Introduction to Machine Learning
NPFL 054

http://ufal.mff.cuni.cz/course/npfl054

Barbora Hladká **Martin Holub** **Ivana Lukšová**

{Hladka | Holub | Luksova}@ufal.mff.cuni.cz

Charles University in Prague,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Outline — Lab session #2

1. **Overview of mathematial foundations**
   - Why statistics?
   - Probability basics
   - Warming exercises
   - Conditional probability and the Bayes rule
   - Random variables and probability distributions
   - Examples of probability distributions
   - Measures of location and spread
2. **Homework**
3. **Literature**
4. **More easy exercises**

Why statistics and probability theory?

Motivation

- In machine learning, models come from data and provide insights for understanding data (unsupervised classification) or making prediction (supervised learning).
- A good model is often a model which not only fits the data but gives good predictions, even if it is not interpretable.

Statistics

- is the science of the collection, organization, and interpretation of data
- uses the probability theory

Two purposes of statistical analysis

Statistics is the study of the collection, organization, analysis, and interpretation of data. It deals with all aspects of this, including the planning of data collection in terms of the design of surveys and experiments.

Description

- describing what was observed in sample data numerically or graphically

Inference

- drawing inferences about the population represented by the sample data

Populations and statistics (plural)

In statistics, a **population** is any finite or infinite collection of items under consideration. Usually the population is an entire set of all entities or events of a given type/kind.

The word *statistics* (the scientific discipline) should not be confused with the word **statistic**, referring to a quantity calculated from a data **sample**. Statistics are calculated by applying a function (or a statistical algorithm) to the values of the items comprising the sample.

More formally, statistical theory defines a statistic as a function of a sample where the function itself is independent of the sample's distribution; that is, the function can be stated before realisation of the data. The term statistic is used both for the function and for the value of the function on a given sample.

Statistical parameters, and estimators

Statistical parameter is a value that describes a property of a statistical population.

Statistical parameters are distinct from statistics. They are generally *not* computable because often the population is too large to examine and measure all its items. *Parameters statistically characterize populations, while statistics characterize samples.*

A statistic, when used to estimate a population parameter, is called an **estimator**. For instance, the *sample mean* is a statistic which estimates the *population mean*, which is a parameter.

Observability
A statistic is an *observable* random variable, which differentiates it from a parameter that is a generally *unobservable* quantity. A parameter can only be computed exactly if the entire population can be observed without error.

Descriptive statistics

Descriptive statistics describes the main characteristics of a collection of data. It summarizes the population data by describing what was observed in the sample numerically (quantitatively) or graphically.

Numerical descriptors include e.g. mean and standard deviation for continuous data types (like heights or weights), while frequency and percentage are more useful in terms of describing categorical data (like race).

A **histogram** is an example of a graphical representation showing a visual impression of the distribution of data. It is an estimate of the probability distribution.

Inferential statistics

Inferential statistics uses patterns in the sample data to draw inferences about the population represented, accounting for randomness. These inferences may take the form of:

- answering yes/no questions about the data (hypothesis testing)
- estimating numerical characteristics of the data (estimation)
- describing associations within the data (correlation)
- modeling relationships within the data (for example, using regression analysis)

Inference can extend to forecasting, prediction and estimation of unobserved values either in or associated with the population being studied; it can include extrapolation and interpolation of time series or spatial data, and can also include data mining.

Two approaches to probability

"classical"

- based on the knowledge of probability of elementary events

"statistical"

- based on empirical observation and relative frequencies

Probability theory – basic terms

- random experiment
- elementary outcomes ω_i
- sample space $\Omega = \bigcup \{\omega_i\}$
- event $A \subseteq \Omega$
- complement of an event $A^c = \Omega \setminus A$
- probability of any event is a non-negative value $\mathrm{P}(A) \geq 0$
- total probability of all elementary outcomes is one

$$\sum_{\omega \in \Omega} \mathrm{P}(\omega) = 1$$

- if two events A, B are *mutually exclusive* (i.e. $A \cap B = \emptyset$), then $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$

Basic formulas to calculate probabilities

Generally, probability of an event A is

$$P(A) = \sum_{\omega \in A} P(\omega)$$

Probability of an complement event is

$$P(A^c) = 1 - P(A)$$

**IF all elementary outcomes have the same probability,
THEN probability of an event is given by the proportion of**

$$\frac{\text{number-of-desired-outcomes}}{\text{total-number-of-outcomes-possible}}$$

What is $\mathrm{P}(A \text{ or } B)$?

$$\mathrm{P}(A \text{ or } B) = \mathrm{P}(A \cup B)$$

For *mutually exclusive* events:

$$\mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B)$$

otherwise (generally):

$$\mathrm{P}(A \text{ or } B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$$

What is P(A and B)?

$$\mathrm{P}(A \text{ and } B) = \mathrm{P}(A \cap B)$$

Two events A and B are independent of each other if the occurrence of one has no influence on the probability of the other.

