Problem sheet 1

Defining the data

Problem 1.1

Define the types of the following variables as specific as possible:

a) monthly household income
b) socioeconomic status (high – middle – low)
c) state of residence
d) temperature in °C
e) height in cm
f) height (short – medium – tall)
g) zip code
h) BMI
i) accidents per month
j) hair colour
k) temperature (pleasant – unpleasant)
l) family size (number of people)
m) number of side effects per patient
n) dog breed

Problem 1.2

On Moodle you will find a sample data set of 40 English students from a statistics course named `StatsCourse.csv`. The following variables were collected:

`gender:`	female (`w`), male (`m`) or diverse (`d`)
`books:`	number of books read during the course
`attendance:`	number of lessons attended
`percentage grade:`	percentage grade of the course (0 – 100)
`grade:`	final (Bologna) grade for the statistics course (A – F)

Determine the variable types of the data set.

Problem 1.3

We all come into contact with statistics every day. Find examples of published statistics. Focus in particular on the following questions:

- Are there differences between statistics for categorical and numerical variables?
- Are there different basic types of statistics?
- Are there „good“ and „bad“ statistics? What are the criteria?

Discuss these questions in small groups before searching the world wide web. Find representative examples for different statistics which you can present to your colleagues.

Problem sheet 2

Introduction to probability theory

Problem 2.1

Determine the sample spaces for the following random experiments, as well as the respective concrete events A, B, C, D.

a) Single coin toss with event A: result is „head“

b) Throwing a die once with event B: result is a prime number

c) Throwing a die twice with event C: both numbers are the same

d) Measuring the height of a tower with event D: The measured height is between 120 m and 121 m (both values included)

Problem 2.2

In a quality control process, 4 light bulbs produced are checked in turn to see whether they are working (w) or defective (d). Determine the following events and their counter-events:

a) Exactly 3 light bulbs are defective.

b) The second light bulb is defective.

c) At least 2 light bulbs are defective.

d) The 2$^{\text{nd}}$ light bulb is working.

Hint: A single elementary event $\omega \in \Omega$ has the form $\omega = (\omega_1, \omega_2, \omega_3, \omega_4)$, where ω_i describes the state of the i^{th} light bulb. There are therefore a total of $2^4 = 16$ elementary events.

Problem 2.3

In a random experiment two dice are thrown simultaneously. (Assume that the numbers 1 to 6 have equal probability.)

a) Describe the sample space of the elementary events.

b) What is the probability of a single elementary event?

c) Calculate the probability of the event E_1 = „the sum of the spots is 7“.

d) What is the probability that event E_2 = „the sum of spots is smaller than 4“ occurs?

e) Determine $\mathrm{P}(E_3)$ of the event E_3 = „both spots are odd“.

f) Calculate $\mathrm{P}(E_2 \cup E_3)$.

Problem 2.4

a) You draw randomly one of the 26 letters of the alphabet. What is the probability that you draw a vowel?

b) You toss a fair coin three times. What is the probability that you get head at least once?

c) A wooden dice with a green surface is cut into $10^3 = 1000$ small dices of equal size. What is the probability for a randomly chosen (small) dice to have exactly two green cube faces?

Problem 2.5

A, B and C are events.

a) Which of the following statements are meaningful?

   i) $P(A \cup (B \cap C))$

   ii) $P(A) + P(B)$

   iii) $P(A^c) \cap P(B)$

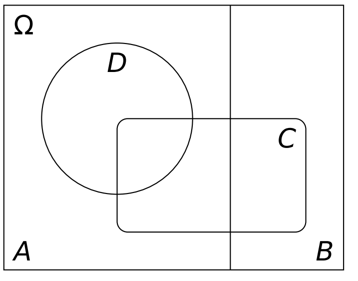
   iv) $(P(B))^c$

b) Display the following events in the given diagram.

   i) $C \cap D$

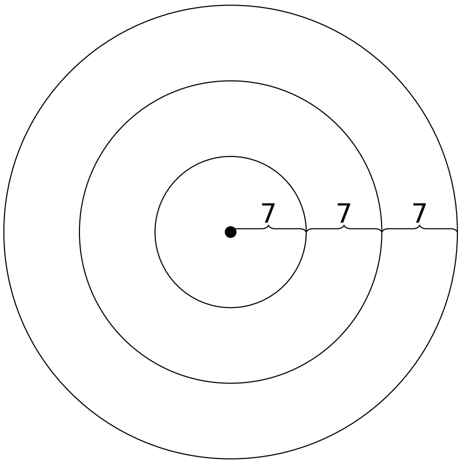
   ii) $(D \setminus C) \cup (C \cap A)$

   iii) $B \cup D$



Problem 2.6

You throw darts at a dartboard consisting of three circular discs (length specifications in centimeters):



On average, every second dart hits the dartboard. Further, the hit probabilities for the single circular discs are proportional to their areas. Let A, B, C describe the events „the innermost“, „the middle“, „the outermost circular disc is hit“. (Note: If the innermost circular disc is hit, then the middle as well as the outermost circular discs are hit simultaneously.)

a) Calculate the hit probability for the innermost, the middle and the outermost circular disc.

b) Calculate the probability to hit the middle circular *ring*.

Problem 2.7

a) In the canton of Bern, the telephone numbers have seven digits (not including the area code). How many telephone connections are possible if the digits 0 to 9 are available, but the first digit cannot be 0?

b) What is the probability that you will reach one of your 45 friends who still have a landline in the canton of Bern by calling a randomly dialed (correctly formed) telephone number?

Problem 2.8

a) Amy Farrah Fowler owns 4 dresses, 3 hats and 5 pairs of shoes. In how many ways ways can she dress to go out if everything matches and wearing a hat is i) mandatory and ii) optional?

b) Amy owns 6 sweaters and 5 pairs of pants. Of the sweaters, three are red, two are brown and one is green; of the pants, two are red, one is brown, one is blue and one is green. Amy reaches into the closet at random and gets dressed. What is the probability that her sweater and pants are the same color?

Problem 2.9

a) You and your 6 best friends go out for an after-work beer. How often do the glasses clink as everyone toasts with each other?

b) How many people are at a cocktail party when the glasses clink 253 times during the toast?

Problem 2.10

Calculate the following probabilities.

a) What is the probability to draw 4 „Bauern“ in the Swiss card game „Schieber“? („Schieber“: 4 people receive 9 cards each; of the total of 36 cards, exactly 4 are „Bauern“.)

b) Calculate the probabilities for 0,1,2 or 3 „Bauern“.

Problem 2.11

There are 8 lamps in a room that can be switched on and off independently of each other. How many types of lighting are there

a) in total?

b) with exactly 5 lamps burning?

c) with at least 5 lamps burning?

d) with a maximum of 3 lamps burning?

Problem 2.12

Calculate the probabilities for the following events.

a) A: In 4 throws with one die, at least one six falls.

b) B: In 24 rolls with two dice, a double six is rolled at least once.

Hint: Consider the counter-event in each case.

Problem 2.13

There are 23 people in a room. What is the probability that (ignoring the year of birth) at least two of these people have their birthday on the same day?

Assume that all birthdays are equally likely and that there are no dependencies between the different birthdays. You can disregard February 29.

(First estimate the order of magnitude of the result!)

Hint: First consider the counter-event, i.e. that all 23 people have birthdays on different dates.

Problem sheet 3

Conditional probability & independence

Problem 3.1

On Fridays Bonnie and Clyde often miss class. Bonnie is absent with probability 0.3 and Clyde with probability 0.45. The probability that both of them attend class is only 0.4. Are the attendances of Bonnie and Clyde independent events?

Problem 3.2

An unskilled hunter hits his target only with a probability of 20%. He fires three times at a rabbit. However, the rabbit knows the hunter. The rabbit remains quiet and sits still, as he believes that the probability of being shot is less than 50%. Has the rabbit calculated the probability correctly?

Problem 3.3

You roll a die. What is the probability that

a) you roll a six, given that the spots are even?

b) you roll a six, given that the spots are odd?

c) you roll even spots, given that the spots are smaller than four?

Problem 3.4

600 out of a sample of 900 people got vaccinated prophylactically against influenza. After a given time, it was analysed who actually got the flu. You find the results in the following table:

Group	B (sick)	B^c (healthy)	sum
A (vaccinated)	60	540	600
A^c (not vaccinated)	120	180	300
sum	180	720	900

We define the events A = „person is vaccinated“ and B = „person got sick“. Calculate the following probabilities and describe the meaning of the corresponding events.

a) $\mathrm{P}(A)$

b) $\mathrm{P}(B)$

c) $\mathrm{P}(A \cap B)$

d) $\mathrm{P}(B|A)$

e) $\mathrm{P}(A|B)$

f) $\mathrm{P}(A^c \cap B)$

g) $\mathrm{P}(B|A^c)$

Problem 3.5

30% of the Swiss population hold a bachelor or a comparable degree. In this population subgroup the unemployment rate is 5%. For the rest of the Swiss population the unemployment rate is 10%.

a) What is the probability for a Swiss to be unemployed?

b) From a specific Swiss guy you know that he is unemployed. What is the probability that he holds a bachelor or a comparable degree?

c) Are „holding a bachelor or comparable degree“ and „being unemployed“ independent events?

