

Data analysis and visualization (DAV)

Lecture 09

Łukasz P. Kozłowski

Warsaw, 2025

Data analysis and visualization (DAV)

Lecture 09
Statistics & machine learning
Part 1

Łukasz P. Kozłowski

Warsaw, 2025

Descriptive Statistics vs. Inferential Statistics

Summary statistics – descriptive statistics used to summarize a set of observations, in order to obtain the largest amount of information as simply as possible

Normally we focus on four measurements:

- **location** (arithmetic mean, median, mode, interquartile mean)

Descriptive Statistics vs. Inferential Statistics

Summary statistics – descriptive statistics used to summarize a set of observations, in order to obtain the largest amount of information as simply as possible

Normally we focus on four measurements:

- **location** (arithmetic mean, median, mode, interquartile mean)
- **spread** (standard deviation, variance, range, interquartile range, absolute deviation, mean absolute difference and the distance standard deviation)

Descriptive Statistics vs. Inferential Statistics

Summary statistics – descriptive statistics used to summarize a set of observations, in order to obtain the largest amount of information as simply as possible

Normally we focus on four measurements:

- **location** (arithmetic mean, median, mode, interquartile mean)
- **spread** (standard deviation, variance, range, interquartile range, absolute deviation, mean absolute difference and the distance standard deviation)
- **shape** (skewness, kurtosis)

Descriptive Statistics vs. Inferential Statistics

Summary statistics – descriptive statistics used to summarize a set of observations, in order to obtain the largest amount of information as simply as possible

Normally we focus on four measurements:

- **location** (arithmetic mean, median, mode, interquartile mean)
- **spread** (standard deviation, variance, range, interquartile range, absolute deviation, mean absolute difference and the distance standard deviation)
- **shape** (skewness, kurtosis)
- **dependence** (Pearson correlation coefficient)

Descriptive Statistics vs. Inferential Statistics

Location or center of the distribution (central tendency) is a central or typical value for a probability distribution

arithmetic mean

$$A = \frac{1}{n} \sum_{i=1}^n a_i = \frac{a_1 + a_2 + \cdots + a_n}{n}$$

$$\frac{2500 + 2700 + 2400 + 2300 + 2550 + 2650 + 2750 + 2450 + 2600 + 2400}{10} = 2530.$$

median

$$\text{median}(a) = \frac{a_{\lfloor \frac{l+1}{2} \rfloor} + a_{\lceil \frac{l+1}{2} \rceil}}{2}$$

1, 3, 3, **6**, 7, 8, 9

Median = **6**

1, 2, 3, **4**, **5**, 6, 8, 9

Median = $(4 + 5) \div 2$
= **4.5**

Descriptive Statistics vs. Inferential Statistics

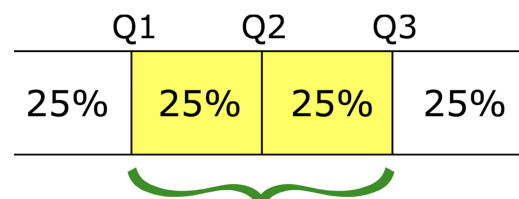
Location or center of the distribution (central tendency) is a central or typical value for a probability distribution

Mode - the value that appears most often

1, **2, 2**, 3, 4, 7, 9

Interquartile mean (IQM, midmean)

$$x_{\text{IQM}} = \frac{2}{n} \sum_{i=\frac{n}{4}+1}^{\frac{3n}{4}} x_i$$



5, 8, 4, 38, 8, 6, 9, 7, 7, 3, 1, 6 dataset

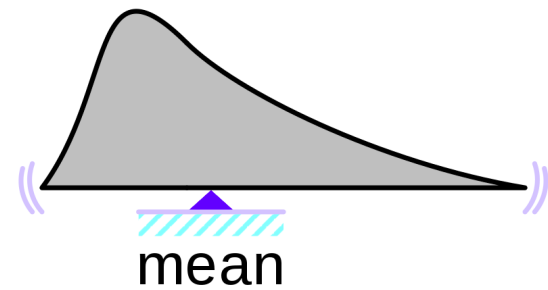
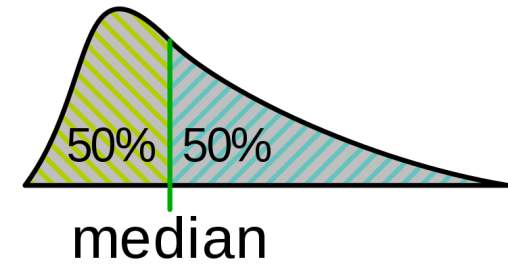
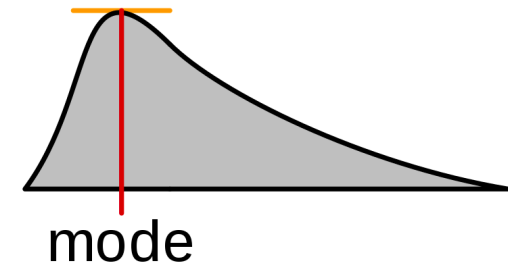
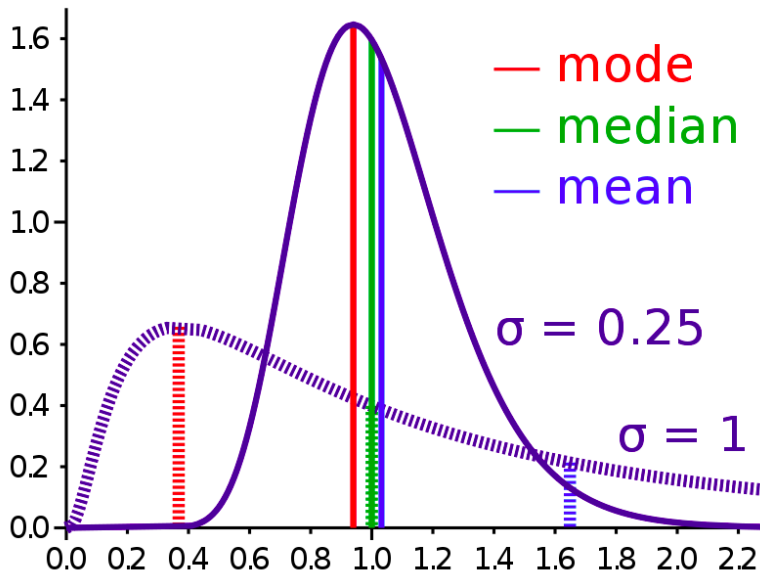
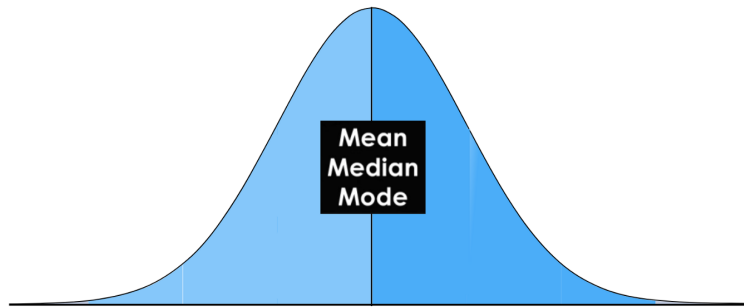
1, 3, 4, 5, 6, 6, 7, 7, 8, 8, 9, 38 sort

~~1, 3, 4~~, 5, 6, 6, 7, 7, 8, ~~8, 9, 38~~ discard

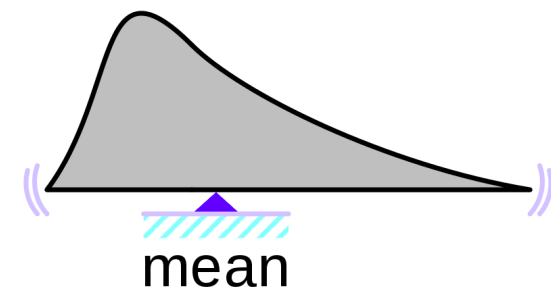
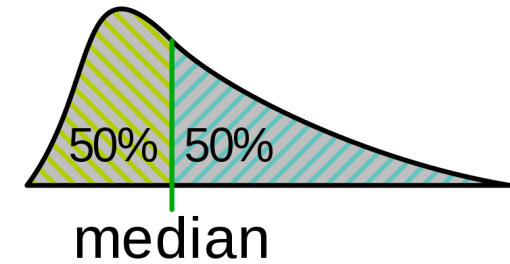
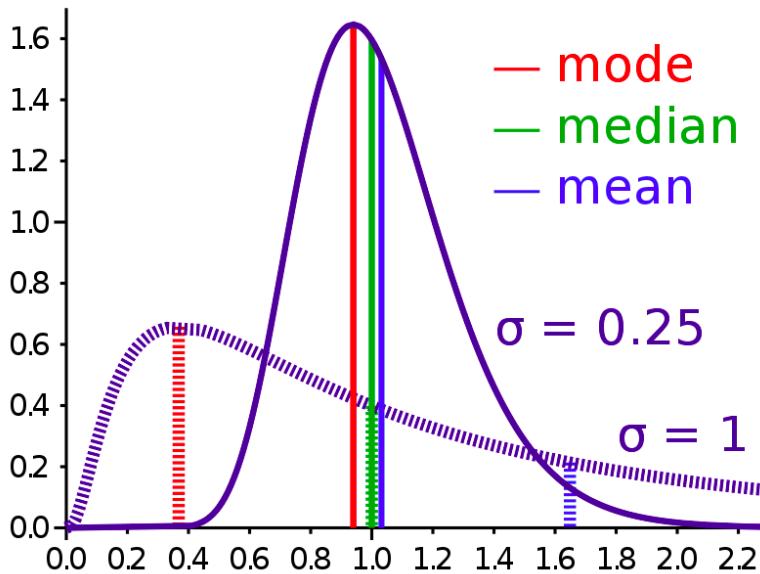
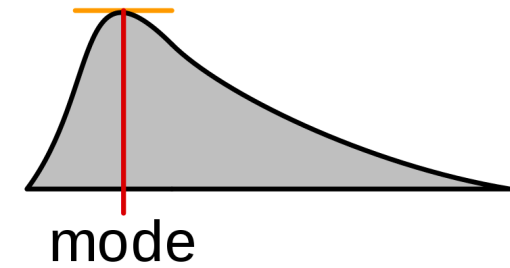
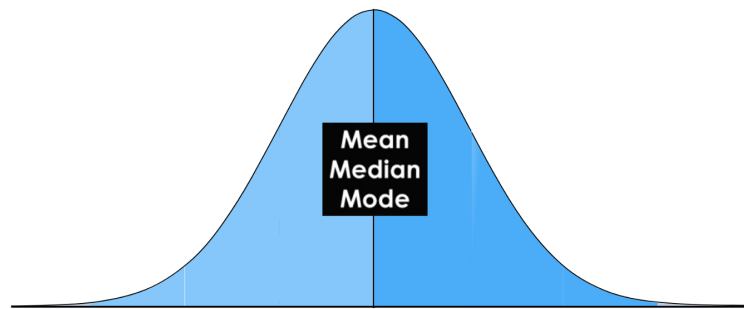
$$\text{IQM} = (5 + 6 + 6 + 7 + 7 + 8) / 6 = 6.5$$