For independent events:

$$\mathrm{P}(A \text{ and } B) = \mathrm{P}(A) \cdot \mathrm{P}(B)$$

otherwise (generally):

$$\mathrm{P}(A \text{ and } B) = \mathrm{P}(A \mid B) \cdot \mathrm{P}(B) = \mathrm{P}(B \mid A) \cdot \mathrm{P}(A)$$

Warming exercises

Rolling two dice, observing the sum. What is likelier?

a) the sum is even

b) the sum is greater than 8

c) the sum is 5 or 7

What is likelier:

a) rolling at least one six in four throws of a single die, OR

b) rolling at least one double six in 24 throws of a pair of dice?

When we roll 3 dice, which sum is likelier, 11 or 12?

- Note that there are only 6 options how to get the sum 11:
  4+4+3 | 4+5+2 | 4+6+1 | 5+5+1 | 5+3+3 | 6+3+2
- And also, there are only 6 options how to get the sum 12:
  4+4+4 | 4+5+3 | 4+6+2 | 5+5+2 | 5+6+1 | 6+3+3

Definition of conditional probability

Conditional probability of the event A given the event B is

$$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)} = \frac{\mathrm{P}(A, B)}{\mathrm{P}(B)}$$

Or, in other words,

$$\mathrm{P}(A, B) = \mathrm{P}(A \mid B)\mathrm{P}(B)$$

Statistically independent events

Definition: The random event B is *independent* of the random event A, if the following holds true at the same time:

$$\mathrm{P}(B) = \mathrm{P}(B \mid A), \quad \mathrm{P}(B) = \mathrm{P}(B \mid A^c).$$

An equivalent definition is that B is independent of A if

$$\mathrm{P}(A) \cdot \mathrm{P}(B) = \mathrm{P}(A \cap B).$$

Computing conditional probability

Exercise

The probability that it is Friday and that a student is absent is 3 %. Since there are 5 school days in a week, the probability that it is Friday is 20 %. What is the probability that a student is absent given that today is Friday?

Solution

Random experiment:

At a random moment we observe the day in working week and the fact if a student is absent.

Events:

- A ... it is Friday
- B ... a student is absent

Probabilities:

- $\mathrm{P}(A, B) = 0.03$
- $\mathrm{P}(A) = 0.2$
- $\mathrm{P}(B \mid A) = \mathrm{P}(A, B)/\mathrm{P}(A) = 0.15$

Correct answer:

- The probability that a student is absent given that today is Friday is 15 %.

Bayes rule

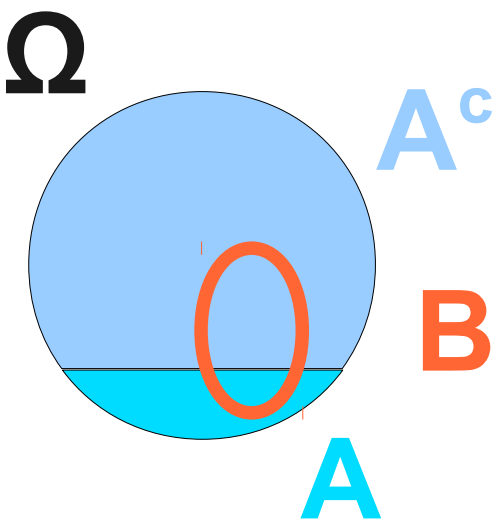
Because of the symmetry $\mathrm{P}(A, B) = \mathrm{P}(B, A)$, we have

$$\mathrm{P}(A, B) = \mathrm{P}(A \mid B)\mathrm{P}(B) = \mathrm{P}(B \mid A)\mathrm{P}(A) = \mathrm{P}(B, A)$$

And thus

$$\mathrm{P}(B \mid A) = \frac{\mathrm{P}(A \mid B)\mathrm{P}(B)}{\mathrm{P}(A)}$$