Problem 3.6

A spare part is manufactured on two machines: machine I delivers 2000 per day and machine II 5000. Experience shows that machine I has a 3.5% reject rate and machine II has a 1.5% reject rate. In a sample, a part is randomly selected from the total production of a period, and it is determined that this part is defective. What is the probability that the defective part was manufactured by machine I?

Problem 3.7

The dishwasher at a beverage company only washes 98% of the bottles cleanly. Therefore, the bottles are checked after washing. However, the checking machine only indicates 90% of the clean bottles as clean, thus erroneously sorting out 10% of the clean bottles. In addition, 5% of the bottles that are still dirty are classified as clean, thus erroneously not being sorted out.

a) What is the probability that a clean bottle will be rejected?

b) What is the probability that a bottle will be classified as clean?

c) What is the probability that a bottle that has not been rejected is actually clean?

Problem 3.8

The following headline was published in the ADAC-Motorwelt: *„Death rides along! Four out of ten fatally injured drivers were not wearing a seat belt!“*

This information is completely irrelevant – what would be interesting is the probability $P(\text{death}|\text{no seat belt})$. Calculate this assuming that 80% of drivers wear a seat belt and 10% of car accidents end fatally.

Problem sheet 4

Simple spam filter

Bayes' *theorem* establishes a relationship between the conditional probability $\mathrm{P}(A|B)$ and the reverse conditional probability $\mathrm{P}(B|A)$. This is useful because you can often estimate/measure conditional probability in one direction well, but are actually interested in the other direction (as we have already seen in some exercises).

For two events A and B with $P(A) \neq 0$ and $\mathrm{P}(B) \neq 0$ it follows from the definition of conditional probability

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

and the total probability formula

$$\mathrm{P}(A) = \mathrm{P}(B) \cdot \mathrm{P}(A|B) + \mathrm{P}(\overline{B}) \cdot \mathrm{P}(A|\overline{B})$$

namely:

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)} = \frac{\mathrm{P}(A|B) \cdot \mathrm{P}(B)}{\mathrm{P}(A|B) \cdot \mathrm{P}(B) + \mathrm{P}(A|\overline{B}) \cdot \mathrm{P}(\overline{B})}$$

This formula goes back to the mathematician Thomas Bayes (1702-1761) and is frequently used in medicine, for example. For example, if you know how often a disease occurs in general, how many people in the population smoke and how many of the people with the disease have smoked, you can simply calculate the probability that someone with the disease has also smoked.

Bayes' theorem can also be used to build filters that recognize spam emails. When emails were invented, nobody thought of people who wanted to send fake luxury watches, medication for eating disorders or stock scams... Sending spam is cheap and easy, and today it is assumed that around 90 percent of all emails sent fall into this category.

The idea behind a so-called Bayesian spam filter is surprisingly simple: the filter uses frequency information from past emails to estimate whether a new incoming email should be classified as spam or non-spam. The filter takes particular account of whether certain words appear in an e-mail. Intuitively, it is clear that certain words (such as „Viagra“ or „Rolex“) will occur significantly more frequently in a spam e-mail than in a non-spam e-mail.

To formulate our considerations in mathematical language, we define the following random events:

- S is the event that an incoming email is a spam email;
- E is the event that an incoming email contains a certain word w.

Using Bayes' theorem, we now obtain the following expression for the probability that an email containing the word w is a spam email

$$\mathrm{P}(S|E) = \frac{\mathrm{P}(E|S) \cdot \mathrm{P}(S)}{\mathrm{P}(E|S) \cdot \mathrm{P}(S) + \mathrm{P}(E|S^c) \cdot \mathrm{P}(S^c)}$$

The counter-event S^c here of course denotes the event that an incoming email is a normal email, i.e. non-spam.

For the probabilities $\mathrm{P}(S)$ and $\mathrm{P}(S^c)$, we use the above-mentioned empirical values 0.9 and $1-0.9 = 0.1$, which at least approximate these probabilities. (We will come back to these values later.)

Of course, we cannot exactly determine the probabilities $P(E|S)$ and $P(E|S^c)$ either. (For this, we would have to have complete information about all emails ever sent in the past and future). However, we can *estimate* these probabilities by examining a large number of emails. The larger this number, the more accurate our estimates of these probabilities will be, and the better our filter will work.

Specifically, we proceed as follows:

1. First, a large number of emails are collected and sorted (by hand) into the categories „spam“ and „non-spam“.

2. The emails in both categories are examined for the occurrence of certain words. The corresponding occurrence frequencies provide an *estimate* for the actual probabilities that a certain word occurs in a spam or non-spam email, i.e. for the conditional probabilities $P(E|S)$ and $P(E|S^c)$.

   We denote the resulting estimates for the unknown actual probabilities $P(x|y)$ as $\hat{p}(x|y)$ in order to clearly distinguish between these two things.

3. Using these estimates and Bayes' theorem, we can now (approximately) calculate the probability that a new incoming email containing a certain word is spam. The spam filter can be calibrated by setting a threshold value for this probability.

Example:

We looked at 2000 spam mails and 1000 non-spam mails and found, that the word „Rolex“ appears in 250 of the spam mails and in 5 of the non-spam mails. Should we discard an email containing the word „Rolex“ as spam or not?

The above frequencies result in the estimates $\hat{p}(E|S) = 250/2000 = 0.125$ and $\hat{p}(E|S^c) = 5/1000 = 0.005$. With the previously mentioned empirical values $P(S) = 0.9$ and $P(S^c) = 0.1$, this results in

$$\hat{p}(S|E) = \frac{0.125 \cdot 0.9}{0.125 \cdot 0.9 + 0.005 \cdot 0.1} \approx 0.996$$

According to our calculation, this email is therefore almost 100% likely to be spam and we will classify this email as spam. Whether we generally reject an email as spam or not depends on where we set the threshold value of our spam filter. If $\hat{p}(S|E)$ is greater than the selected threshold, we categorize the email as spam. The higher the selected threshold value, the more emails the filter will let through.

The following table is called *confusion matrix* or truth matrix and shows the four cases that can occur when classifying emails.

		actual category	
		spam	*non- Spam*
result of the spam filter	*spam*	correctly identified (true positive)	incorrectly identified (false positive)
	non-spam	falsely identified (false negative)	correctly identified (true negative)

Problem 4.1

You will now program and test a Bayesian spam filter using a prepared data set. This dataset was prepared at Hewlett-Packard Labs and contains data on 4601 emails that were classified as spam or non-spam by users. The data rows correspond to the individual emails and whose columns are to be interpreted as follows:

- The first four columns contain the occurrence frequencies of the following four words in the email (in percent): „will", „remove", „you" and "free" (in that order). The corresponding entry is

$$100 \cdot \frac{\text{number of occurrences of the word}}{\text{number of words in the entire email}}.$$

- The next two columns contain the frequency of occurrence of the following two characters in the email: Exclamation mark (!) and dollar sign ($). The corresponding entry is

$$100 \cdot \frac{\text{number of occurrences of the character}}{\text{number of characters in the entire email}}.$$

- The last column contains the classification into spam and non-spam: 1 stands for spam, 0 stands for non-spam.

On Moodle you will find the Jupyter notebook `spamfilter_template`, which loads the data automatically for you using the `ucimlrepo` package and then *randomly* splits it into two sets: a training data set (80% of the data) and a test data set (20% of the data). As the names suggest, the training dataset is to be used to develop/train our spam filter and the test dataset to test it afterwards. These two datasets must not be mixed!

The test dataset is therefore not required for the time being. For the training dataset, histograms are now created for the six tabulated frequencies, separately for the spam and non-spam emails. These provide a first impression of the characteristic features of spam emails.

Now it's your turn:

a) Determine the estimated values $\hat{p}(E|S)$, $\hat{p}(E|S^c)$ and $\hat{p}(S|E)$ for the event $E =$„contains at least 3% exclamation marks" analogous to the example above. So should we classify emails with the property E as spam if we use a threshold value of 98% for sorting?