Also: Geometric mean, Harmonic mean, Weighted arithmetic mean, Truncated mean, Midrange, Midhinge, Trimean, Winsorized mean, Geometric median, Quadratic mean, Simplicial depth, Tukey median

Location



Location



Comparison of common averages of values { 1, 2, 2, 3, 4, 7, 9 }

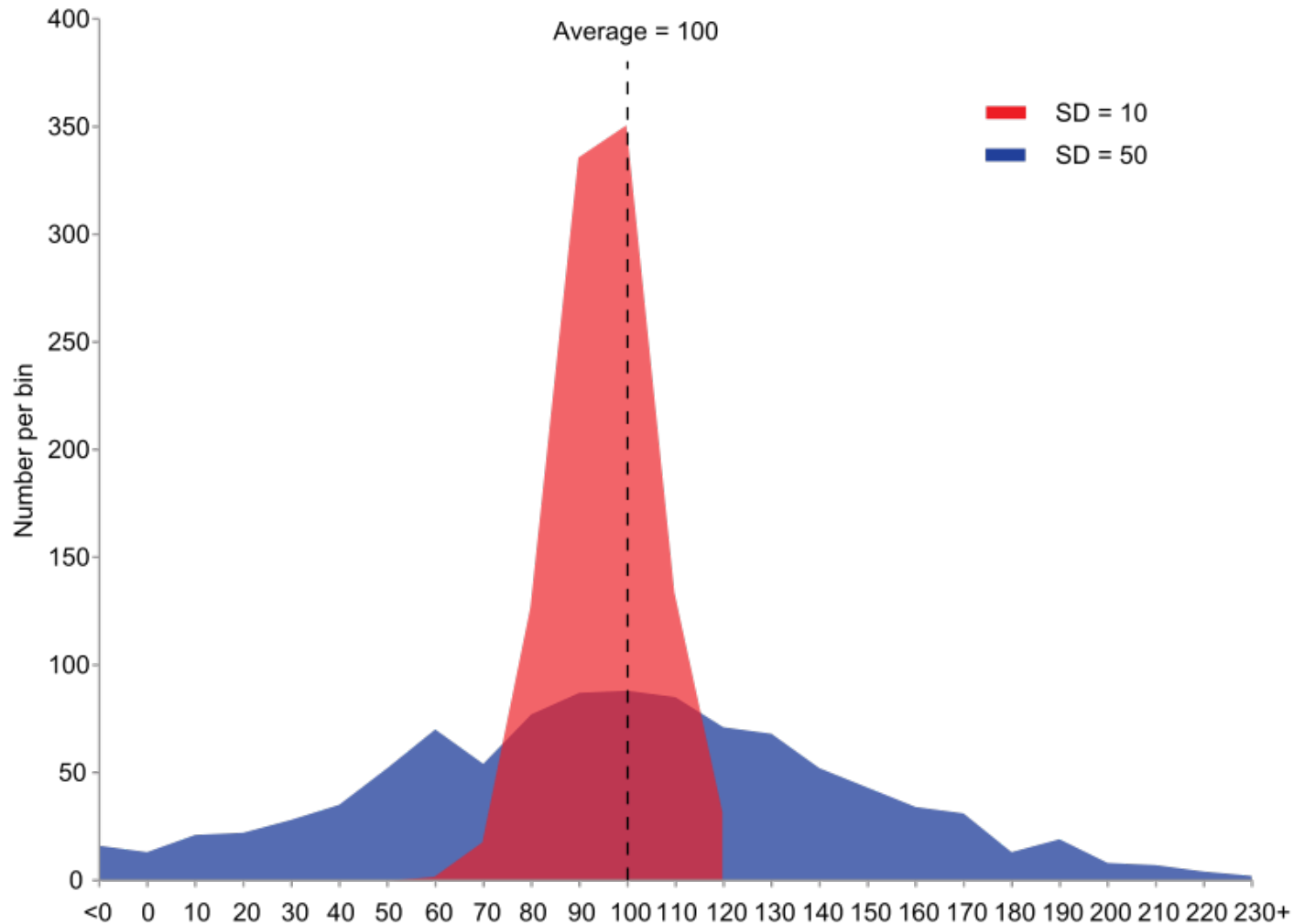
Type	Description	Example	Result
Arithmetic mean	Sum of values of a data set divided by number of values: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	$(1+2+2+3+4+7+9) / 7$	4
Median	Middle value separating the greater and lesser halves of a data set	1, 2, 2, 3 , 4, 7, 9	3
Mode	Most frequent value in a data set	1, 2 , 2 , 3, 4, 7, 9	2

Spread (also called variability, scatter, or dispersion)

Examples of dispersion measures include:

- Standard deviation
- Interquartile range (IQR)
- Range
- Mean absolute difference (also known as Gini mean absolute difference)
- Median absolute deviation (MAD)
- Average absolute deviation (or simply called average deviation)
- Distance standard deviation

Spread (also called variability, scatter, or dispersion)



Two populations with the same mean but different dispersion

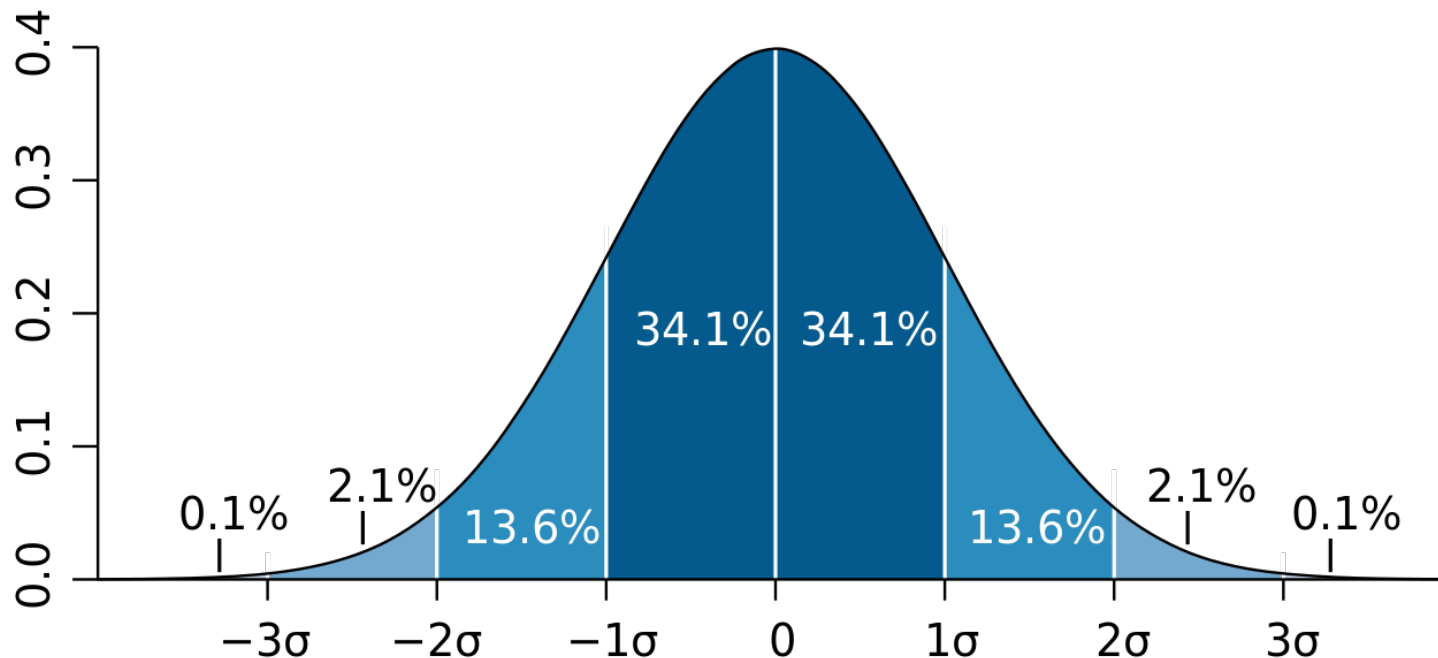
The blue population is much more dispersed than the red population

Spread (also called variability, scatter, or dispersion)