Computing $P(B)$ when you know both $P(A)$ and the two probabilities $P(B \mid A)$ and $P(B \mid A^c)$



$$\Omega = A \cup A^c$$

$$B = (B \cap A) \cup (B \cap A^c)$$

$$P(B) = P(B, A) + P(B, A^c)$$

$$P(B) = P(B \mid A)\, P(A) + P(B \mid A^c)\, P(A^c)$$

Generalized Bayes rule

Generally, when the set of n events $A_1, A_2, \ldots A_n$ form a *mutually exclusive* and *exhaustive* system, i.e. $\Omega = \bigcup A_i$ and $A_i \cap A_j = \emptyset$ for $i \neq j$, then for any event $B \subset \Omega$, $\mathrm{P}(B) \neq 0$ we have

$$B = \bigcup_{i=1}^{n} B \cap A_i,$$

$$\mathrm{P}(B) = \sum_{i=1}^{n} \mathrm{P}(B \cap A_i) = \sum_{i=1}^{n} \mathrm{P}(B \mid A_i)\mathrm{P}(A_i).$$

Generalized Bayes rule says

$$\mathrm{P}(A_i \mid B) = \frac{\mathrm{P}(B \cap A_i)}{\mathrm{P}(B)} = \frac{\mathrm{P}(B \mid A_i)\mathrm{P}(A_i)}{\sum_{j=1}^{n} \mathrm{P}(B \mid A_j)\mathrm{P}(A_j)}$$

Using Bayes rule

Exercise

One coin in a collection of 65 has two heads. The rest are fair. If a coin, chosen at random from the lot and then tossed, turns up heads 6 times in a row, what is the probability that it is the two-headed coin?

Random variables

A **random variable** (or sometimes stochastic variable) is, roughly speaking, a variable whose value results from a measurement/observation on some type of random process. Intuitively, a random variable is a numerical or categorical description of the outcome of a random experiment (or a random event).

Random variables can be classified as either

- **discrete**
  = a random variable that may assume either a finite number of values or an infinite sequence of values (countably infinite)
- **continuous**
  = a variable that may assume any numerical value in an interval or collection of intervals.

Probability description

Probability distributions are used to describe the probabilities of different values occurring.

Discrete probability distributions are usually characterised by the *probability mass function* that assigns a probability to each possible value of the random variable (often it can be described by a table or by a histogram).

Continuous probability distributions are usually characterised either by the *cumulative distribution function* or by the *probability density function*.

Probability distribution

Probability distribution identifies either

- the probability of each value of a random variable (discrete case), or
- the probability of the value falling within a particular interval (continuous case)

You should definitely know

- probability mass function (PMF)
- histograms
- cumulative distribution function (CDF)
- probability density function (PDF)
- quantile function

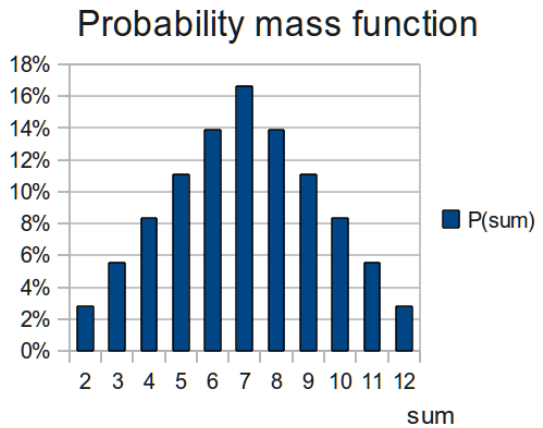
Example – PMF

- When you roll two dice and add the result together, what will be the total?

Rolling two dice – the sum

die 1 \ die 2	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

Histogram



Frequency list

sum	count	P(sum)
2	1	2.8% = 1/36
3	2	5.6% = 2/36
4	3	8.3% = 3/36
5	4	11.1% = 4/36
6	5	13.9% = 5/36
7	6	16.7% = 6/36
8	5	13.9% = 5/36
9	4	11.1% = 4/36
10	3	8.3% = 3/36
11	2	5.6% = 2/36
12	1	2.8% = 1/36