   Note: The event E does not have the exact form „contains a certain word (at least once)" discussed above; however, this does not change the general procedure.

b) Now create the confusion matrix for the training data set with regard to the criterion E. How do you assess the result?

c) Now vary the percentage limit (i.e. the 3%) in the event E and try to improve your results from a) and b) in this way. (Continue to work only on the training data set.)

d) If you are satisfied with the result, now test the criterion you have found on the test data set. Does the result meet your expectations?

e) Now replace the estimates 0.9 and 0.1 for $P(S)$ and $P(S^c)$ with 0.5 in each case and repeat the previous steps. What do you find? Do you still get meaningful results?

f) Now try to further improve the spam filter yourself by defining and testing other criteria E. These can and should now include several columns of the data set, e.g. in the form „contains at least 3% exclamation marks and at least 1% dollar signs".

Problem sheet 5

Discrete random variables

Problem 5.1

You're being offered the following game: You roll a die and if the number of spots is even, you win the same amount in CHF – otherwise you lose the corresponding amount. Are you gonna play, i.e. do you expect to win?

Problem 5.2

The random variable X has the following probability mass function:

x_i	-3	-1	2	5
$P(X = x_i)$	$3a^2$	a	$5a$	$4a^2$

Calculate a.

Problem 5.3

A class has the following age distribution:

Age x_i	15	16	17	18
Frequency H_i	2	7	13	3

Determine the probability distribution for the age X of a randomly selected student.

Problem 5.4

You're being offered the following dice game with two tetrahedra (four-sided dice with number of spots 1 to 4): Show both of the dice the same number of spots, you get five times your stake (i.e. you win 4 times your stake). Otherwise you lose x times your stake, if the difference between the number of spots is x.

a) Calculate the expected value and the variance of your gain.

b) What would be a fair payout for the described game (indicated as multiple of your stake)?

Problem 5.5

An unfair coin with P(tail) = 1/3 and P(head) = 2/3 is tossed three times in succession. The random variable X assigns the frequency of „tails" to each result that is observed.

a) Determine the result space Ω and the probabilities of the elementary events.

b) Determine the value set of the random variable X and calculate the corresponding probability distribution.

c) Calculate the expected value $E(X)$ using the distribution you have just determined. Are you surprised by the result?

d) Now calculate the expected value $E(X)$ with the following more elegant consideration: Write X as the sum of three random variables X_i=„Frequency of tails in the ith throw", and use the linearity of the expected value.

Problem 5.6

Calculate the expected value, the variance and the standard deviation of the random variable X with the following probability distribution:

x_i	-6	-2	1	3
$P(X = x_i)$	$\frac{1}{2}$	$\frac{1}{8}$	$\frac{1}{4}$	$\frac{1}{8}$

Problem 5.7

a) We throw a fair die once. What is the expected value and variance of the result?

b) We throw a fair die twice and add the two numbers. What is the expected value and variance of the result?

c) We throw a fair die once and multiply the number thrown by 2. What is the expected value and variance of the result?

d) We throw a fair die once and add the number of points on the upper side to the number of points on the lower side. What is the expected value and variance of the result?

Problem 5.8

A fair die is thrown 100 times. Give a formula for each of the following probabilities and then calculate the actual values using `python`:

a) What is the probability that not a single six appears?

b) What is the probability that exactly one six appears?

c) What is the probability that at most 12 sixes appear?

d) What is the probability that at least 17 sixes appear?

Problem 5.9

A manipulated coin is tossed 100 times. The probability for heads is 0.52, and the probability for tails is therefore 0.48. Give a formula for each of the following probabilities and then calculate the actual values using `python`:

a) What is the probability that not a single head appears?

b) What is the probability that a head appears at least once?

c) What is the probability that head appears at most 45 times?

d) What is the probability that head appears at most 45 times on a fair coin?

Problem 5.10

Each member of a committee with 9 members comes to the meeting with a probability of 0.5. What is the probability that a two-thirds majority (i.e. six or more people) will be present?

You can assume that the individual members are present independently of each other.

Problem 5.11

A friend of you tosses a coin 12 times. He bets on getting exactly six times tail. What is his winning probability?

Problem 5.12

For an inspection, water samples (10 ml) are tested for contamination. As only 2% of all samples are contaminated, it's proposed to mix ten samples together. From every sample, 5 ml are mixed into a collective sample containing 50 ml. Now the collective sample is tested for contamination. If the sample is not contaminated, the inspection for the ten samples is over. Otherwise, the 10 single samples must be tested separately.

a) What is the probability that there is no contamination in the collective sample (50 ml) (assuming the ten single samples are independent)?

b) Let the random variable Y be the number of analyses needed. Which are the possible values for Y? Calculate the probability mass function of Y.

c) How many analyses need to be done on average (what is the expected value of Y)? How many analyses can be saved by mixing the samples into the collective sample, on average?

Problem 5.13

In a city two serious accidents happen per week on average. In particular, we assume that the number of serious accidents is Poisson distributed.

a) Calculate the probability that more than five serious accidents happen per week

b) Calculate the probability that more than one serious accident happens per *day*.

Problem 5.14

A technical support center is being called 12 times per hour on average. We assume that the number of calls is Poisson distributed. Calculate the probability that the support centre

a) gets no calls during the next five minutes.

b) gets at least 5 calls during the next 10 minutes.

c) gets at most 6 calls during the next 20 minutes.

Problem 5.15

A factory produces screws which are defective with probability $p = 0.001$. Calculate the probability that a batch of 500 screws contains at least two defective ones.

a) Calculate the *exact* probability using the binomial distribution.

b) Calculate the *approximate* probability using the Poisson distribution.

Problem 5.16

Plot in `python` the probability mass functions and the corresponding cdf's of the following distributions:

a) Binomial distribution for $n = 5$ and $\pi = 0.05$

b) Poisson distribution for $\lambda = 0.5$ (choose $x \in [0, 5]$ as plotting area)

c) Poisson distribution for $\lambda = 5$ (choose $x \in [0, 10]$ as plotting area)

d) Binomial distribution for $n = 500$ and $\pi = 0.01$ (choose $x \in [0, 10]$ as plotting area)

Problem 5.17

The probability that a passenger who has reserved a seat will not show up for the flight is 4%. The airline knows this and sells 75 tickets for 73 available seats. What is the probability that all seats are filled and no one has to wait for the next flight?

a) Solve the problem exactly with a binomial distribution.

b) Solve the problem approximately with a Poisson distribution.

Problem sheet 6

Continuous random variables

Problem 6.1

The continuous random variable X is uniformly distributed between 5 and 7, and thus has the density function

$$f_X(x) = \begin{cases} \frac{1}{2}, & \text{for } x \in [5,7] \\ 0, & \text{otherwise} \end{cases}$$

a) Determine the distribution function F_X for f_X and visualize both functions in a graph.

b) Calculate the probabilities $\mathrm{P}(X < 5.5)$ and $\mathrm{P}(5.8 < X < 6.8)$.

c) Calculate the mean and variance of X.

d) Calculate the median and the 90^{th} percentile of the distribution.

Problem 6.2

Let f_X be the function

$$f_X(x) = \frac{a}{1+x^2}$$

given, where $a \in \mathbb{R}_+$.

a) Determine the parameter a such that f_X is a density function.

   Note: $\int \frac{1}{1+x^2}\,dx = \arctan(x) + C$, $C \in \mathbb{R}$. Determine the distribution function F_X for f_X and visualize the two functions in a graph.

b) Calculate the following probabilities for a continuous random variable X that is distributed according to the above density function f_X:

$$\mathrm{P}\left(0 \le X \le \frac{1}{2}\right), \quad \mathrm{P}\left(X \le -\frac{1}{4}\right), \quad \mathrm{P}\left(X \ge \frac{3}{2}\right)$$

c) Calculate the median and the 90%-quantile of the distribution.

Problem 6.3

Let X be a (discretely or continuously distributed) random variable with standard deviation $0 < \sigma = \sigma(X) < \infty$ and expected value $\mu = \mathrm{E}(X) \in \mathbb{R}$. We consider the following transformed random variable:

$$Z := \frac{X - \mathrm{E}(X)}{\sigma(X)} = \frac{X-\mu}{\sigma}$$

Calculate $\mathrm{E}(Z)$, $\mathrm{Var}(Z)$ and $\sigma(Z)$ using the rules for calculating the expected value and variance.

Problem 6.4

Let X be a uniformly distributed random variable on $[0,1]$. Determine the density function f_Y of $Y := e^X$.

Problem 6.5

Let X be any continuous random variable with a distribution function F_X and a density function f_X. Determine the distribution and density function of the transformed random variable $Y = e^X$.

Problem 6.6

You get an email every 15 seconds on average. (For simplicity, assume that this rate is constant over time.)

a) What is the probability that you wait less than 30 seconds for your next email?

b) What is the maximum waiting time between two emails with 95% probability?

Problem 6.7

Car accidents occur on average 4 times a week.

a) Calculate the probability that more than 5 accidents occur in one week.

b) Calculate the probability that at least two weeks elapse between two accidents.