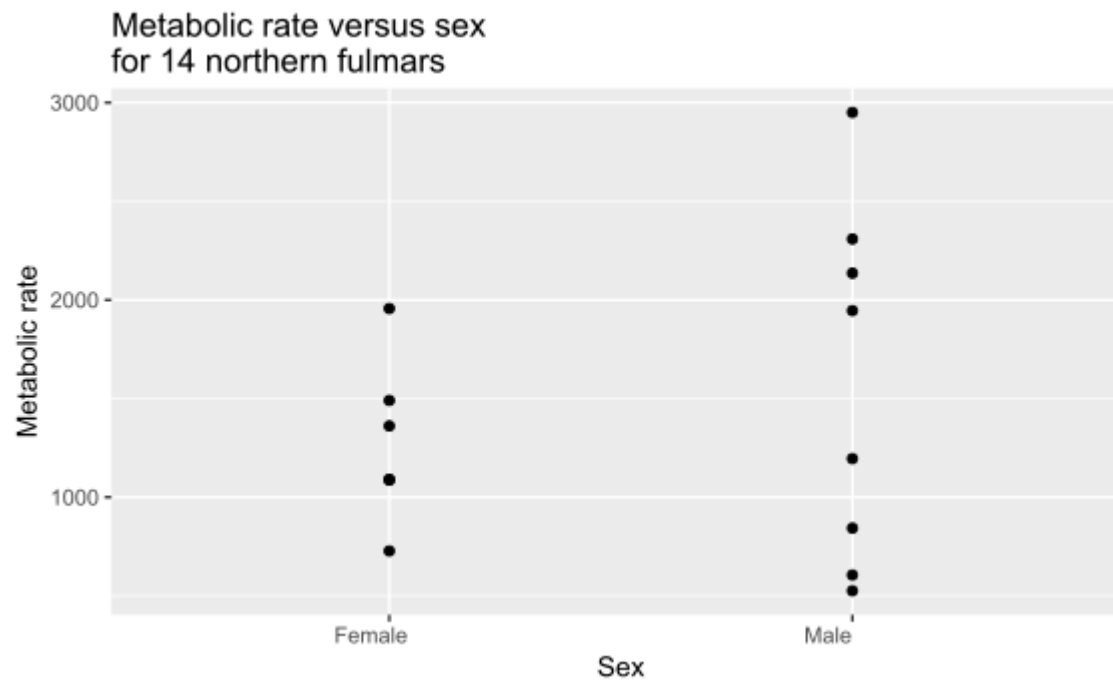
Standard deviation (SD or the lower case Greek letter sigma σ)

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

where $\{x_1, x_2, \dots, x_N\}$ are the observed values of the sample items, \bar{x} is the mean value of these observations, and N is the number of observations in the sample

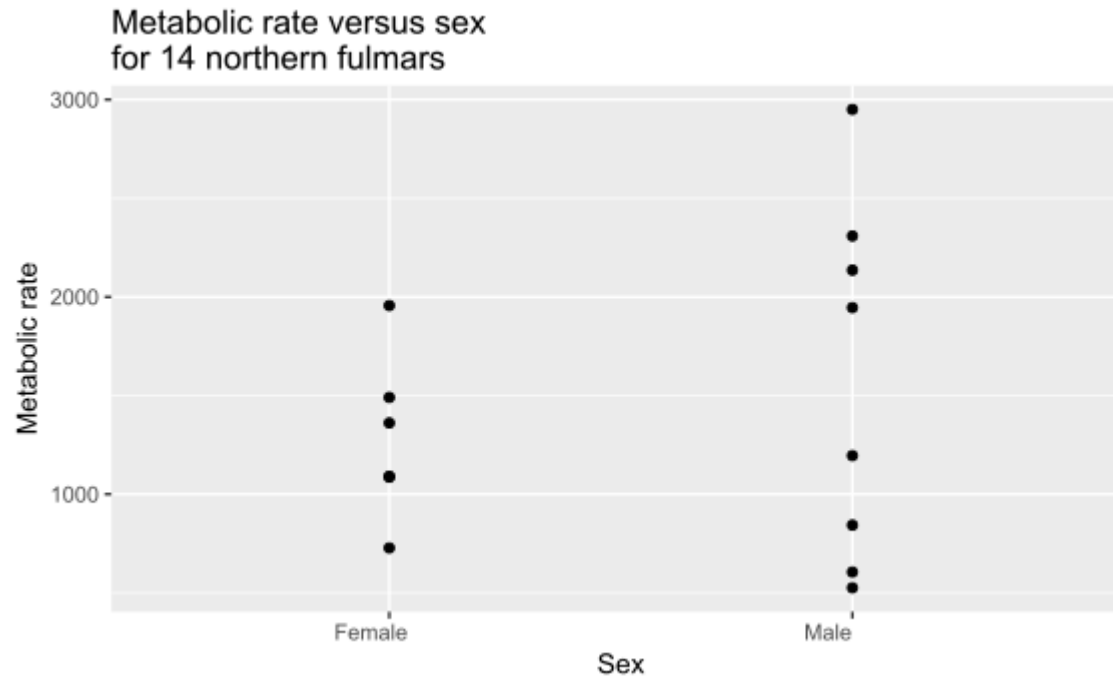


Standard deviation



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

Standard deviation



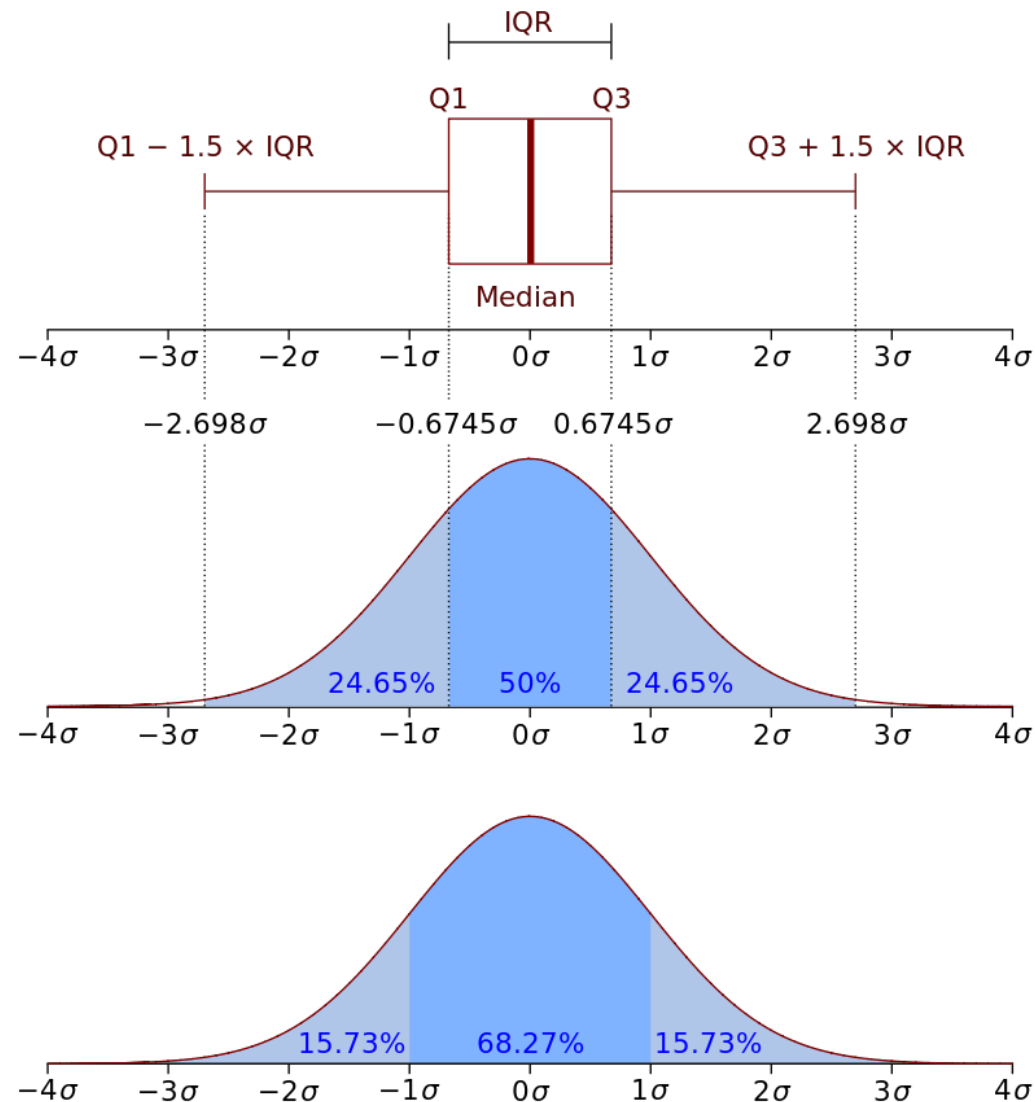
Sum of squares calculation for female fulmars

Animal	Sex	Metabolic rate	Mean	Difference from mean	Squared difference from mean
1	Female	727.7	1285.5	-557.8	311140.84
2	Female	1086.5	1285.5	-199.0	39601.00
3	Female	1091.0	1285.5	-194.5	37830.25
4	Female	1361.3	1285.5	75.8	5745.64
5	Female	1490.5	1285.5	205.0	42025.00
6	Female	1956.1	1285.5	670.6	449704.36
Mean of metabolic rates			1285.5	Sum of squared differences	886047.09

$$s = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}} = \sqrt{\frac{886047.09}{5}} = 420.96.$$

Interquartile range (IQR), also called the midspread, middle 50%, or H-spread, is equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles

$$\text{IQR} = Q_3 - Q_1$$



Interquartile range (IQR), also called the midspread, middle 50%, or H-spread, is equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles

$$IQR = Q_3 - Q_1$$

i	x[i]	Median	Quartile
1	7	$Q_2=87$ (median of whole table)	$Q_1=31$ (median of upper half, from row 1 to 6)
2	7		
3	31		
4	31		
5	47		
6	75		
7	87		$Q_3=119$ (median of lower half, from row 8 to 13)
8	115		
9	116		
10	119		
11	119		
12	155		
13	177		

For the data in this table the interquartile range is $IQR = Q_3 - Q_1 = 119 - 31 = 88$.