Mean and variance

- **Expected value or mean** of a (numerical) random variable is given by

$$\mathrm{E}\, X = \sum_{\text{all } x} x \cdot \mathrm{P}(X = x)$$

- **Variance** of a (numerical) random variable is defined as

$$\operatorname{var} X = \mathrm{E}\,(X - \mathrm{E}\, X)^2$$

- Variance is often computed using the equality $\operatorname{var} X = \mathrm{E}\, X^2 - (\mathrm{E}\, X)^2$
- **Standard deviation** is often denoted by $\sigma = \sqrt{\operatorname{var} X}$

Binomial distribution

- The binomial distribution is a discrete distribution of values $1, \ldots, n$ ($n \in \mathbb{N}$). If a random variable X gets only values $1, \ldots, n$ with probabilities

$$\mathrm{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

  where $p \in (0, 1)$, then X has binomial distribution

$$X \sim \mathrm{Bi}(n, p).$$

- If $q = 1 - p$, then $\mathrm{E}\, X = np$ and $\mathrm{var}\, X = npq$.

- If $n = 1$, we call it *Bernoulli distribution*.

Examples/exercises

- When you roll a die 50 times, how many sixes will you get?

Examples/exercises

- When you roll a die 50 times, how many sixes will you get?
- How many times do you need to roll a die to get the first six?

Examples/exercises

- When you roll a die 50 times, how many sixes will you get?
- How many times do you need to roll a die to get the first six?
- How many times do you need to roll a die to get just six sixes?

Examples/exercises

- When you roll a die 50 times, how many sixes will you get?
- How many times do you need to roll a die to get the first six?
- How many times do you need to roll a die to get just six sixes?
- When you roll six dice, how many different numbers will you get?

Normal distribution

- The normal distribution is a continuous distribution and has density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

where μ is the mean of the distribution and σ is the standard deviation.

When a random variable X has normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, we write

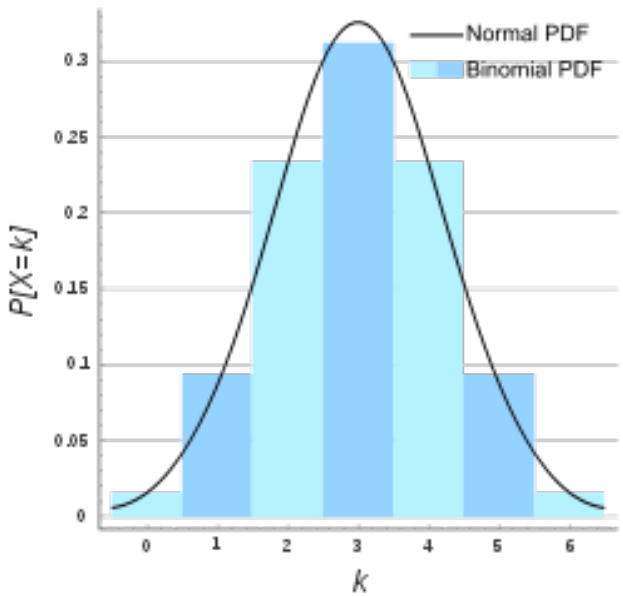
$$X \sim \mathrm{N}(\mu, \sigma^2).$$

Then $\mathrm{E}\, X = \mu$ and $\operatorname{var} X = \sigma^2$.

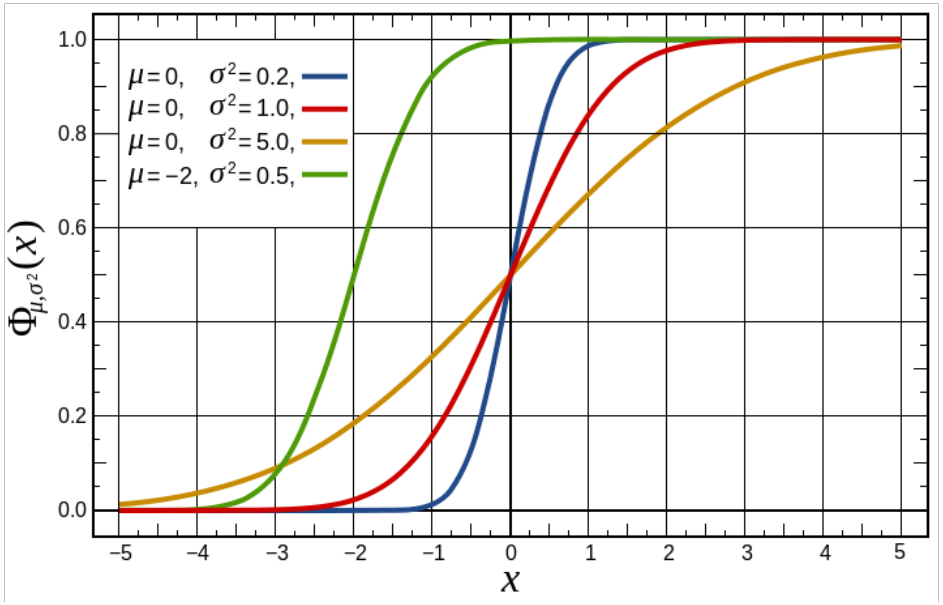
- The distribution $\mathrm{N}(0, 1)$ is called *standard* normal distribution.
It has density $\phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \mathrm{e}^{-x^2/2}$
and cumulative distribution $\Phi(x) = \int_{-\infty}^{x} \phi(t)\,\mathrm{d}t.$

