Assessment 1
Statistics and probability theory

SS 2007

Prof. Dr. M.H. Faber

ETH Zurich

Thursday 3rd of May 2007
08:15 – 09:45

Surname: .....

Name: .....

Stud. Nr.: .....

Course of studies: .....

Date and duration:

Thursday, 3rd of. May 2007
Start: 8:15
End: 9.45
Duration: 90 minutes

Aids:

- No communication medium (e.g. cell phones, calculators with Bluetooth etc.)

Hints:

- Please control first, if you have received all the materials (listed under: Contents).

- Please place your Legi on your desk.
-
- Please write your **name on every sheet of paper,** at the bottom left side**.**
-
- Use only the provided sheets of paper.

- When you have finished, place **all materials** back in the envelope and raise your hand to call an assistant to collect it. You are allowed to leave till 9.15. If you finish later than this, wait quietly until the end time of the assessment (9.45).

- **Do not open** the paper fastener.

Contents

- General information and exercises (12 pages).
- 1 sheet of paper (checkered)

Part 1: „Multiple Choice“

In answering the following multiple choice questions it should be noted that for some of the questions several answers may be correct. Tick **ALL** correct alternatives in every question as: ☑

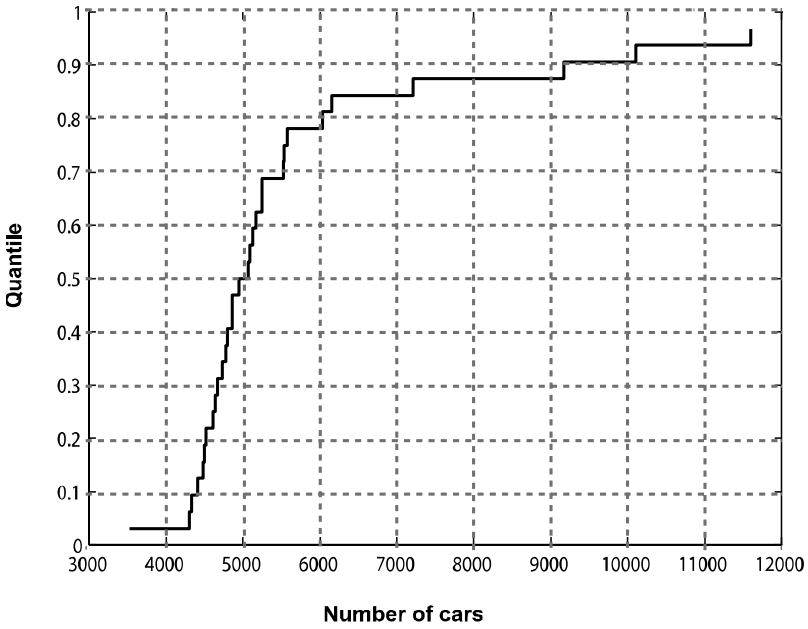
1.1 90 Persons are questioned about their income. 55 of them earn 10000 CHF a year, 34 of them earn 20000 CHF and 1 person, director of a bank, earns 2000000 CHF a year. Which one(s) of the following statement(s) is(are) correct?

The mode is 10000 CHF. ☑

The mean is 33300 CHF. ☐

None of the above. ☐

1.2 The following figure shows a cumulative distribution plot of the daily traffic flow through the Gotthard tunnel in January 1997.



The median is equal to 5000 cars. ☑

The mode is equal to 5000 cars. ☐

The mean is equal to 5000 cars. ☐

1.3 Which one(s) of the following statements is(are) correct?

The coefficient of variation is a measure of comparison of the dispersion of a data set. ☑

The sample covariance is a measure of the correlation of two data sets. ☑

None of the above. ☐

1.4 Two sets of observed values, A and B, are being compared. The observed values are shown in the following table. Which statement(s) is(are) correct?