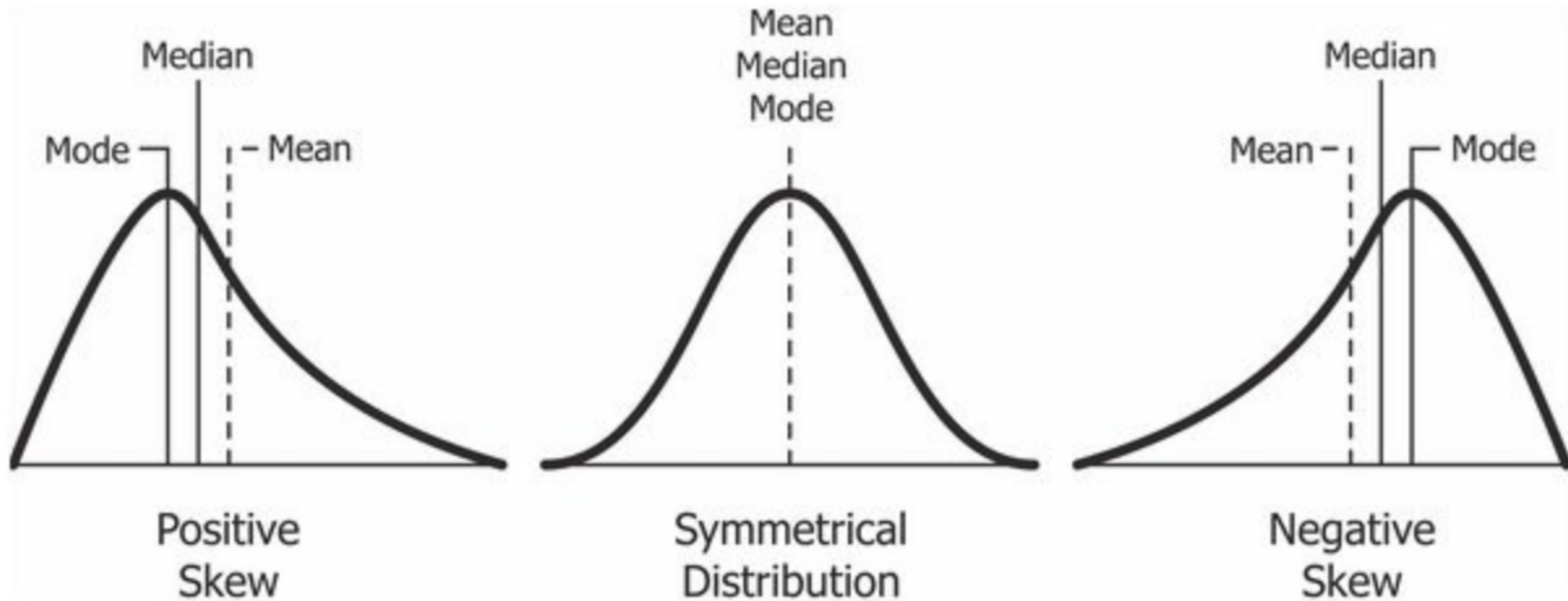
Shape

Examples of shape measures include:

- Skewness - a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean
- kurtosis - a measure of the "tailedness" of the probability distribution of a real-valued random variable

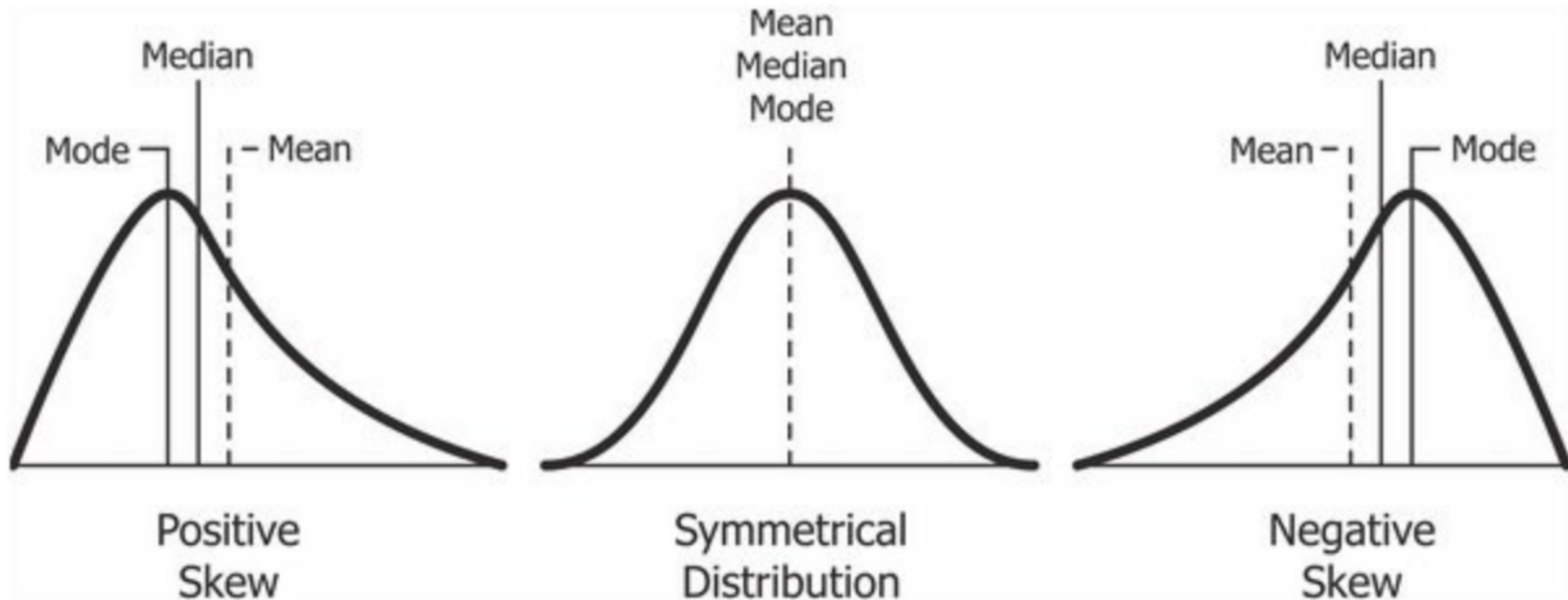
Shape

Skewness is a measure of the **asymmetry** of the probability distribution of a real-valued random variable about its mean. It can be positive, zero, negative, or undefined



Shape

Skewness is a measure of the **asymmetry** of the probability distribution of a real-valued random variable about its mean. It can be positive, zero, negative, or undefined



positive skew - the right tail is longer; the mass of the distribution is concentrated on the left. The distribution is said to be right-skewed, right-tailed, or skewed to the right (right refers to the right tail being drawn out and, often, the mean being skewed to the right of a typical center of the data)

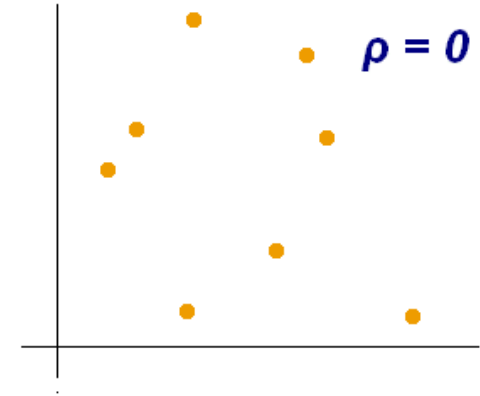
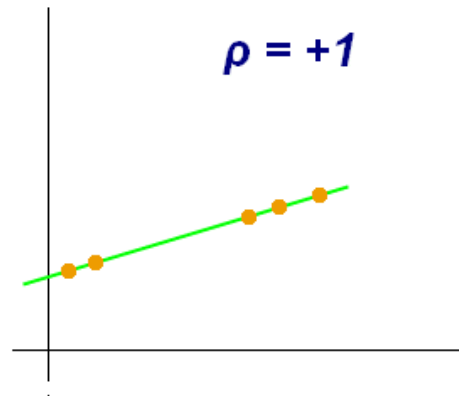
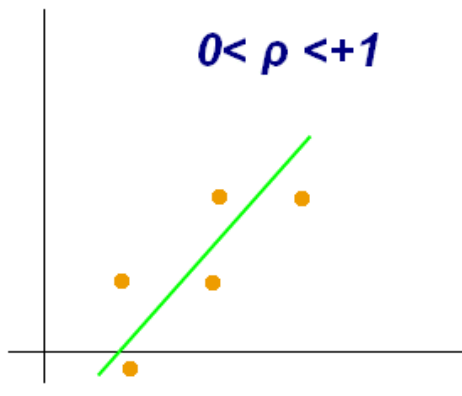
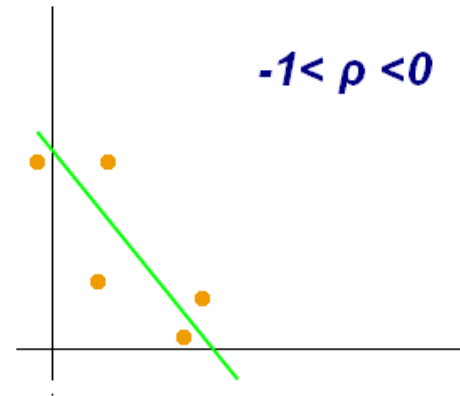
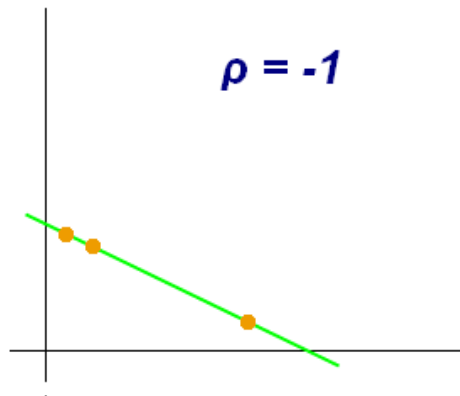
negative skew - the left tail is longer; the mass of the distribution is concentrated on the right. The distribution is said to be left-skewed, left-tailed, or skewed to the left (left refers to the left tail being drawn out and, often, the mean being skewed to the left of a typical center of the data)