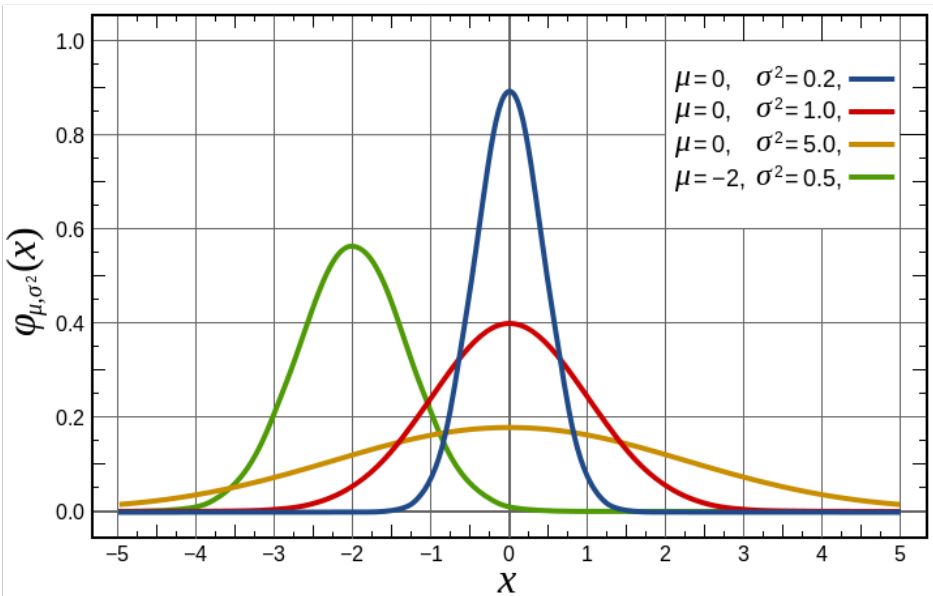
Normal approximation

If n is large enough, then a reasonable approximation to $\mathrm{Bi}(n, p)$ is given by the normal distribution $\mathrm{N}(np, np(1-p))$.



Normal distribution – CDF and PDF





Example – who is typical?

Exercise

In 2001 it was found that the average male height in the Czech Republic is 180.3 cm, while the average female height is 167.2 cm. We assume that both distributions are normal.

Question: Who is more typical: a man with the height of 178 cm, or a woman with the height of 169 cm?

Normal distribution – example

Question:
Having a normally distributed population with mean μ and variance σ^2, what is the probability that a randomly selected value will fall into the interval $(\mu - \sigma, \mu + \sigma)$?

Normal distribution – example

Question:
Having a normally distributed population with mean μ and variance σ^2, what is the probability that a randomly selected value will fall into the interval $(\mu - \sigma, \mu + \sigma)$?

Answer:

- about 68% of values drawn from a normal distribution are within one standard deviation σ away from the mean
- about 95% of the values lie within two standard deviations
- about 99.7% are within three standard deviations.

This fact is known as the 68-95-99.7 rule, or the empirical rule, or the 3-sigma rule.

Normal distribution – example

Question:
Having a normally distributed population with mean μ and variance σ^2, what is the probability that a randomly selected value will fall into the interval $(\mu - \sigma, \mu + \sigma)$?

