships



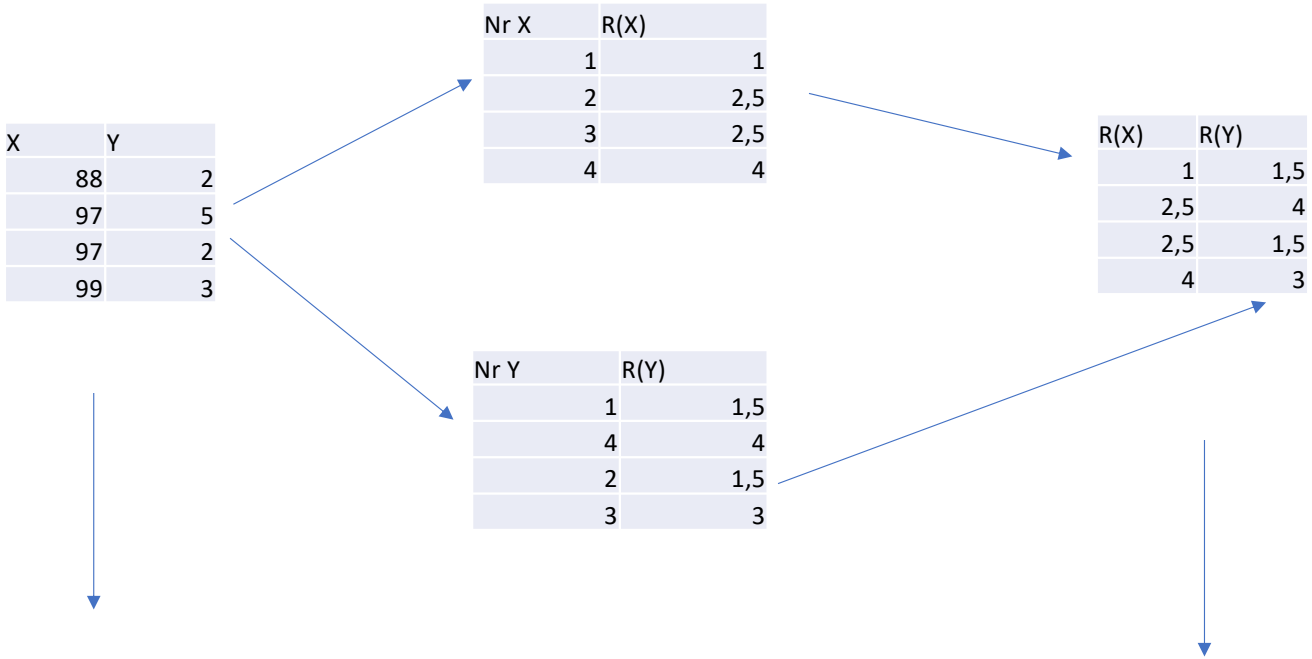
Sperman Rank (correlation)

- While Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships
- Rank the variables *X, Y* and use Correlation (Pearson correlation)

$$r_s = \rho_{R(X)R(Y)} = \frac{\sum_{k=1}^{N}(R(x_k) - \overline{R(x)})(R(y_k) - R(\overline{y}))}{(n-1)R(s_x)R(s_y)} = \frac{cov(R(X),R(Y))}{R(s_x)R(s_y)}$$

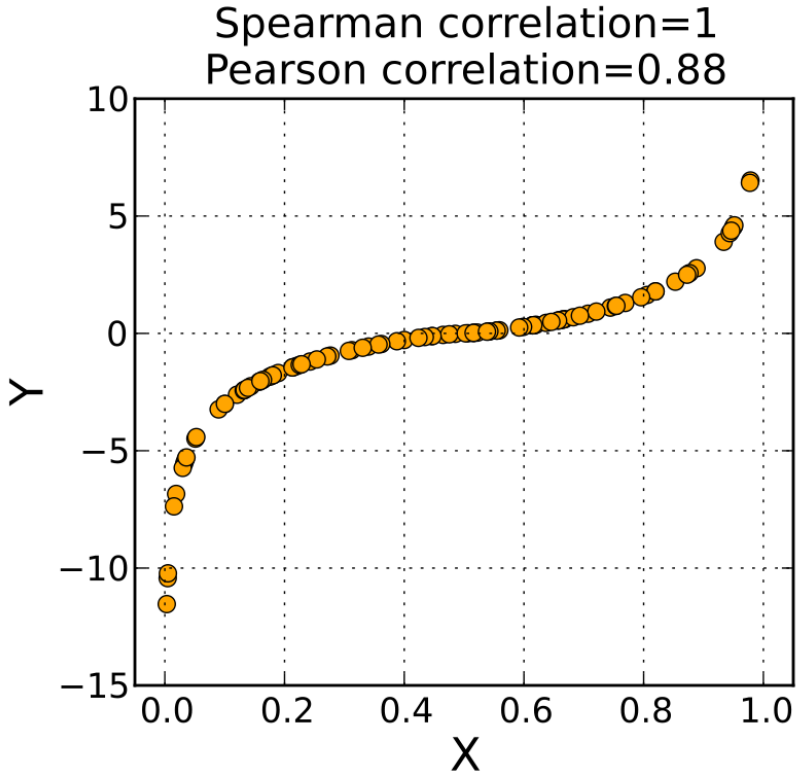
- Works with rankings instead of absolute values