Pearson correlation coefficient

A statistic that measures linear correlation between two variables X and Y

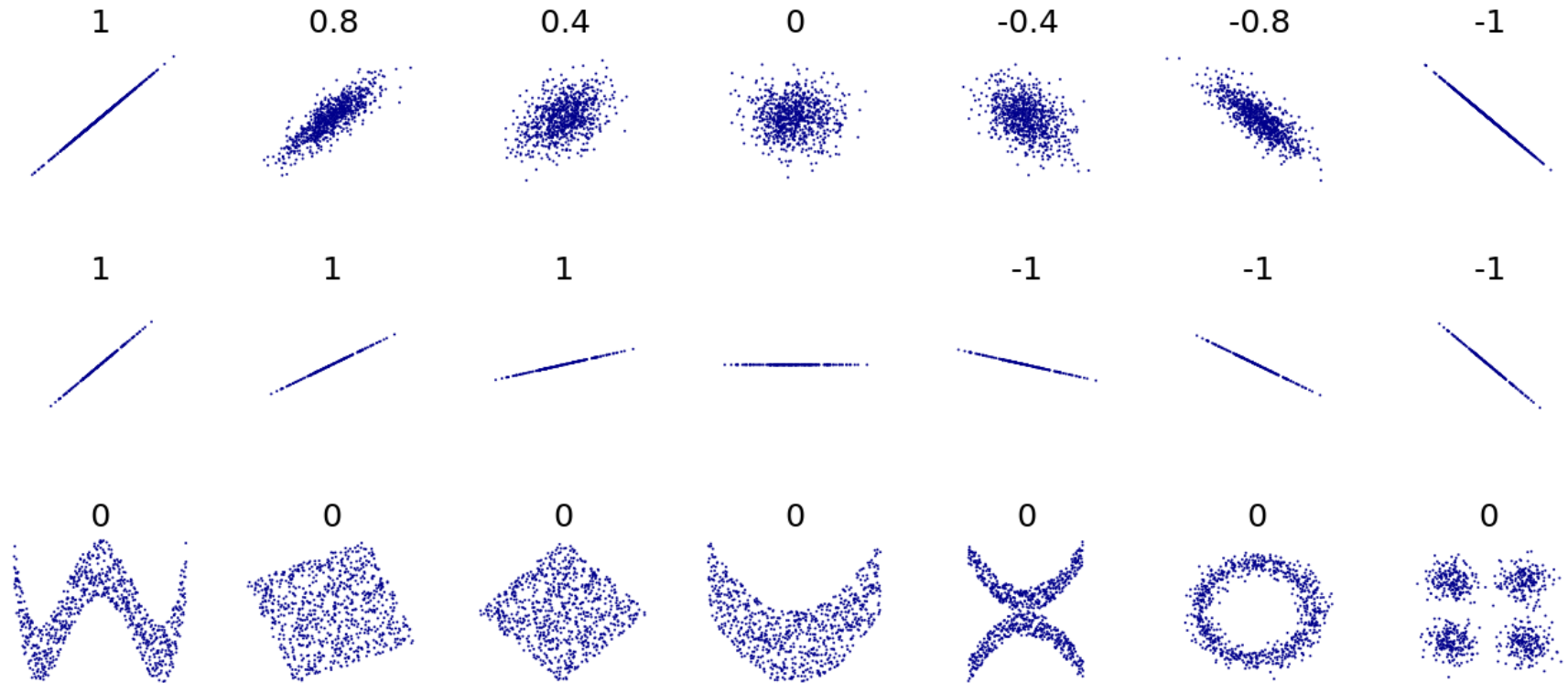
It has a value between +1 and -1, where:

- 1 is total positive linear correlation
- 0 is no linear correlation
- -1 is total negative linear correlation



Dependence

Used if more than one variable is measured



Several sets of (x, y) points, with the correlation coefficient of x and y for each set. Note that the correlation reflects the strength and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.

Pearson correlation coefficient

A statistic that measures **linear** correlation between two variables X and Y

It has a value between +1 and -1, where:

- 1 is total positive linear correlation
- 0 is no linear correlation
- -1 is total negative linear correlation

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the **covariance**
- σ_X is the **standard deviation** of X
- σ_Y is the **standard deviation** of Y

Population

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

where:

- n is sample size
- x_i, y_i are the individual sample points indexed with i
- $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ (the sample **mean**); and analogously for \bar{y}

Sample

Spearman's correlation coefficient

Two variables can be related by a nonlinear relationship

Two variables being considered may have a non-Gaussian distribution.

$$r_s = \rho_{\text{rg}_X, \text{rg}_Y} = \frac{\text{cov}(\text{rg}_X, \text{rg}_Y)}{\sigma_{\text{rg}_X} \sigma_{\text{rg}_Y}},$$

where

ρ denotes the usual Pearson correlation coefficient, but applied to the rank variables,

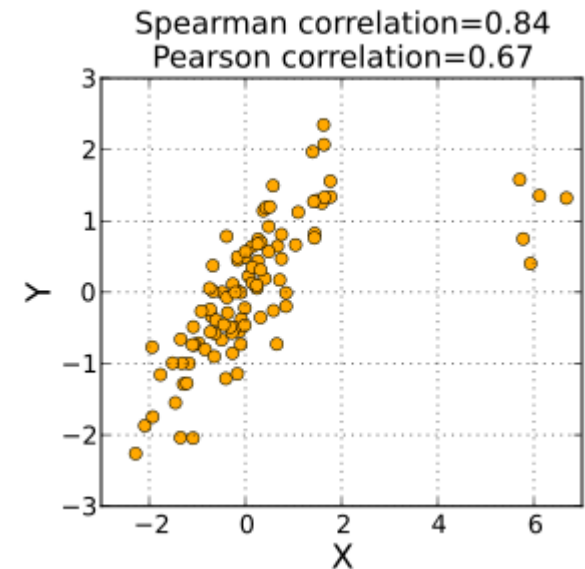
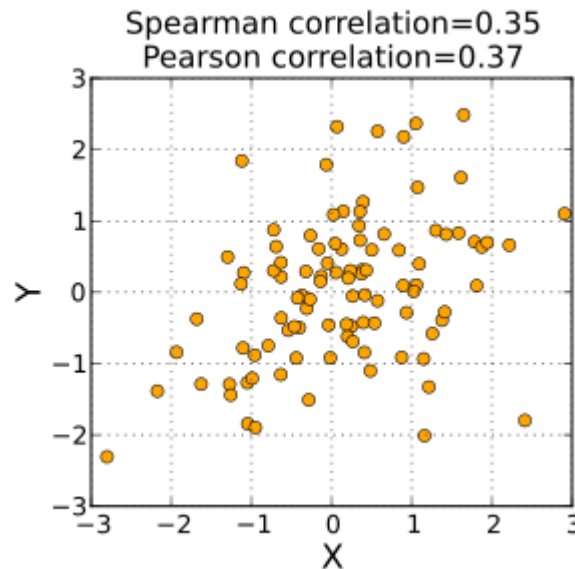
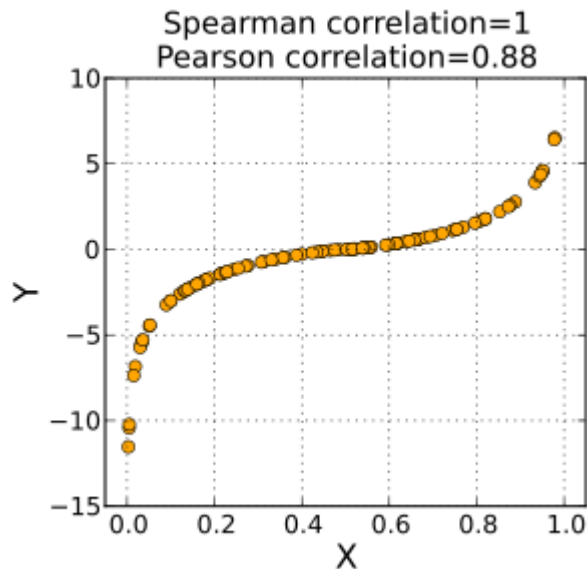
$\text{cov}(\text{rg}_X, \text{rg}_Y)$ is the covariance of the rank variables, σ_{rg_X} and σ_{rg_Y} are the standard deviations of the rank variables.

Spearman's correlation coefficient

Spearman's correlation coefficient (often denoted by the Greek letter ρ (rho)) is a nonparametric measure of **rank correlation** (statistical dependence between the rankings of two variables)

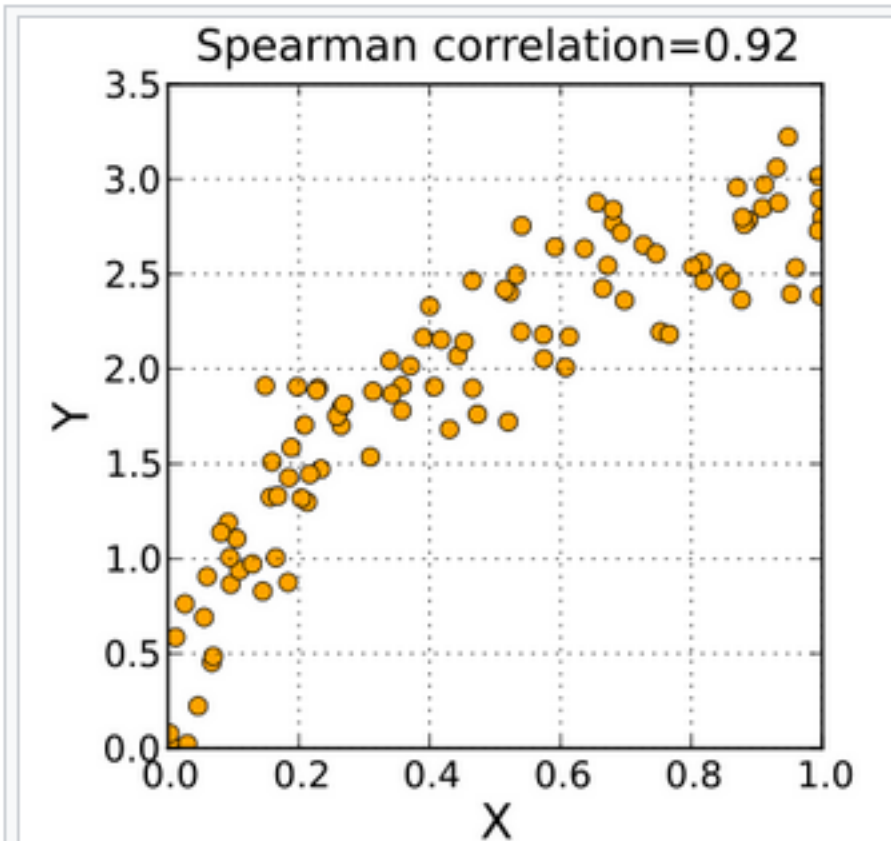
It assesses how well the relationship between two variables can be described using a monotonic function

while Pearson's correlation assesses linear relationships, Spearman's correlation assesses monotonic relationships (whether linear or not)