Answer:

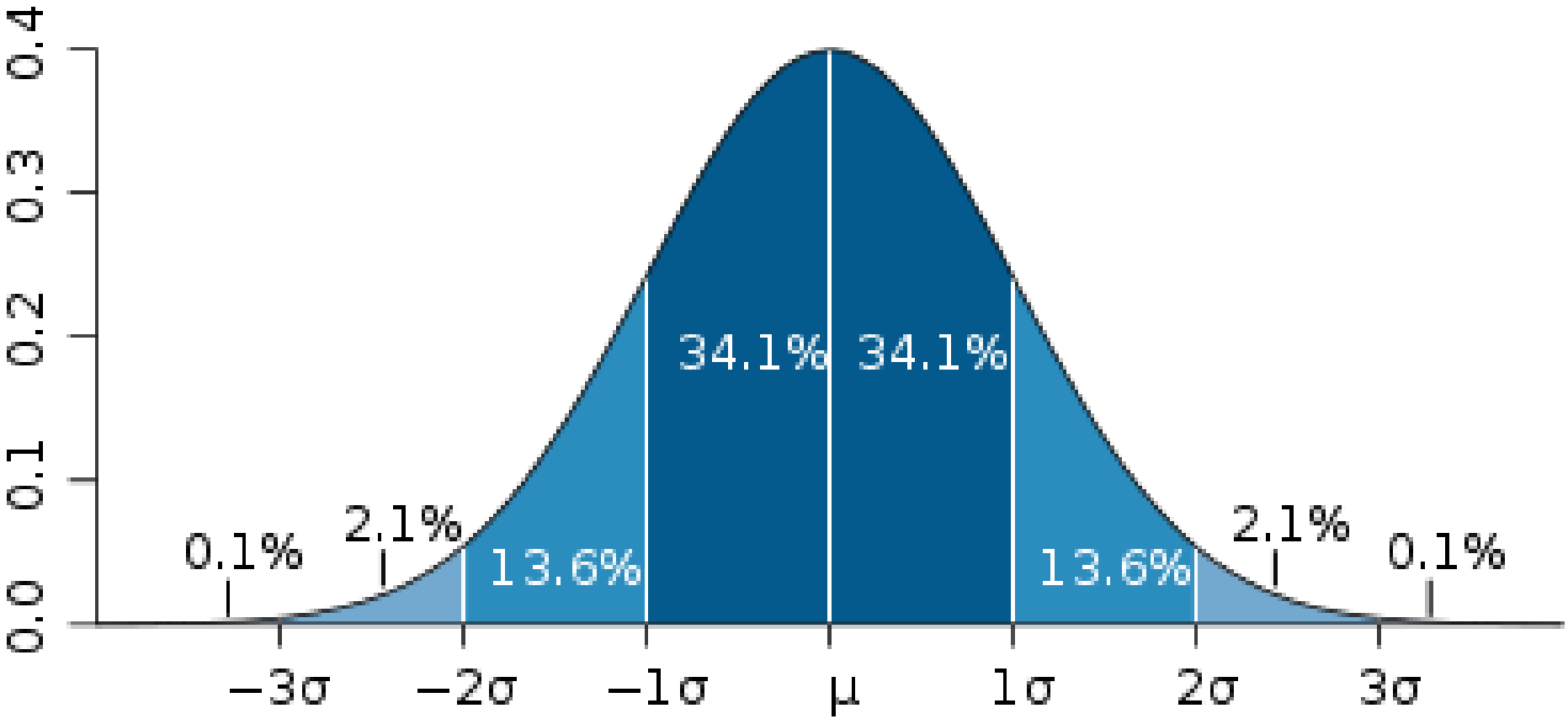
- about 68% of values drawn from a normal distribution are within one standard deviation σ away from the mean
- about 95% of the values lie within two standard deviations
- about 99.7% are within three standard deviations.

This fact is known as the 68-95-99.7 rule, or the empirical rule, or the 3-sigma rule.

Exercise

Assume that the variance of men's height is 100, while the variance of women's height is only 50. Then answer the question from the previous page.

Normal distribution – the area under the bell



Measures of location and spread

To describe a set (or sample) of data $x_1, \ldots, x_n$ we use several basic statistics that can characterize the location and the spread of the values.