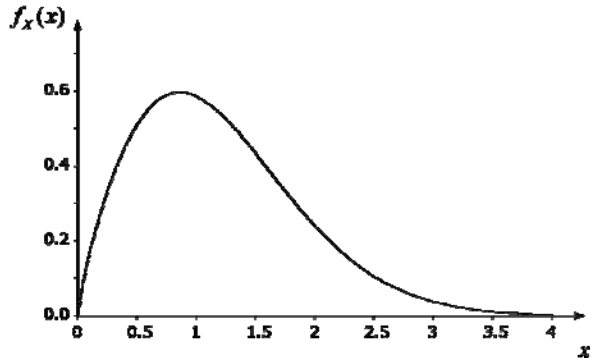
Set A	2	4	6	8
Set B	8	10	12	14

The sample variances of the two data sets are the same. ☑

The sample covariance of data set A is higher than the one of data set B. ☐

The coefficients of variation of the two data sets are the same. ☐

1.5 The probability density function of a continuous random variable X is shown in the following figure.



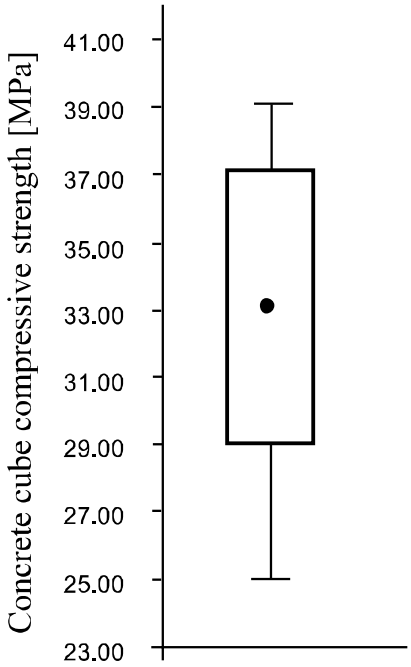
Which of the following statement(s) is(are) correct?

The distribution of the random variable is symmetric. ☐

The distribution of the random variable is skewed to the right. ☑

The mode of this data set is larger than the mean value. ☐

1.6 In the following figure a Tukey Box Plot is shown. Which statement(s) is(are) correct?



The outside values of this data set are lying within the interquartile range. ☐

The 0.25 quantile is 25 MPa. ☐

50% of the data lie between 25 MPa and 39 MPa. ☐

50% of the data lie between 29 MPa and 37 MPa. ☑

1.7 The 0.25 quantile of a data set corresponds to a value of the data set for which:

25% of the data values in the data set are larger. ☐

75% of the data values in the data set are larger. ☑

1.8 Which of the following statement(s) is(are) correct?

The sample correlation coefficient is defined in the interval $0 \le r_{xy} \le 1$. ☐

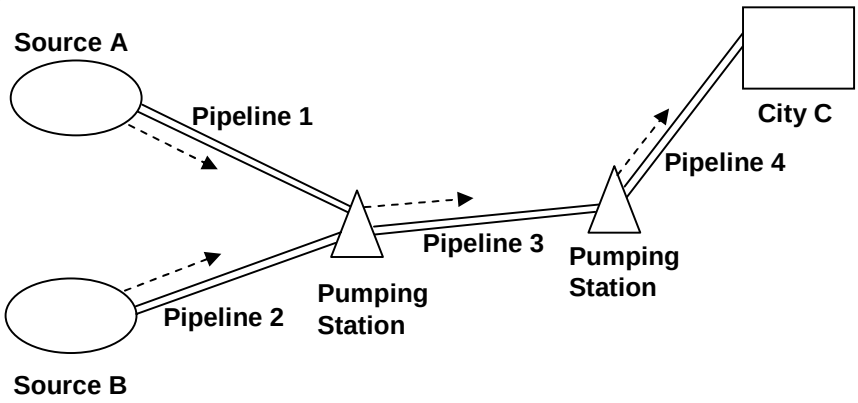
A perfect correlation between two data sets exists if r_{xy} is equal to 0. ☐

A perfect correlation between two data exists if r_{xy} is equal to 1. ☑

1.9 An easy and cheap on-site test method to determine whether the reinforcement in a concrete beam is corroded or not has been developed by an engineer. Based on experiments conducted in the laboratory, the engineer found that the test accurately predicts the occurrence of corrosion with a probability of 0.75. In order to account for differences in laboratory and actual on-site conditions, he however decides that the prediction probability of the test method should be set lower at 0.6. The assignment of this probability is based on which of the following definitions of probability:

Frequentistic	☐
Classical	☐
Bayesian	☑
None of the above	☐

1.10 The water supply for a city C comes from two sources A and B as shown in the figure below. Water is transported by pipelines 1, 2, 3 and 4. If required, either one of the two sources, by itself, is sufficient to supply the water for the city.



The failure of a pipeline will lead to a shortage of water in the city. If events E_1, E_2, E_3 and E_4 denote the failure of pipelines 1, 2, 3 and 4 respectively, the event that there is no shortage of water in city C can be written in the following way(s):

$\{\bar{E}_1 \cup \bar{E}_2\} \cap \{\bar{E}_3 \cap \bar{E}_4\}$	☑
$\overline{\{E_1 \cap E_2\} \cup \{E_3 \cup E_4\}}$	☑
$\overline{\{E_1 \cup E_2\} \cup \{E_3 \cup E_4\}}$	☐
$\overline{\{E_1 \cap E_2\}} \cap \overline{\{E_3 \cup E_4\}}$	☑

1.11 Two events A and B are mutually exclusive. Which of the following expressions is(are) correct?

$P(A|B) = P(A)$ ☐

$P(A|B) = 0$ ☑

$P(A \cup B) = P(A) + P(B)$ ☑

$P(A \cup B) = 1$ ☐

1.12 An experiment has four possible mutually exclusive outcomes – A, B, C and D. Which of the following assignments of probability values is(are) NOT correct?