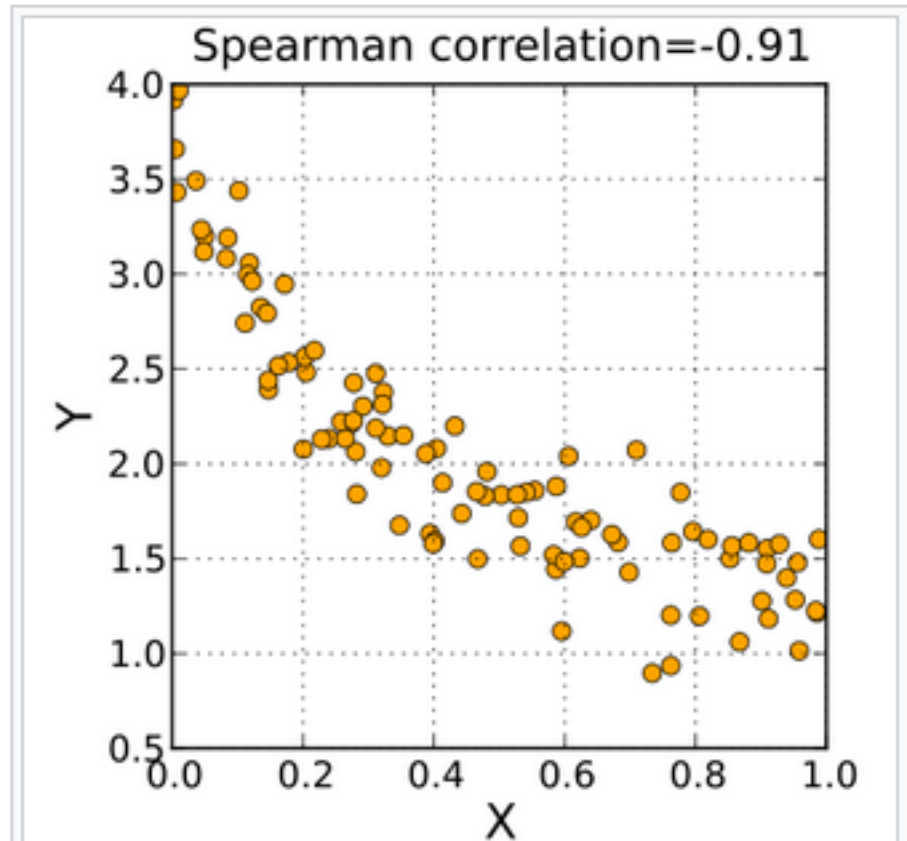


Spearman's correlation coefficient

Positive and negative Spearman rank correlations



A positive Spearman correlation coefficient corresponds to an increasing monotonic trend between X and Y .



A negative Spearman correlation coefficient corresponds to a decreasing monotonic trend between X and Y .

Inferential Statistics

Inferential Statistics use the data to **learn** about the population that the sample of data is thought to represent

Statistical inference is the process of using data analysis to deduce properties of an underlying distribution of probability

We include here:

- point estimate
- interval estimate
- credible interval
- rejection of a hypothesis
- clustering or classification

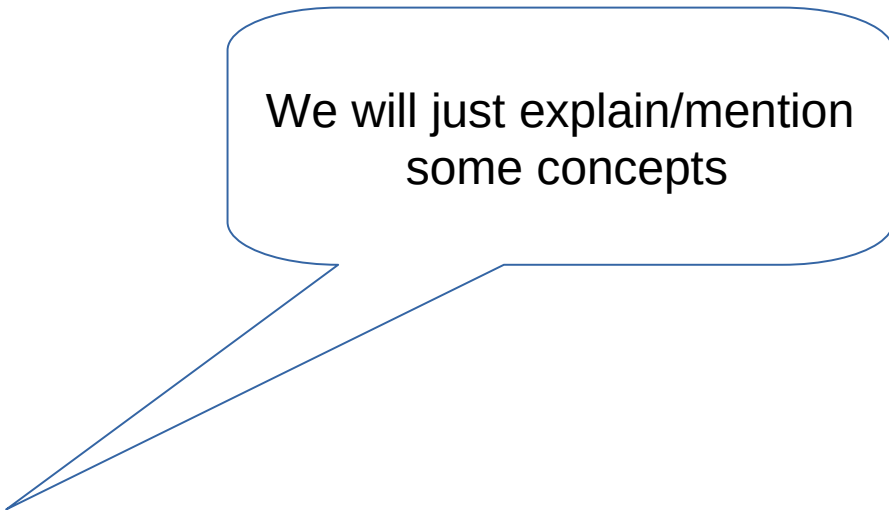
Inferential Statistics

Inferential Statistics use the data to **learn** about the population that the sample of data is thought to represent

Statistical inference is the process of using data analysis to deduce properties of an underlying distribution of probability

We include here:

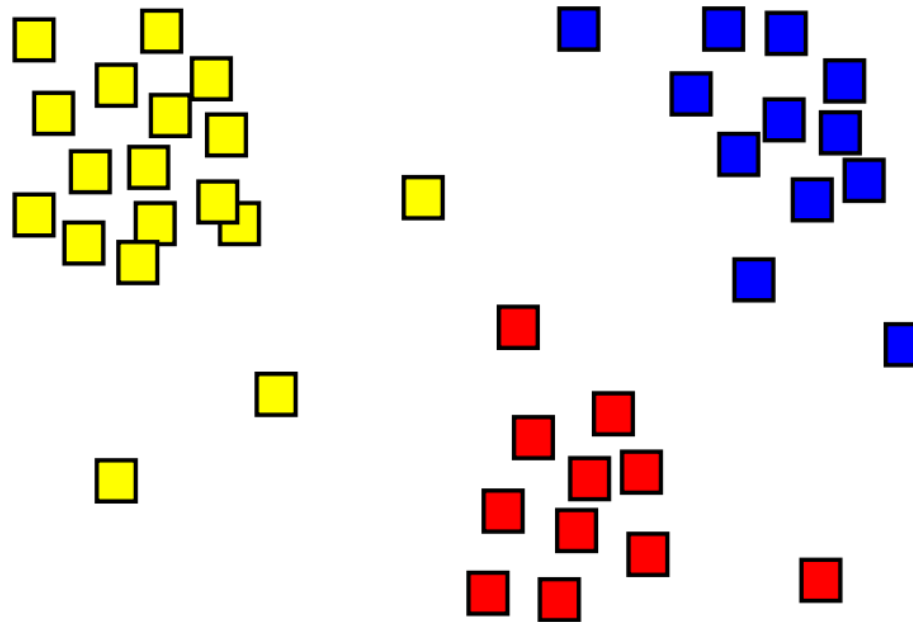
- point estimate
- interval estimate
- credible interval
- rejection of a hypothesis
- clustering or classification



We will just explain/mention
some concepts

Clustering (unsupervised learning)

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters)



Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters)

There are many algorithms:

- hierarchical clustering (connectivity-based)
- k-means clustering (centroid-based)
- EM clustering (distribution-based)
- DBSCAN (density-based)

and few other

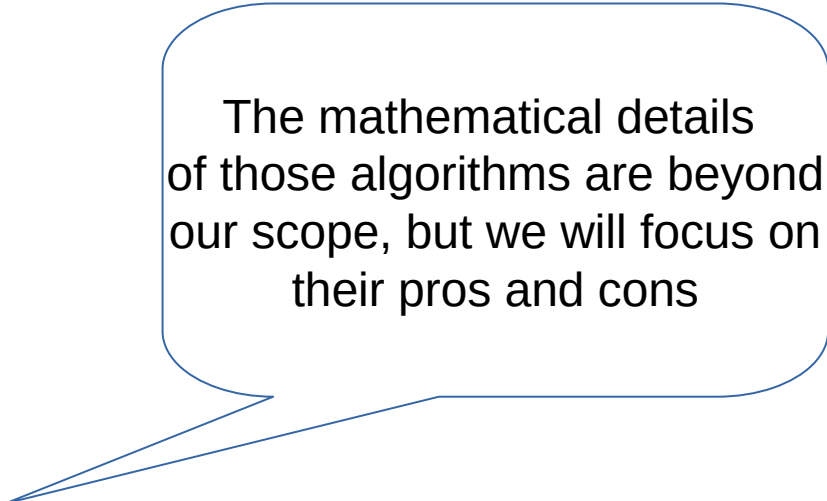
Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters)

There are many algorithms:

- hierarchical clustering (connectivity-based)
- k-means clustering (centroid-based)
- EM clustering (distribution-based)
- DBSCAN (density-based)

and few other



The mathematical details of those algorithms are beyond our scope, but we will focus on their pros and cons

Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters)

There are many algorithms:

- hierarchical clustering (connectivity-based)
- k-means clustering (centroid-based)
- EM clustering (distribution-based)
- DBSCAN (density-based)

and few other

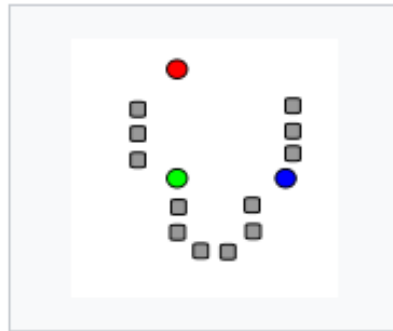
The mathematical details of those algorithms are beyond our scope, but we will focus on their pros and cons

Remember: most of those algorithms are already implemented in python, you need only to know when and how to apply each one

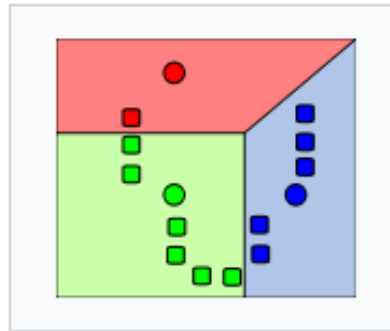
Clustering

k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest **mean** (cluster centers or cluster **centroid**), serving as a prototype of the cluster

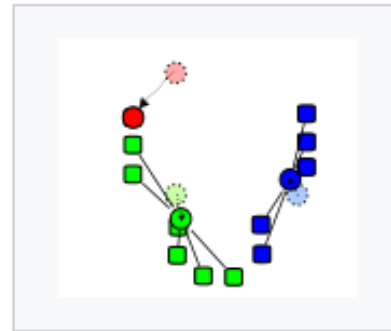
Demonstration of the standard algorithm



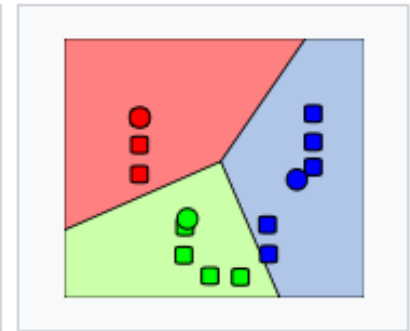
1. k initial "means" (in this case $k=3$) are randomly generated within the data domain (shown in color).