Problem 6.8

Calculate the following probabilities for a standard normally distributed random variable Z.

a) $P(-1 \leq Z \leq 1)$

b) $P(-2 \leq Z \leq 2)$

c) $P(-3 \leq Z \leq 3)$

d) $P(Z \leq 1)$

e) $P(|Z| \geq 0.5)$

f) $P(-3 \leq Z \leq 1)$

Problem 6.9

A pharmaceutical company wants to study women with an „exceptional“ height for their study. Specifically, they only want to include the 5% „most extreme“ heights in the study.

The height of women is approximately normally distributed with an expected value of 168 cm and a standard deviation of 6 cm. So what criterion must the female probands fulfill to be allowed to participate in the study?

Problem 6.10

It is assumed that the IQ of the total population is normally distributed with an expected value of 100 and a standard deviation of 15. We denote the corresponding random variable by X.

a) Specify an interval $[a, b]$ in which X lies with approximately 95% probability, i.e. for which $P(a \leq X \leq b) \approx 0.95$ applies.

b) Your highly gifted neighbor has an IQ of 145. What percentage of people are more intelligent (at least according to IQ tests) than your neighbor?

Problem 6.11

In the production of condensers, the capacity is a normally distributed random variable with expected value $\mu = 5$ [μF] and variance $\sigma^2 = 0.02^2$ [$(\mu\text{F})^2$]. What is the expected rejection rate, if the capacity

a) needs to be at least 4.98 μF?

b) is not allowed to be higher than 5.05 μF?

c) is allowed to differ from the nominal value $\mu = 5$ μF by maximally 0.03 μF?

Problem 6.12

A workpiece is of the desired quality if the deviation of its mass does not deviate more than 3.45 mm from the nominal value. The random deviations from the nominal value are normally distributed with standard deviation $\sigma = 3$ mm. There are no systematic deviations, i.e. the expected value of the deviation is exactly 0. Determine the expected number of workpieces with the desired quality if 24 workpieces are produced.

Solutions

Solution 1.1

a) monthly household income: numerical (continuous)

b) socioeconomic status (high – middle – low): ordered categorical

c) state of residence: nominal categorical

d) temperature in °C: numerical (continuous)

e) height in cm: numerical (continuous)

f) height (short – medium – tall): ordered categorical

g) zip code: nominal categorical / ordered categorical

h) BMI: numerical (continuous)

i) accidents per month: numerical (discrete)

j) hair colour: nominal categorical / ordered categorical via wave length

k) temperature (pleasant – unpleasant): binary

l) family size (number of people): numerical (discrete)

m) number of side effects per patient: numerical (discrete)

n) dog breed: nominal categorical

Solution 1.2

- `gender`: categorical nominal
- `books`: numerical discrete
- `attendance`: numerical discrete
- `percentage grade`: numerical continuous
- `grade`: categorical ordinal

Solution 1.3

see discussion in class

Solution 2.1

a) $\Omega_A = \{„\text{head}“, „\text{tail}“\}$ with $A = \{„\text{head}“\}$

b) $\Omega_B = \{1, 2, 3, 4, 5, 6\}$ with $B = \{2, 3, 5\}$

c) $\Omega_C = \{(1, 1), (1, 2), (1, 3), \dots, (5, 6), (6, 6)\}$ (Ω has $6^2 = 36$ elements) with $C = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}$

d) $\Omega_D =]0, \infty[$ with $D = [120, 121]$

Solution 2.2

a) $A = \{(d, d, d, w), (d, d, w, d), (d, w, d, d), (w, d, d, d)\}$
$A^c = \{(d, d, d, d), (w, w, d, d), (w, d, w, d), (w, d, d, w), (d, w, w, d), (d, w, d, w), (d, d, w, w), (w, w, w, d), (w, w, d, w), (w, d, w, w), (d, w, w, w), (w, w, w, w)\}$

To save us a little paperwork, we will abbreviate d/w for a defective *or* working light bulb. Of course, each of the 16 elementary events always occurs either in (e.g.) B or the complement B^c; spelled out, we would therefore have to list all 16 elementary events each time as in a).

b) $B = \{(d/w, d, d/w, d/w)\}$ and $B^c = \{(d/w, w, d/w, d/w)\}$

c) $C = \{(d, d, d/w, d/w), (d, d/w, d, d/w), (d, d/w, d/w, d), (d/w, d, d/w), (d/w, d, d/w, d), (d/w, d/w, d, d)\}$
$C^c = \{(w, w, w, d/w), (w, w, d/w, w), (w, d/w, w, w), (d/w, w, w, w)\}$

d) $D = B^c$ and therefore $D^c = B$

Solution 2.3

If the elementary events have equal probability we can calculate the probability of event A as the number of elementary events in A divided by the total number of elementary events.

a) $\Omega = \{(1, 1), (1, 2), \ldots, (1, 6), (2, 1), (2, 2), \ldots, (2, 6), \ldots, (6, 6)\}$, $|\Omega| = 36$.

b) $\mathrm{P}(\{\text{elementary event}\}) = \frac{1}{|\Omega|} = \frac{1}{36}$.

c) $E_1 = \{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\}$;
number of favourable cases: $|E_1| = 6$;
number of possible cases: $|\Omega| = 36$;
$\mathrm{P}(E_1) = \frac{|E_1|}{|\Omega|} = \frac{6}{36} = \frac{1}{6}$.

d) $E_2 = \{(1, 1), (2, 1), (1, 2)\}$;
$\mathrm{P}(E_2) = \frac{|E_2|}{|\Omega|} = \frac{3}{36} = \frac{1}{12}$.

e) $E_3 = \{(1, 1), (1, 3), (1, 5), (3, 1), (3, 3), (3, 5), (5, 1), (5, 3), (5, 5)\}$;
$\mathrm{P}(E_3) = \frac{|E_3|}{|\Omega|} = \frac{9}{36} = \frac{1}{4}$.

f)

$$\begin{aligned} \mathrm{P}(E_2 \cup E_3) &= \mathrm{P}(E_2) + \mathrm{P}(E_3) - \mathrm{P}(E_2 \cap E_3) \\ &= \mathrm{P}(E_2) + \mathrm{P}(E_3) - \mathrm{P}(\{(1, 1)\}) \\ &= \frac{3}{36} + \frac{9}{36} - \frac{1}{36} = \frac{11}{36}. \end{aligned}$$

Solution 2.4

a) $\mathrm{P}(A) = 5/26 \approx 0.19$

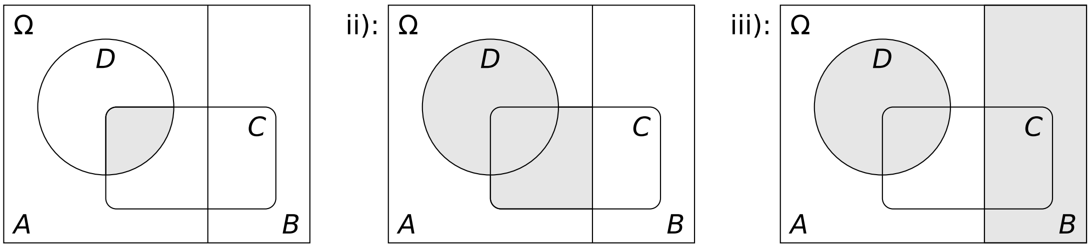
b) Out of the $2^3 = 8$ (equally probable) elementary events 7 are favourable, so, $\mathrm{P}(B) = 7/8 = 0.875$.

c) For each of the 12 cube borders, only 8 dice have exactly two green faces, so, $\mathrm{P}(C) = (12 \cdot 8)/1000 = 0.096$.

Solution 2.5

a) The operators union ($\cup$), intersection ($\cap$) and complement (^{c}) operate on sets and the addition (+) on numbers. We get that i) and ii) are meaningful and iii) and iv) are not.

b) i): ii): iii):



Solution 2.6

a) We can easily see that $A \subseteq B \subseteq C$ and $A \cup B \cup C = C$. Using the fact that the hit probabilities are proportional to the areas on the dartboard, the ratio of the probabilities must be $7^2\pi : 14^2\pi : 21^2\pi = 1 : 4 : 9$. Additionally, we know that only every second dart hits the board, therefore, $\mathrm{P}(C) = 0.5$. Consequently,

$$\mathrm{P}(B) = \frac{4}{9} \cdot \mathrm{P}(C) = \frac{2}{9}, \qquad \mathrm{P}(A) = \frac{1}{9} \cdot \mathrm{P}(C) = \frac{1}{18} .$$

b) The probability to hit the middle circular ring is $\mathrm{P}(B \setminus A) = \mathrm{P}(B) - \mathrm{P}(A) = \frac{2}{9} - \frac{1}{18} = \frac{1}{6}$, using the fact that $A \subseteq B$.

Solution 2.7

a) We have 10 possibilities for each digit of the telephone number; however, there are only 9 for the first digit. We get $9 \cdot 10^6$, i.e. 9 million different telephone numbers.

b) The probability we are looking for is $45/(9 \cdot 10^6) = 5 \cdot 10^{-6}$.