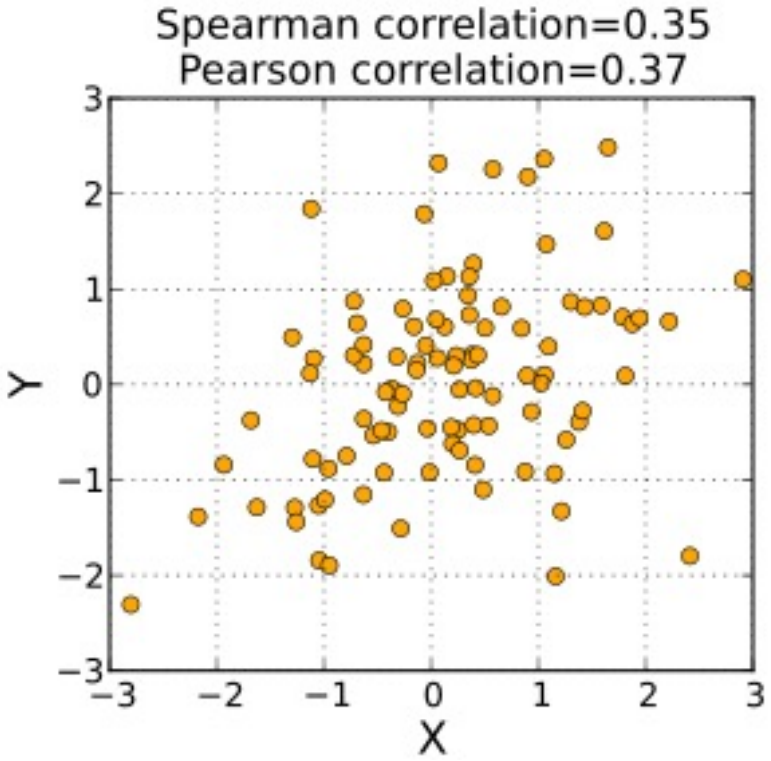
Rank *X* and *Y*

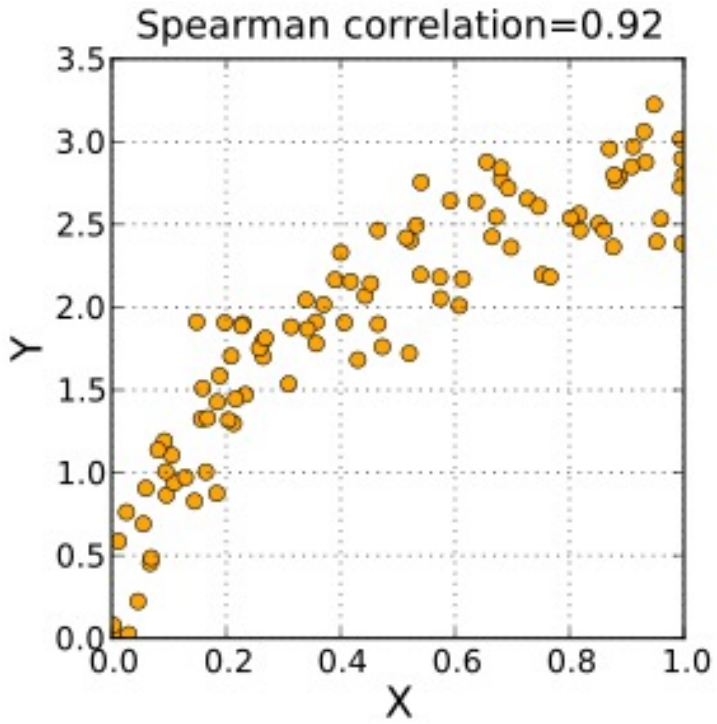


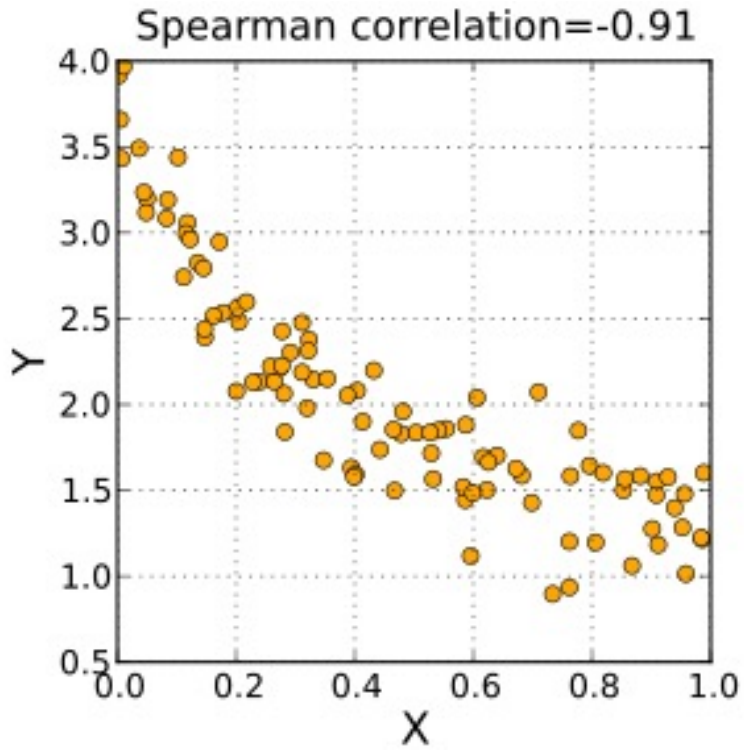
$$r_{xy}=\frac{\sum_{k=1}^{N}(x_k-\bar{x})(y_k-\bar{y})}{(n-1)s_x s_y}=\frac{cov(X,Y)}{s_x s_y}$$

$$r_s=\rho_{R(X)R(Y)}=\frac{\sum_{k=1}^{N}(R(x_k)-\overline{R(x)})(R(y_k)-R(\overline{y}))}{(n-1)R(s_x)R(s_y)}=\frac{cov(R(X),R(Y))}{R(s_x)R(s_y)}$$


Spearman correlation=1
Pearson correlation=0.88
10
5
0
−5
−10
−15
Y
0.0
0.2
0.4
0.6
0.8
1.0
X

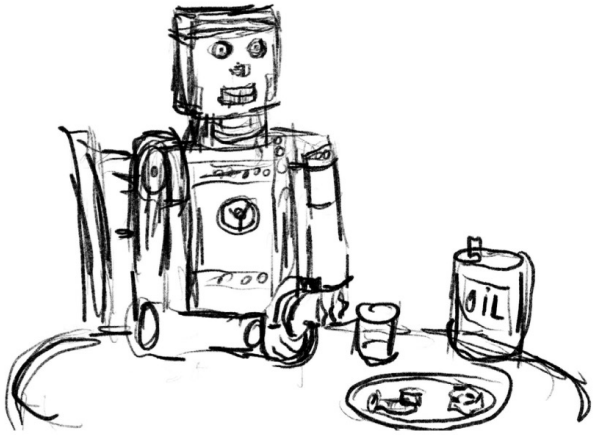

Spearman correlation=0.35
Pearson correlation=0.37
3
2
1
0
−1
−2
−3
Y
−3
−2
−1
0
1
2
3
X


Spearman correlation=0.92
Y
X
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0


Spearman correlation=-0.91
Y
X
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0.0
0.2
0.4
0.6
0.8
1.0

Before *Machine* Learning…..

- ~~Step 1: Select and **Understand your Data**~~
- Step 2: Preprocess/Prepare your Data
- Step 3: Transform Data



2. Why Prepare Data?

- Some data preparation is needed for machine learning
- Data in the real world is dirty
  - incomplete: lacking attribute values, lacking certain attributes of interest
  - noisy: containing errors or outliers
  - inconsistent: containing discrepancies in codes or names
- Data discretization
  - Part of data reduction but with particular importance, especially for numerical data
- Data cleaning
  - Fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies

Variables

- **Categorical** (or qualitative) variables
  - values are categories, e.g. eye color
  - can either be **nominal**/symbolic or **ordinal** (e.g. low, average, high)
    - Nominal: Names of people
  - **binary** variables are variables with two categories (whether nominal or ordinal)
  - variable **cardinality** = number of categories

- **Numerical** (or quantitative) variables
  - values are quantities
  - can be either be **discrete** (e.g. integers) or **continuous** (e.g. real values)

- Convert ordinal fields to numeric to be able to use “>” and “<“ comparisons on such fields.
- For example American grades:

  - A - into the numerical value 4.0
  - A- into the numerical value 3.7
  - B+ into the numerical value 3.3
  - B into the numerical value 3.0