2. k clusters are created by associating every observation with the nearest mean. The partitions here represent the **Voronoi diagram** generated by the means.



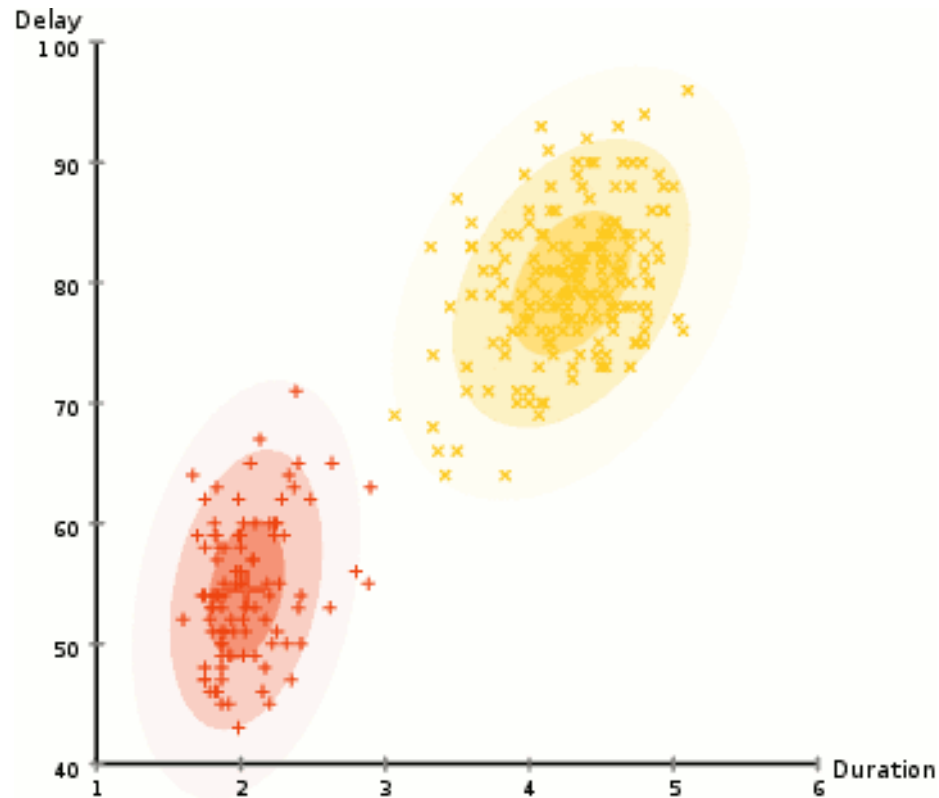
3. The **centroid** of each of the k clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.

Clustering

Expectation-maximization (EM) is an iterative algorithm to find maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables

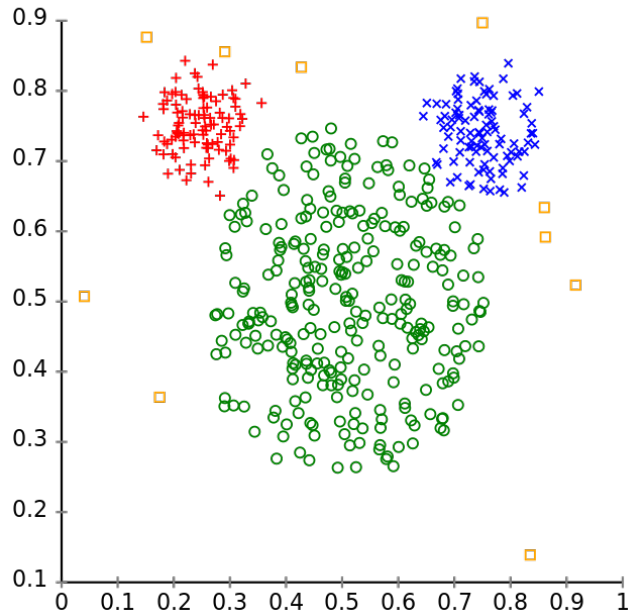


Clustering

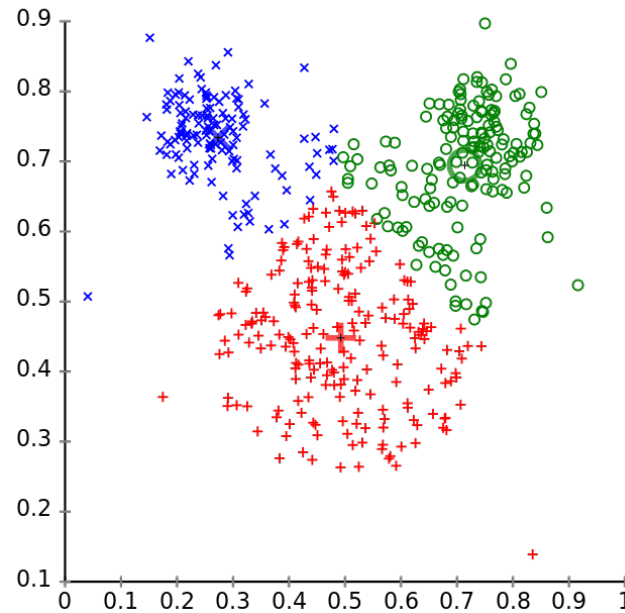
k-means clustering vs EM

Different cluster analysis results on "mouse" data set:

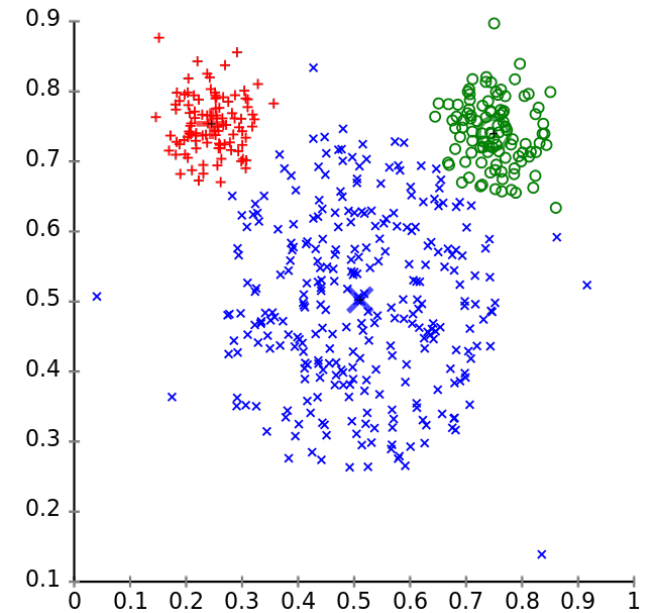
Original Data



k-Means Clustering



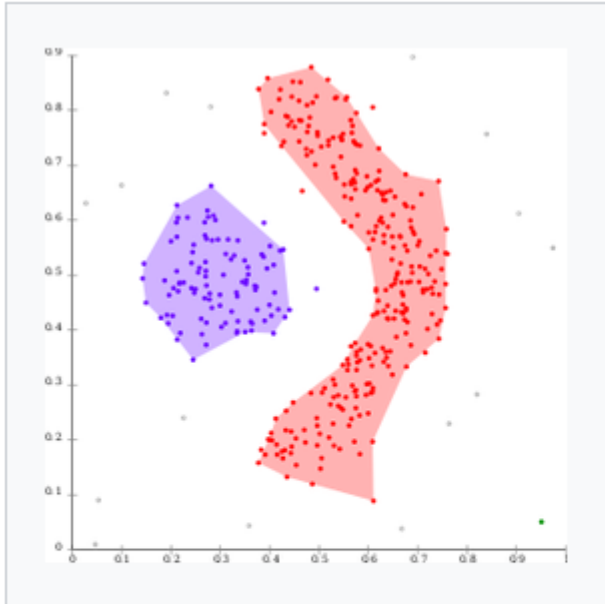
EM Clustering



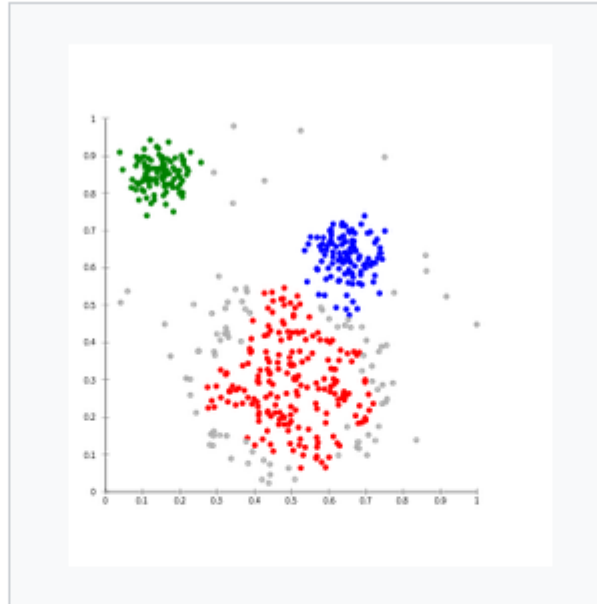
K-means works good when your dataset can be clustered into k equal size clusters, the way the algorithm works is that it will balance the size of the clusters until they are more or less similar in size. On the other hand EM does not have such problem.