Solution 2.8

a) i) $4 \cdot 3 \cdot 5 = 60$

ii) $4 \cdot (3+1) \cdot 5 = 80$

b) Of the $6 \cdot 5 = 30$ (equally probable!) combinations, $3 \cdot 2 + 2 \cdot 1 + 0 \cdot 1 + 1 \cdot 1 = 9$ are favorable. The required probability is therefore $9/30 = 0.3$.

Solution 2.9

a) The glasses clink $\binom{7}{2} = \frac{7 \cdot 6}{2} = 21$ times.

b) 23 people, because $\binom{23}{2} = \frac{23 \cdot 22}{2} = 253$.

Solution 2.10

a) The number of possible hands you can get is $\binom{36}{9}$, of which only $\binom{32}{5}$ contain the 4 required „Bauern“ as the 5 remaining cards can be any from the 32 remaining cards. The probability of having 4 „Bauern“ is therefore approx. 0.002 = 0.2%.

b)
- Probability for 0 „Bauern“: The number of favourable hands in this case is $\binom{32}{9}$. For the probability we are looking for, we get $0.298 \approx 30\%$.
- Probability for 1 „Bauern“: The number of favourable hands in this case is $\binom{4}{1} \cdot \binom{32}{8}$. The probability therefore is $0.447 \approx 45\%$.
- Probability for 2 „Bauern“: Analogously, here the number of favourable hands is $\binom{4}{2} \cdot \binom{32}{7}$. The probability therefore is $0.215 \approx 22\%$.
- Probability for 3 „Bauern“: Analogously, the number of favourable hands is $\binom{4}{3} \cdot \binom{32}{6}$. The probability therefore is $0.039 \approx 4\%$.

Of course, all 5 calculated probabilities add up to 100%!

Solution 2.11

a) Each of the 8 lamps can be switched on or off. So there are a total of $2 \cdot \ldots \cdot 2 = 2^8 = 256$ lighting types.

b) There are $\binom{8}{5} = 56$ types of lighting in which exactly 5 lamps are burning.

c) There are $\binom{8}{5} + \binom{8}{6} + \binom{8}{7} + \binom{8}{8} = 56 + 28 + 8 + 1 = 93$ types of lighting in which at least 5 lamps are burning.

d) There are $\binom{8}{3}+\binom{8}{2}+\binom{8}{1}+\binom{8}{0} = 56+28+8+1 = 93$ types of lighting in which at most 3 lamps are lit; compare with c).

Solution 2.12

a) With the counter-event A^c = „All 4 throws are at most 5“ we get

$$P(A) = 1 - P(A^c) = 1 - \frac{5^4}{6^4} \approx 0.52.$$

b) With the counter-event B^c = „All 24 throws are not double six“ we get

$$P(B) = 1 - P(B^c) = 1 - \frac{35^{24}}{36^{24}} \approx 0.49.$$

Solution 2.13

There are 365^{23} equally probable ways in which the birthdays can be distributed. The number of possibilities in which all 23 people have birthdays on different dates is $365 \cdot 364 \cdot \ldots \cdot 344 \cdot 343$ (23 factors in total).

The probability we are looking for is therefore

$$1 - \frac{365 \cdot 364 \cdot \ldots \cdot 343}{365^{23}} \approx 0.51,$$

a little more than 50 %. Would you have expected that?

Solution 3.1

P(„Bonnie attends class“) $=$ P$(B) = 0.7$ and P(„Clyde attends class“) $=$ P$(C) = 0.55$. We know that

$$\mathrm{P}(B \cap C) = 0.4 \neq 0.7 \cdot 0.55 = 0.385.$$

So, the attendances of Bonnie and Clyde are *dependent* events.

Solution 3.2

P(„rabbit is shot“) $= 1 -$ P(„rabbit survives“) $= 1 - 0.8 \cdot 0.8 \cdot 0.8 = 1 - 0.512 = 0.488$. So, the rabbit was correct – the probability of being shot is less than 50%!

Solution 3.3

a) $\mathrm{P}(6\,|\,\text{„even spots“}) = \frac{\mathrm{P}(6 \cap \text{„even spots“})}{\mathrm{P}(\text{„even spots“})} = 2 \cdot \frac{1}{6} = \frac{1}{3}$

b) $\mathrm{P}(6\,|\,\text{„odd spots“}) = \frac{\mathrm{P}(6 \cap \text{„odd spots“})}{\mathrm{P}(\text{„odd spots“})} = \frac{0}{1/2} = 0$

c) $\mathrm{P}(\text{„even spots“}\,|\,\text{„spots<4“}) = \frac{\mathrm{P}(\text{„even spots“} \cap \text{„spots<4“})}{\mathrm{P}(\text{„spots<4“})} = \frac{1}{6} / \frac{1}{2} = \frac{1}{3}$

Solution 3.4

a) $\mathrm{P}(A) = \frac{600}{900} \approx 0.667$: The probability that a randomly selected person from the sample is vaccinated is approx. 0.667.

b) $\mathrm{P}(B) = \frac{180}{900} = 0.2$: The probability that a randomly selected person from the sample got the flu is 0.2.

c) $\mathrm{P}(A \cap B) = \frac{60}{900} \approx 0.067$: The probability that a randomly selected person from the sample got sick, even though he or she is vaccinated, is approx. 0.067.

d) $P(B|A) = \frac{P(A\cap B)}{P(A)} = \frac{60}{600} = 0.1$: Given that a person is vaccinated, he or she got sick with probability 0.1.

e) $P(A|B) = \frac{P(A\cap B)}{P(B)} = \frac{60}{180} \approx 0.333$: Given that a person got the flu, he or she is vaccinated with probability 0.333 (approx.)

f) $P(A^c \cap B) = \frac{120}{900} \approx 0.133$: The probability that a randomly selected person from the sample is not vaccinated and got sick is approx. 0.133.

g) $P(B|A^c) = \frac{P(A^c\cap B)}{P(A^c)} = \frac{120}{300} = 0.4$: Given that a person is not vaccinated, he or she got the flu with probability 0.4.

Solution 3.5

Let us define the following events: B =„Swiss resident holds a bachelor or comparable degree“ and E =„Swiss resident has a job“. Therefore, the following information is known:

$P(B) = 0.3$, also $P(B^c) = 0.7$, as well as $P(E^c|B) = 0.05$ and $P(E^c|B^c) = 0.1$.

a) $P(E^c) = P(E^c|B) \cdot P(B) + P(E^c|B^c) \cdot P(B^c) = 0.05 \cdot 0.3 + 0.1 \cdot 0.7 = 0.085$

b) $P(B|E^c) = \frac{P(B\cap E^c)}{P(E^c)} = \frac{P(E^c|B)\cdot P(B)}{P(E^c)} = \frac{0.05\cdot 0.3}{0.085} \approx 0.1765$

c) No, because $P(B \cap E^c) = 0.05 \cdot 0.3 = 0.015 \neq P(B) \cdot P(E^c) = 0.3 \cdot 0.085 = 0.0255$.

Solution 3.6

$$P(\text{„machine I“}|\text{„defective“}) = \frac{P(\text{„machine I“} \cap \text{„defective grqq})}{P(\text{„defective“})} = \frac{2/7 \cdot 0.035}{2/7 \cdot 0.035 + 5/7 \cdot 0.015} \approx 0.483$$

Solution 3.7

a) $P(\text{„rejected“}|\text{„clean“}) = 0.1$

b) $P(\text{„not rejected“}) = 0.02 \cdot 0.05 + 0.98 \cdot 0.9 = 0.883$

c) $P(\text{„clean“}|\text{„not rejected“}) = \frac{P(\text{„clean“}\cap\text{„not rejected“})}{P(\text{„not rejected“})} = \frac{0.98\cdot 0.9}{0.02\cdot 0.05+0.98\cdot 0.9} \approx 0.999$

Solution 3.8

We know that P(no seat belt|death) = 4/10 = 0.4. However, we are interested in the conditional probability P(death|no seat belt) – that is, the inverse. The relationship between these two conditional probabilities is given the so-called Bayes' theorem. With

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

it follows that $P(A \cap B) = P(A|B) \cdot P(B)$. Now, inserting this into the definition of $P(B|A)$ gives the **Bayes' Theorem**, which states

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

It follows:

$$P(\text{death}|\text{no seat belt}) = \frac{P(\text{no seat belt}|\text{death}) \cdot P(\text{death})}{P(\text{no seat belt})} = \frac{4/10 \cdot 1/10}{2/10} = 0.2 = 20\%$$

The probability of a fatal car accident if you don't wear a seat belt doubles!