Measures of location

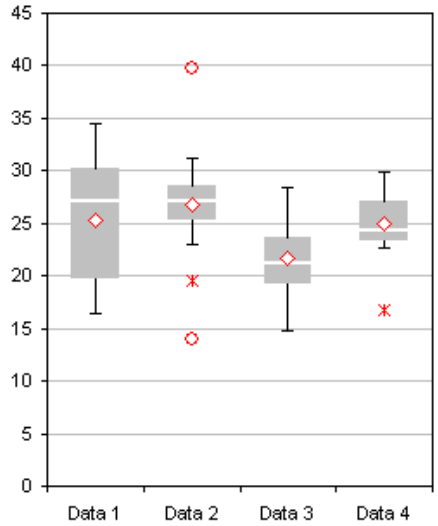
- Extreme values — minimum, maximum
- Sample mean (average) is $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$
- Other measures of central tendency — midrange, median, mode, quartiles
- Five number summary: Min, Q_1, Median, Q_3, Max

Measures of spread

- Range
- Inter-quartile range is $\mathrm{IQR} = Q_3 - Q_1$
- Sample variance is $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})$
- Sample standard deviation is $s = \sqrt{s^2}$

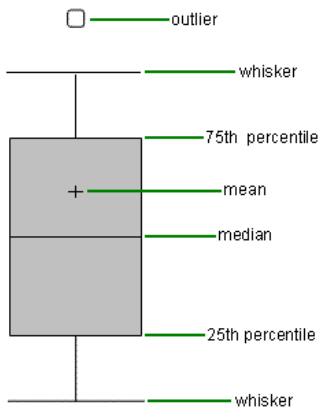
Box plots

In descriptive statistics, a boxplot (also known as a box-and-whisker diagram or plot) is a convenient way of graphically depicting groups of numerical data through their lower quartile (Q1), median (Q2), and upper quartile (Q3). A boxplot may also indicate which observations might be considered outliers.



The ends of the whiskers can represent several possible alternative values:

- the minimum and maximum of all the data
- the lowest datum still within 1.5 IQR of the lower quartile, and the highest datum still within 1.5 IQR of the upper quartile
- one standard deviation above and below the mean of the data



Homework

- Go through all exercises carefully and make sure that you are clear about their correct solution! So that you get ready for the tests.

- Install the R system on your home computer (freely available for both Linux or Windows).

Recommended literature

- Larry Gonick & Woollcott Smith: *The Cartoon Guide to Statistics*

Some easy exercises

Exercise

In a certain population, 30% of the persons smoke and 8% have a certain type of heart disease. Moreover, 12% of the persons who smoke have the disease.

a) What percentage of the population smoke and have the disease?

b) What percentage of the population with the disease also smoke?

Formulate this task and its solution in the terms of probability and conditional probability!

Some easy exercises

Exercise

Consider the experiment that consists of rolling 2 fair dice. Let X denote the first die score, and S denote sum of the scores. Now consider the following random events A, B, C, and D:

A ... $X = 3$

B ... $X = 2$

C ... $S = 5$

D ... $S = 7$

a) Find the probability of each event.

b) Find the conditional probabilities P(A|C), P(C|A), P(A|D), P(D|A), P(B|C), P(C|B).

c) Are some of the events A, B, C, D independent?

Some easy exercises

Exercise

In a town, 48% of all teenagers own a skateboard and 39% of all teenagers own a skateboard and roller blades. What is the probability that a teenager owns roller blades given that the teenager owns a skateboard?

Some easy exercises

Exercise

In a certain population, the probability a woman lives to at least seventy years is 0.70 and is 0.55 that she will live to at least eighty years. If a woman is seventy years old, what is the conditional probability she will survive to eighty years?

Some easy exercises

Exercise

In a survey, 85 percent of the employees say they favor a certain company policy. Previous experience indicates that 20 percent of those who do not favor the policy say that they do, out of fear of reprisal. What is the probability that an employee picked at random really does favor the company policy? It is reasonable to assume that all who favor say so.

Some easy exercises

Exercise

Two fair dice are rolled.

- a) What is the (conditional) probability that one turns up two spots, given they show different numbers?
- b) What is the (conditional) probability that the first turns up six, given that the sum is 10?
- c) What is the (conditional) probability that at least one turns up six, given that the sum is 10?

Some easy exercises

Exercise

Four persons are to be selected from a group of 12 people, 7 of whom are women. What is the probability that the first and third selected are women?

Some easy exercises

Exercise

A doctor assumes that a patient has one of three diseases d1, d2, or d3. Before any test, he assumes an equal probability for each disease. He carries out a test that will be positive with probability 0.8 if the patient has d1, 0.6 if he has disease d2, and 0.4 if he has disease d3. Given that the outcome of the test was positive, what probabilities should the doctor now assign to the three possible diseases?