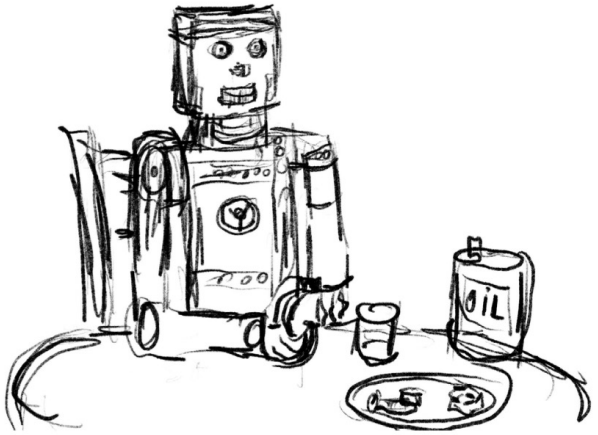
Rank *X* and *Y*

education level (X)	**income** (Y)	R(X)	R(Y)
Preparatory	**25**	5	3
Primary	**10**	6	5.5
University	**8**	1.5	7
Secondary	**10**	3.5	5.5
Secondary	**15**	3.5	4
Illiterate	**50**	7	2
University	**60**	1.5	1

- r_s denotes the magnitude of association
- $r_s = -0.17$
  - No strong correlation

Before *Machine* Learning…..

- ~~Step 1: Select and **Understand your Data**~~
- ~~Step 2: Preprocess/Prepare your Data~~
- Step 3: Transform Data
  - Normalization and Data reduction
  - Obtains reduced representation that produces the same or similar analytical



Discretization

- Divide the range of a continuous attribute into intervals
- Some methods require discrete values, e.g. **decision trees**, most versions of *Naïve Bayes*
- Reduce data size by discretization
  - Prepare for further analysis
  - Discretization is very useful for generating a summary of data
  - Also called “binning”

Variables

- [**discretization**] numeric variables can be discretized into ordinal variables
  - e.g. age categories of 0-10, 11-20, 21-30, 31-40…
- [**normalization**] numeric variables can be normalized
  - comparability between variables with different domains Type equation here.
- [**aggregation**] categoric variables with high cardinality can be aggregated
  - 100 colors can be aggregated into coarser categories in accordance with hue
- [**imputation**] missing values can occur
  - unobserved, error or noisy measurements
  - missings can be imputed using variable expectations

Normalization

- What is the scale?
- For distance-based methods, normalization helps to prevent that attributes with large ranges out-weight attributes with small ranges

  - min-max normalization
  - z-score normalization
  - normalization by decimal scaling

Data Transformation: Normalization

- Min-max normalization: to [new_min_A, new_max_A]

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

  - Ex. Let income range \$12,000 to \$98,000 normalized to [0.0, 1.0]. Then \$73,000 is mapped to $\frac{73{,}600 - 12{,}000}{98{,}000 - 12{,}000}(1.0 - 0) + 0 = 0.716$

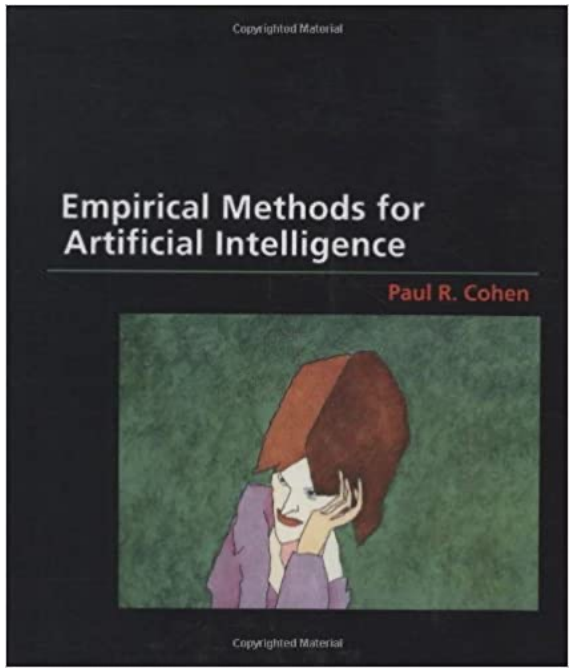
- Z-score normalization (μ : mean, σ : standard deviation):

$$v' = \frac{v - \mu_A}{\sigma_A}$$

  - Ex. Let μ = 54,000, σ = 16,000. Then $\frac{73{,}600 - 54{,}000}{16{,}000} = 1.225$

- Normalization by decimal scaling

Literature



★★★★★ **Just what I need.**

Reviewed in the United States on January 16, 2022

Verified Purchase

This is what I need for my data analysis. I have read a few chapters and I am loving this. I recommend this for anyone doing serious research.

- **Paul R. Cohen, Empirical Methods for Artificial Intelligence (Bradford Books),** Reprint edition (May 26, 2017)