Solution 4.1

a) See file `spamfilter_solution.ipynb` on Moodle. Because of $\hat{p}(S|E) \approx 0.9743 < 0.98$ the criterion E (barely) does not give us enough certainty to classify an email as spam.

b) See file `spamfilter_solution.ipynb` on Moodle.

c) If you set the limit at 1% instead of 3%, you get $\hat{p}(S|E) \approx 0.9879$; this is now a criterion that mathematically gives us the required 98% certainty. At the same time, it is clear that we will of course sort out more mails as spam than in a), i.e. we must also expect more false positives. (This is also clearly recognizable in the confusion matrix.)

d) —

e) —

f) —

Solution 5.1

Let X be your profit with possible values -5, -3, -1, 2, 4, 6. Obviously, $\mathrm{P}(X = x_i) = \frac{1}{6}$ for $i = 1, 2, \ldots, 6$. Therefore, $\mathrm{E}(X) = \frac{1}{6}(-5-3-1+2+4+6) = 0.5$. Of course you're gonna play – on average you win 0.5 CHF per game!

Solution 5.2

The sum of the single probabilities must be 1 in order to define a probability mass function, therefore: $7a^2 + 6a = 1$. It follows that $a = \frac{1}{7}$ (the solution -1 of the quadratic equation does obviously not make sense).

Solution 5.3

Age x_i	15	16	17	18
$\mathrm{P}(X = x_i)$	0.08	0.28	0.52	0.12

Solution 5.4

a) It is a Laplace experiment with 16 elementary events (i, k) for $i, k = 1, 2, 3, 4$. Your gain is denoted by the random variable X with possible values $4, -1, -2, -3$ and

$$\begin{aligned}
\{X = 4\} &= \{(1,1), (2,2), (3,3), (4,4)\} \\
\{X = -1\} &= \{(1,2), (2,1), (2,3), (3,2), (3,4), (4,3)\} \\
\{X = -2\} &= \{(1,3), (3,1), (2,4), (4,2)\} \\
\{X = -3\} &= \{(1,4), (4,1)\}
\end{aligned}$$

It follows that $\mathrm{E}(X) = 4 \cdot \frac{4}{16} - 1 \cdot \frac{6}{16} - 2 \cdot \frac{4}{16} - 3 \cdot \frac{2}{16} = -\frac{1}{4}$ and $\mathrm{Var} = \left(4 + \frac{1}{4}\right)^2 \cdot \frac{4}{16} + \left(-1 + \frac{1}{4}\right)^2 \cdot \frac{6}{16} + \left(-2 + \frac{1}{4}\right)^2 \cdot \frac{4}{16} + \left(-3 + \frac{1}{4}\right)^2 \cdot \frac{2}{16} \approx 6.438$.

b) $\mathrm{E}(X) = x \cdot \frac{4}{16} - 1 \cdot \frac{6}{16} - 2 \cdot \frac{4}{16} - 3 \cdot \frac{2}{16} = 0 \implies x = 5$, i.e. the payout would need to be 6 times your stake.

Solution 5.5

a) $\Omega = \{(T,T,T), (T,T,H), (T,H,T), (H,T,T), (T,H,H), (H,T,H), (H,H,T), (H,H,H)\}$ with associated probabilities $\mathrm{P}((T,T,T)) = \left(\frac{1}{3}\right)^3 = \frac{1}{27}$, $\mathrm{P}((T,T,H)) = \mathrm{P}((T,H,T)) = P((H,T,T)) = \left(\frac{1}{3}\right)^2 \cdot \frac{2}{3} = \frac{2}{27}$, $\mathrm{P}((T,H,H)) = \mathrm{P}((H,T,H)) = \mathrm{P}((H,H,T)) = \frac{1}{3} \cdot \left(\frac{2}{3}\right)^2 = \frac{4}{27}$ and $\mathrm{P}((H,H,H)) = \left(\frac{2}{3}\right)^3 = \frac{8}{27}$.

b) The random variable X can take the values 0,1,2,3. The probability distribution is according to a) $\mathrm{P}(X=0) = \left(\frac{2}{3}\right)^3 = \frac{8}{27}$, $\mathrm{P}(X=1) = 3 \cdot \left(\frac{2}{3}\right)^2 \cdot \frac{1}{3} = \frac{4}{9}$, $\mathrm{P}(X=2) = 3 \cdot \frac{2}{3} \cdot \left(\frac{1}{3}\right)^2 = \frac{2}{9}$ and $\mathrm{P}(X=3) = \left(\frac{1}{3}\right)^3 = \frac{1}{27}$.

c) $E(X) = 0 \cdot \frac{8}{27} + 1 \cdot \frac{4}{9} + 2 \cdot \frac{2}{9} + 3 \cdot \frac{1}{27} = 1$. The result is in line with our intuition: In 3 throws we will observe on average $1/3 \cdot 3 = 1$ times "'tail"'. See also subtask d).

d) With X_i=„frequency of tails in the ith throw" (= 0 or 1), $E(X) = E(X_1 + X_2 + X_3) = E(X_1) + E(X_2) + E(X_3) = \frac{1}{3} + \frac{1}{3} + \frac{1}{3} = 1$. (Random variables that can only take the value 0 or 1 are also called *indicator variables*; as you can easily see, for indicator variables $E(X) = P(X = 1)$).

Solution 5.6

$E(X) = -2.625$, $Var(X) \approx 12.984$ and $\sigma(X) \approx 3.603$

Solution 5.7

a) Let X be the number of points the die shows. Therefore, $E(X) = \frac{1}{6} \cdot (1+2+3+4+5+6) = \frac{7}{2} = 3.5$ and $Var(X) = \frac{1}{6} \cdot ((1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2) = \frac{35}{12} \approx 2.9$

b) Let X_1 be the number of the first roll and X_2 the number of the second roll. Then $E(X_1 + X_2) = E(X_1) + E(X_2) = 2 \cdot E(X_1) = 2 \cdot 3.5 = 7$. Since the two casts are independent, the following applies to the variance: $Var(X_1 + X_2) = Var(X_1) + Var(X_2) = 2 \cdot Var(X_1) \approx 2 \cdot 2.9 = 5.8$.

c) In this case, let X be the number of points of a roll. We are interested in the expected value and variance of $2 \cdot X$. According to the calculation rules for expected values, $E(2 \cdot X) = 2 \cdot E(X) = 7$ and $Var(2 \cdot X) = 4 \cdot Var(X) \approx 11.6$.

d) Obviously, the sum of the points *always* is 7; the expected value is therefore 7 and the variance 0. More formally: Let X_1 be the number of points at the top and X_2 the number of points at the bottom of the die. For the expected value, you can either calculate as in b) or observe that $X_1 + X_2 = 7$ always applies and therefore $E(X_1+X_2) = E(7) = 7$ follows. (The second variant would be incorrectly argued for b). Similarly, $Var(X_1 + X_2) = Var(7) = 0$.

Note: The expected value is always 7 in all three examples; the variance, however, is different each time.

Solution 5.8

Let the random variable X be the number of sixes in 100 throws of a fair die. Then X is binomially distributed with $n = 100$ and $\pi = \frac{1}{6}$.

a) $P(X = 0) = \left(\frac{5}{6}\right)^{100}$

b) $P(X = 1) = 100 \cdot \left(\frac{1}{6}\right) \cdot \left(\frac{5}{6}\right)^{99}$

c) $P(X \leq 12) = \sum_{k=0}^{12} \binom{100}{k} \cdot \left(\frac{1}{6}\right)^k \cdot \left(\frac{5}{6}\right)^{100-k}$

d) $P(X \geq 17) = 1 - P(X \leq 16) = 1 - \sum_{k=0}^{16} \binom{100}{k} \cdot \left(\frac{1}{6}\right)^k \cdot \left(\frac{5}{6}\right)^{100-k}$

In python, we calculate the solutions as follows:

```
# import libraries
#--------------------------------------
from scipy.stats import binom, poisson
import matplotlib.pyplot as plt
import numpy as np
#--------------------------------------
>>> n, p = 100, 1/6
>>> binom.pmf(0,n,p)
1.2074673472413602e-08
>>> binom.pmf(1,n,p)
2.4149346944827303e-07
>>> binom.cdf(12,n,p)
0.1296708011552209
```

```
>>> 1 - binom.cdf(16,n,p)
0.5058410243147228
```

Solution 5.9

Let the random variable X be the number of heads in a 100-times coin toss.

a) $P(X=0)=(0.48)^{100}$

b) $P(X\geq 1)=1-P(X=0)=1-(0.48)^{100}$

c) $P(X\leq 45)=\sum_{k=0}^{45}\binom{100}{k}\cdot(0.52)^k\cdot(0.48)^{100-k}$

d) $\sum_{k=0}^{45}\binom{100}{k}\cdot(0.5)^k\cdot(0.5)^{100-k}=\sum_{k=0}^{45}\binom{100}{k}\cdot(0.5)^{100}$

In python we find the following values:

```
>>> n, p = 100, 0.52
>>> binom.pmf(0,n,p)
1.3308335400789563e-32
>>> 1 - binom.pmf(0,n,p)
1.0
>>> binom.cdf(45,n,p)
0.09665335032776938
>>> p = 0.5
>>> binom.cdf(45,n,p)
0.18410080866334788
```

Solution 5.10

The number of people present is binomially distributed with parameters 9 and 0.5. The probability you are looking for is given by $\sum_{k=6}^{9}\binom{9}{k}(0.5)^9=\frac{\sum_{k=6}^{9}\binom{9}{k}}{2^9}\approx 0.2539$.