Clustering

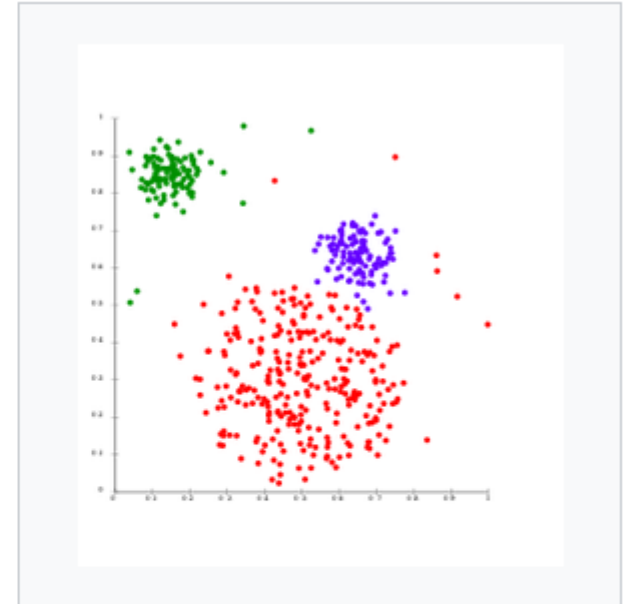
Density-based clustering examples



Density-based clustering with **DBSCAN**.



DBSCAN assumes clusters of similar density, and may have problems separating nearby clusters



OPTICS is a DBSCAN variant, improving handling of different densities clusters

Frequentist inference

draw conclusions from sample data by emphasizing the frequency or proportion of the data

Examples of frequentist inference

- p-value
- confidence interval

p-value

p-value (probability value) is the probability of obtaining test results at least as extreme as the results actually observed during the test, assuming that the null hypothesis is correct

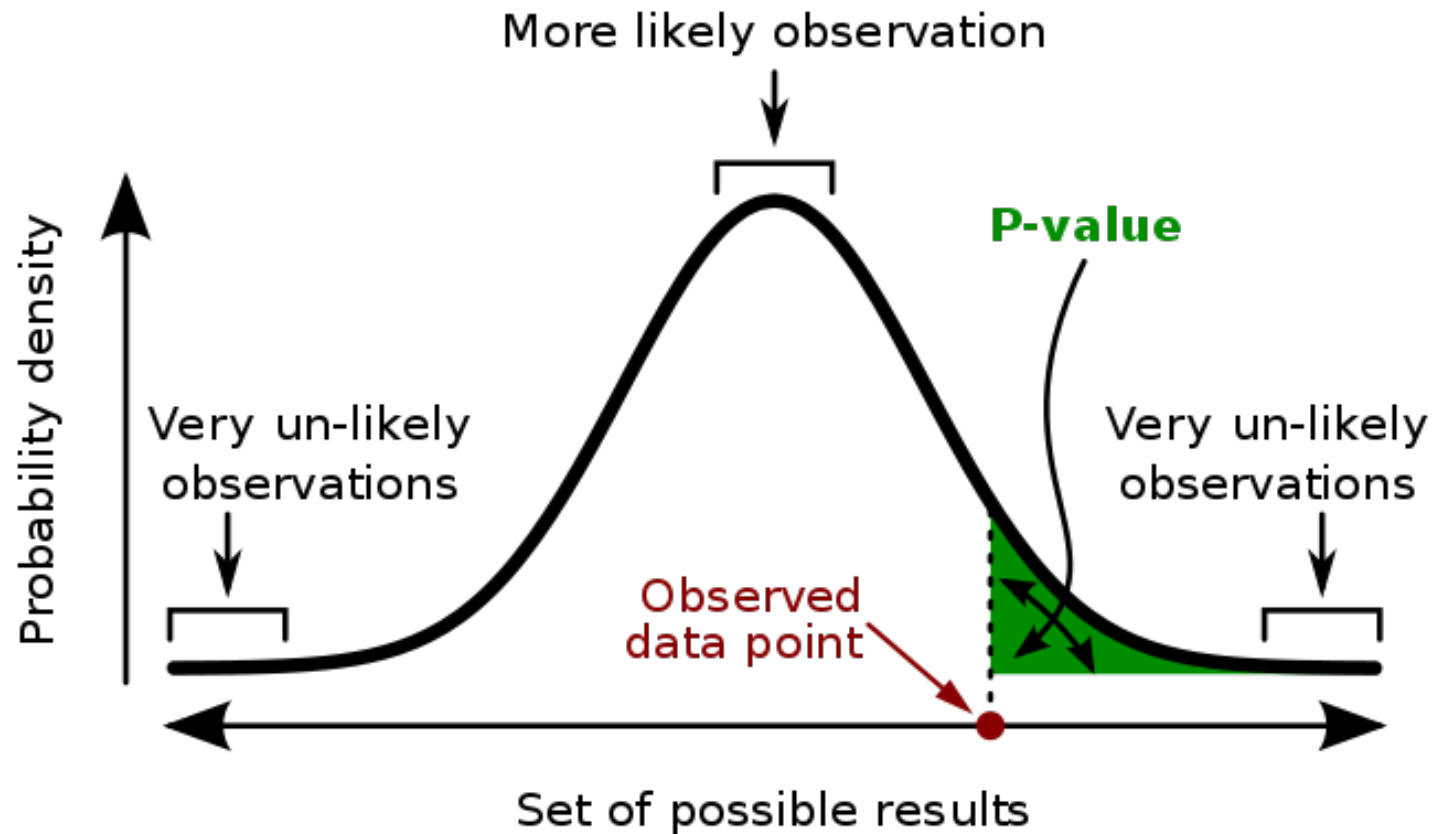
Important:

$\Pr(\text{observation} \mid \text{hypothesis}) \neq \Pr(\text{hypothesis} \mid \text{observation})$

The probability of observing a result given that some hypothesis is true is *not equivalent* to the probability that a hypothesis is true given that some result has been observed.

Using the p-value as a “score” is committing an egregious logical error:
the transposed conditional fallacy.

p-value



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

Clarifications about p-values

1) The p-value is not the probability that the null hypothesis is true, or the probability that the alternative hypothesis is false. A p-value can indicate the degree of compatibility between a dataset and a particular hypothetical explanation (such as a null hypothesis). Specifically, the p-value can be taken as the prior probability of obtaining an effect that is at least as extreme as the observed effect, given that the null hypothesis is true. This should not be confused with the posterior probability that the null hypothesis is true given the observed effect (see prosecutor's fallacy). In fact, frequentist statistics does not attach probabilities to hypotheses.

2) The p-value is not the probability that the observed effects were produced by random chance alone. The p-value is computed under the assumption that a certain model, usually the null hypothesis, is true. This means that the p-value is a statement about the relation of the data to that hypothesis.

3) The 0.05 significance level is merely a convention. The 0.05 significance level (alpha level) is often used as the boundary between a statistically significant and a statistically non-significant p-value. However, this does not imply that there is generally a scientific reason to consider results on opposite sides of any threshold as qualitatively different.

4) The p-value does not indicate the size or importance of the observed effect. A small p-value can be observed for an effect that is not meaningful or important. In fact, the larger the sample size, the smaller the minimum effect needed to produce a statistically significant p-value (see effect size).

Confidence interval

Confidence interval (CI) is a type of estimate computed from the statistics of the observed data. This proposes a range of plausible values for an unknown parameter (for example, the mean)

The interval has an associated **confidence level** that the true parameter is in the proposed range. Given observations x_1, \dots, x_n and a confidence level γ (gamma), a valid confidence interval has a γ probability of containing the true underlying parameter.

The interval has an associated confidence level that the true parameter is in the proposed range

In general terms, a confidence interval for an unknown parameter is based on sampling the distribution of a corresponding estimator

For example: if the confidence level (CL) is 90% then in hypothetical indefinite data collection, in 90% of the samples the interval estimate will contain the population parameter

Statistical classification (supervised learning)

The classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known

Examples:

- assigning a given email to the "spam" or "non-spam" class
- assigning a diagnosis to a given patient based on observed characteristics of the patient (sex, blood pressure, presence or absence of certain symptoms, etc.)

Statistical classification (supervised learning)

The classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known

Approaches:

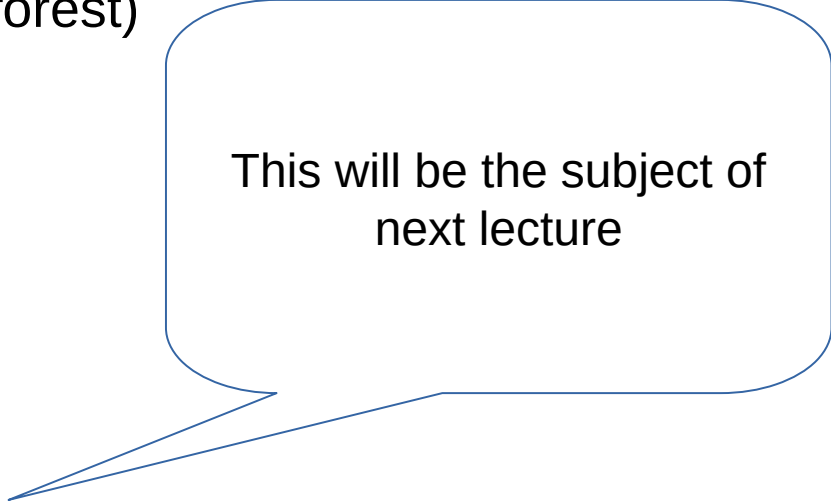
- Decision trees
- Ensembles (Bagging Boosting, Random forest)
- k-NN
- Linear regression
- Logistic regression
- Naive Bayes
- Relevance vector machine (RVM)
- Support vector machine (SVM)
- Perceptron
- Artificial neural networks
- Deep learning

Statistical classification (supervised learning)

The classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known

Approaches:

- Decision trees
- Ensembles (Bagging Boosting, Random forest)
- k-NN
- Linear regression
- Logistic regression
- Naive Bayes
- Relevance vector machine (RVM)
- Support vector machine (SVM)
- Perceptron
- Artificial neural networks
- Deep learning



This will be the subject of next lecture

Thank you for your time
and
See you at the next lecture

Any other
questions & comments

l.kozlowski@mimuw.edu.pl