Solution 5.11

The count of tail is binomially distributed with parameters 12 and 0.5. The winning probability is therefore $\binom{12}{6}\cdot 0.5^{12}\approx 0.2256$.

Solution 5.12

a) Let X be the number of contaminated samples in one collective sample. The probability that a sample is contaminated is $\pi=0.02$. Under the assumption that all samples are independent, X is binomially distributed: $X\sim \text{Bin}(n=10,\pi=0.02)$.

   The probability to find no contamination in the sample is given by

$$P(X=0)=\binom{10}{0}\cdot 0.02^0\cdot 0.98^{10}=0.98^{10}=0.8171.$$

   In python we calculate $P(X=0)$ as

```
>>> n, p = 10, 0.02
>>> binom.pmf(0,n,p)
0.8170728068875469
```

   Another possible solution: Each sample is clean with a probability of 0.98, independently of the other samples. Therefore we have

$$P(\text{all samples are clean})=\prod_{i=1}^{10}P(i\text{-th sample is clean})=0.98^{10}=0.8171.$$

b) The random variable Y can only have the values 1 or 11, because:

   i) If all samples are clean, we are done after one analysis: $Y = 1$.

   ii) If at least one sample is contaminated, then the collective sample is contaminated and we need to check all 10 samples separately: $Y = 11$.

   Hence,

$$\begin{aligned} P(Y=1) &= P(\text{no sample is contaminated}) = 0.8171\,, \\ P(Y=11) &= 1 - P(Y=1) = 0.1829\,. \end{aligned}$$

c) The average number of analyses for one collective sample is given through the expected value of Y:

$$E(Y) = \sum_{k=0}^{\infty} k \cdot P(Y=k) = 1 \cdot P(Y=1) + 11 \cdot P(Y=11) = 1 \cdot 0.8171 + 11 \cdot 0.1829 = 2.8293\,.$$

   On average, we save $10 - 2.8293 = 7.1707 \approx 7$ analyses.

Solution 5.13

a) Let X be the number of serious accidents per week: $X \sim \text{Po}(2)$ and the probability is $P(X > 5) = 1 - P(X \leq 5) \approx 0.017$.

b) Let Y be the number of serious accidents per day: $Y \sim \text{Po}(2/7)$ and the probability is $P(Y > 1) = 1 - P(Y \leq 1) \approx 0.034$.

```
>>> lmd = 2
>>> 1-poisson.cdf(5,lmd)
0.016563608480614445
>>> 1-poisson.cdf(1,lmd/7)
0.03381490890320382
```

Solution 5.14

a) The number of calls per 5 minutes X is Poisson distributed with parameter $\lambda_X = \frac{12}{60} \cdot 5 = 1$. Therefore, the probability is $P(X = 0) \approx 0.368$.

b) The number of calls per 10 minutes Y is Poisson distributed with parameter $\lambda_Y = 2 \cdot \lambda_X = 2$. Therefore, the probability is $P(Y \geq 5) = 1 - P(Y < 5) = 1 - P(Y \leq 4) \approx 0.053$.

c) The number of calls per 20 minutes Z is Poisson distributed with parameter $\lambda_Z = 4 \cdot \lambda_X = 2 \cdot \lambda_Y = 4$. Therefore, the probability is $P(Z \leq 6) \approx 0.889$

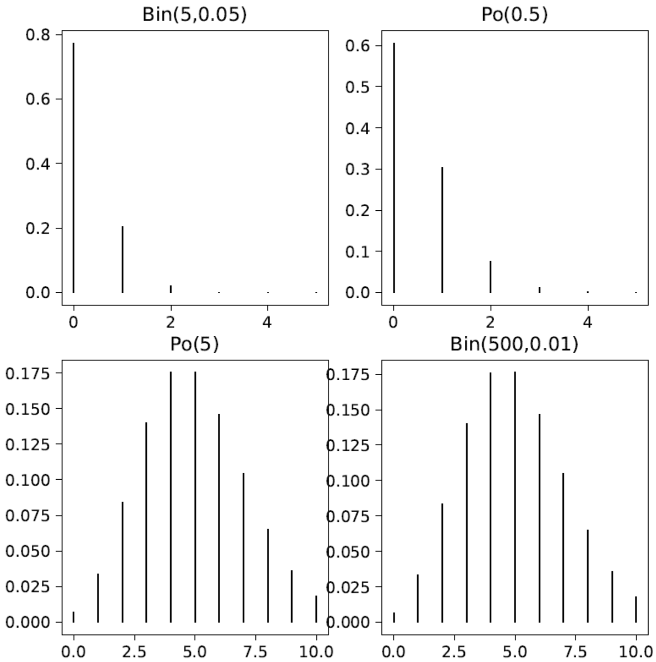
Solution 5.15

a) The number of defective screws X is binomially distributed with parameters $n = 500$ and $\pi = 0.001$:

$$P(X \geq 2) = 1 - P(X < 2) = 1 - F_X(1) = 1 - \sum_{k=0}^{1} \binom{500}{k} 0.001^k \cdot 0.999^{500-k} \approx 0.0901$$

b) The number of defective screws is approximately Poisson distributed with parameter $n \cdot \pi = 500 \cdot 0.001 = 0.5$. Therefore, $P(X \geq 2) \approx 1 - \sum_{k=0}^{1} e^{-0.5} \frac{0.5^k}{k!} \approx 0.0902$.

Solution 5.16



The corresponding python-code is:

```
fig, axs = plt.subplots(2, 2)

# creating a numy array for x-axis
x1 = np.arange(0, 6, 1)
x2 = np.arange(0, 6, 1)
x3 = np.arange(0, 11, 1)
x4 = np.arange(0, 11, 1)

# poisson distribution data for y-axis
y1 = binom.pmf(x1,5,0.05)
y2 = poisson.pmf(x2,0.5)
y3 = poisson.pmf(x3,5)
y4 = binom.pmf(x4,500,0.01)

# plotting the graph
axs[0,0].vlines(x1,0,y1, colors=’k’, linestyles=’-’, lw=1)
axs[0,0].set\_title(’Bin(5,0.05)’)

axs[0,1].vlines(x2,0,y2, colors=’k’, linestyles=’-’, lw=1)
axs[0,1].set\_title(’Po(0.5)’)

axs[1,0].vlines(x3,0,y3, colors=’k’, linestyles=’-’, lw=1)
axs[1,0].set\_title(’Po(5)’)

axs[1,1].vlines(x4,0,y4, colors=’k’, linestyles=’-’, lw=1)
axs[1,1].set\_title(’Bin(500,0.01)’)

# showing the graph
plt.show()
```

Solution 5.17

Let X be the number of passengers who show up for the flight, $n = 75$ and $\pi = 0.96$.

a) If X is binomially distributed, the result is $\mathrm{P}(X = 73) \approx 0.226$

b) Since the probability that a passenger will show up is very high, for a reasonable Poisson approximation we must consider the counter-event Y, which describes the number of passengers who do not show up for the flight. Then Y is Poisson distributed with parameter $\lambda = 75 \cdot 0.04$ and it follows $\mathrm{P}(Y = 2) \approx 0.224$.

Solution 6.1

a) Obviously, $F_X(x) = \mathrm{P}(X \leq x)$ is equal to 0 for $x \leq 5$ and equal to 1 for $x \geq 7$. For $x \in [5, 7]$ we get

$$F_X(x) = \mathrm{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)\mathrm{d}t = \int_{5}^{x} \frac{1}{2}\mathrm{d}t = \frac{x-5}{2}.$$

So overall,

$$F_X(x) = \begin{cases} 0, & \text{for } x < 5 \\ \frac{x-5}{2}, & \text{for } x \in [5,7] \\ 1, & \text{for } x > 7 \end{cases}$$

b) $\mathrm{P}(X < 5.5) = F_X(5.5) = \frac{1}{4}$, $\mathrm{P}(5.8 < X < 6.8) = \int_{5.8}^{6.8} \frac{1}{2}\mathrm{d}t = \frac{1}{2}$. (In principle, you can always choose to calculate either via the density function f_X or via the distribution function F_X.

c) $\mathrm{E}(X) = \int_5^7 x \cdot \frac{1}{2}\mathrm{d}x = [\frac{x^2}{4}]_{x=5}^{7} = 6$

$\mathrm{Var}(X) = \int_{-\infty}^{\infty} \left(x - \mathrm{E}(X)\right)^2 \cdot f_X(x)\mathrm{d}x = \int_5^7 (x-6)^2 \cdot \frac{1}{2}\mathrm{d}x = \int_{-1}^{1} t^2 \cdot \frac{1}{2}\mathrm{d}t = \frac{1}{3}$

or: $\mathrm{E}(X^2) = \int_5^7 x^2 \cdot \frac{1}{2}\mathrm{d}x = [\frac{x^3}{6}]_{x=5}^{7} = 36\frac{1}{3} \quad \Rightarrow \quad \mathrm{Var}(X) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 = \frac{1}{3}$

d) $q_{0.5} = 6$, $q_{0.9} = 6.8$

Solution 6.2

a) It applies $\int_{-\infty}^{\infty} \frac{a}{1+x^2}\mathrm{d}x = 2 \cdot \int_0^{\infty} \frac{a}{1+x^2}\mathrm{d}x = 2a \cdot \lim_{T\to\infty} \arctan(T) = 2a \cdot \pi/2 = a\pi$. For this to be 1, $a = \frac{1}{\pi}$ must apply.

b) $F_X(x) = \frac{1}{2} + \frac{1}{\pi}\arctan(x)$ (the antiderivative of $f_X = \frac{1}{\pi}\frac{1}{1+x^2}$ with $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$)

c) $P\left(0 \leq X \leq \frac{1}{2}\right) = \frac{1}{\pi}\arctan(\frac{1}{2}) \approx 0.148$, $P\left(X \leq -\frac{1}{4}\right) = \frac{1}{2} - \frac{1}{\pi}\, arctan(\frac{1}{4}) \approx 0.422$, $P\left(X \geq \frac{3}{2}\right) = \frac{1}{2} - \frac{1}{\pi}\arctan(\frac{3}{2}) \approx 0.187$

d) $q_{0.5} = F_X^{-1}(0.5) = 0$ (obviously for reasons of symmetry), $q_{0.9} = F_X^{-1}(0.9) = \tan(0.4\pi) \approx 3.078$

Solution 6.3

$\mathrm{E}(Z) = \mathrm{E}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma}\left(\mathrm{E}(X) - \mu\right) = 0$
$\mathrm{Var}(Z) = \mathrm{Var}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2} \cdot \mathrm{Var}(X - \mu) = \frac{1}{\sigma^2} \cdot \mathrm{Var}(X) = 1$
$\sigma(Z) = \sqrt{\mathrm{Var}(Z)} = 1$ (or $\sigma(Z) = \sigma\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma} \cdot \sigma(X-\mu) = \frac{1}{\sigma} \cdot \sigma(X) = 1$)

For any random variable X with a finite expected value μ and finite standard deviation σ, the associated *standardized random variable* $Z = \frac{X-\mu}{\sigma}$ has an expected value of 0 and a variance of 1 (and thus also a standard deviation of 1). This will be particularly important for us in the case of normal distribution.

Solution 6.4

The distribution function of X is

$$F_X(x) = \begin{cases} 0, & \text{for } x < 0 \\ x, & \text{for } x \in [0,1] \\ 1, & \text{for } x > 1 \end{cases}$$

The relationship between the two distribution functions is

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(\mathrm{e}^X \leq y) = \mathrm{P}(X \leq \ln(y)) = F_X(\ln(y)).$$

Thus, the distribution function F_Y of Y is given by

$$F_Y(y) = \begin{cases} 0 & \text{for } \ln(y) < 0, \text{ thus } y < 1, \text{ or} \\ \ln(y), & \text{for } \ln(y) \in [0,1], \text{ thus } y \in [1, \mathrm{e}] \\ 1, & \text{for } \ln(y) > 1, \text{ thus } y > \mathrm{e} \end{cases}$$

The density function we are looking for is obtained by differentiation:

$$f_Y(y) = \begin{cases} 0, & \text{for } y < 1 \\ \frac{1}{y}, & \text{for } y \in [1, \mathrm{e}] \\ 0, & \text{for } y > \mathrm{e} \end{cases}$$

Solution 6.5

As already shown in the previous task, the relationship between the distribution functions is given by

$$F_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(\mathrm{e}^X \leq y) = \mathrm{P}(X \leq \ln(y)) = F_X(\ln(y)).$$

The relationship between the density functions is obtained by differentiation:

$$\begin{aligned} f_Y(y) = \tfrac{\mathrm{d}}{\mathrm{d}y} F_Y(y) &= \tfrac{\mathrm{d}}{\mathrm{d}y} F_X(\ln(y)) \\ &= F_X'(\ln(y)) \cdot \frac{1}{y} \\ &= f_X(\ln(y)) \cdot \frac{1}{y} \end{aligned}$$

Solution 6.6

Let T be the waiting time in minutes. Therefore, T is exponentially distributed with parameter $\bar{\lambda} = 4$.

a)

$$\mathrm{P}(T \leq 0.5) = \int_0^{0.5} 4\mathrm{e}^{-4t} \mathrm{d}t = 1 - \mathrm{e}^{-2} \approx 0.8647$$

b) We need to find k such that $\mathrm{P}(T \leq k) = 0.95$, i.e. we need the 95%-quantile of the random variable T. The probability that there will elapse less than 0.7489 minutes (that is about 45 seconds) between two emails is 0.95.

Solution 6.7

a) Let X be the number of accidents in one week. We assume that $X \sim \mathrm{Po}(4)$: $\mathrm{P}(X > 5) = 1 - \mathrm{P}(X \leq 5) \approx 0.2149$.

b) Let T be the waiting time between two accidents. We assume that $T \sim \text{Exp}(4)$

$$P(T > 2) = \int_2^\infty 4e^{-4t} dt = e^{-8} \approx 0.0003$$

Solution 6.8

a) $P(-1 \leq Z \leq 1) = P(Z \leq 1) - P(Z \leq -1) = \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 \approx 0.6827$
(**Note**: $\Phi(x) = F_Z(x)$ denotes the cdf of Z)

b) $P(-2 \leq Z \leq 2) = 2\Phi(2) - 1 \approx 0.9545$

c) $P(-3 \leq Z \leq 3) = 2\Phi(3) - 1 \approx 0.9973$

Compare your results to the well-known rule of thumb:

- For a normally distributed variable, approximately **68%** of the measured values lie within the interval mean ± one standard deviation.
- For a normally distributed variable, approximately **95%** of the measured values lie within the interval mean ± twice the standard deviation.
- For a normally distributed variable, approximately **99%** of the measured values lie within the interval mean ± three times the standard deviation.

d) $P(Z \leq 1) = \Phi(1) \approx 0.8413$

e) $P(|Z| \geq 0.5) = 1 - P(-0.5 \leq Z \leq 0.5) = 2 - 2\Phi(0.5) \approx 0.6171$

f) $P(-3 \leq Z \leq 1) = \Phi(1) - \Phi(-3) = \Phi(1) + \Phi(3) - 1 \approx 0.8400$

Solution 6.9

Since the height $X \sim \mathcal{N}(168, 6^2)$, we know that approximately 95% of women have a height in the range $[168 - 2 \cdot 6, 168 + 2 \cdot 6] = [156, 180]$. To be eligible to participate in the study, a woman must therefore be either smaller than 156 cm or taller than 180 cm.

Solution 6.10

a) We know that the IQ $X \sim \mathcal{N}(100, 15^2)$. Thus, approximately 95% of the population have an IQ in the range $100 \pm 2 \cdot 15$. The interval we are looking for is therefore $[70, 130]$. An IQ of 145 is 3 standard deviations away from the expected value. Due to the symmetry of the normal distribution, only about 0.15% of the population are more intelligent than their neighbor.

Solution 6.11

a) $1 - P(X \geq 4.98) = P(X \leq 4.98) = P\left(\frac{X-5}{0.02} \leq \frac{4.98-5}{0.02}\right) = P(Z \leq -1) \approx 0.159$

b) $1 - P(X \leq 5.05) \approx 0.006$

c) $P(|X - 5| \geq 0.03) = 1 - P(-0.03 \leq X - 5 \leq 0.03) \approx 0.134$

Solution 6.12

Let X be the deviation from the nominal value. Thus, $P(|X| \leq 3.45) = P(|Z| \leq 3.45/3) = P(Z \leq 1.15) - P(Z \leq -1.15) = 2 \cdot \Phi(1.15) - 1 \approx 0.7498$. The expected number of workpieces with the desired quality is therefore $0.7498 \cdot 24 \approx 17.9952$ (so about 18 – but remember that expected values are generally *not* rounded to whole numbers!).