$P(A) = 0.38, P(B) = 0.19, P(C) = 0.11, P(D) = 0.32$ ☐

$P(A) = 0.34, P(B) = 0.29, P(C) = 0.28, P(D) = 0.19$ ☑

$P(A) = 0.3, P(B) = 0.3, P(C) = 0.28, P(D) = 0.19$ ☑

1.13 In a class of 225 civil engineering graduate students, 134 are enrolled for a course in statistics (event A), 87 are enrolled for a course in mechanics (event B) and 28 are enrolled in both of these courses. The number of students who are not enrolled in either course is:

4 ☐

32 ☑

16 ☐

1.14 If A and B are two independent events, which of the following expressions is(are) correct?

$P(\bar{A} \cap \bar{B}) = P(\bar{A})P(\bar{B})$ ☑

$P(\bar{A} \cap \bar{B}) = 1 - P(A \cup B)$ ☑

$P(\bar{A} \cap \bar{B}) = 1 - P(A)P(B)$ ☐

1.15 From past experience, the probability that a painter A can do a painting work is 60%. The probability that a painter B can do a painting work is 80%.The probability that either one of the two painters can do a painting work is 90%. A and B are not independent. What is the probability that painter B can do a painting work given that painter A cannot do the same painting work?

0.45 ☐

0.30 ☐

0.90 ☐

0.75 ☑

1.16 It costs 60 CHF to test a certain component used in a machine and determine whether the component is defective before installing it. However if the component is installed without testing and later found to be defective, it costs 1200 CHF to repair the resulting damage to the machine. It is more profitable to install the component without testing if it is known that:

1% of all components produced are defective ☑

2% of all components produced are defective ☑

4% of all components produced are defective ☑

6% of all components produced are defective ☐

1.17 Probability distribution functions may be defined in terms of their moments. If X is a discrete random variable which one(s) of the following is(are) correct?

The second moment of X corresponds to its mean value, σ_X^2. ☐

The second central moment of X corresponds to its variance, σ_X^2. ☑

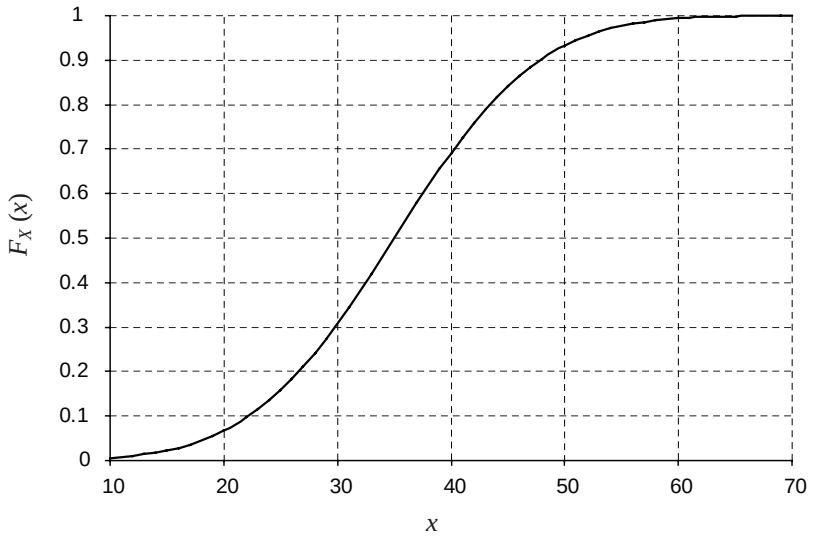
The first moment of X corresponds to its mean value, μ_X. ☑

1.18 The coefficient of variation $CoV[X]$ of a random variable X is a descriptor of:

the variability of the random variable around its expected value. ☑

the variability of the random variable around the median. ☐

1.19 The cumulative distribution function of a continuous random variable X is illustrated in the following diagram.



The probability of X exceeding the value of 30 is approximately equal to:

$P(X > 30) = 0.7$ ☑

$P(X > 30) = 0.3$ ☐

1.20 X is a random variable with mean value equal to $\mu_X = 18$ and standard deviation equal to $\sigma_X = 1$. The random variable X is now standardized into the random variable Z. Then Z:

has a mean value equal to $\mu_Z = 0$ and standard deviation equal to $\sigma_Z = 1$ ☑

has a mean value equal to $\mu_Z = 1$ and standard deviation equal to $\sigma_Y = 0$ ☐

1.21 A structure is designed to withstand wind speeds with a return period of 40 years. The probability that the design wind speed will be exceeded for the first time on the eleventh year after the structure is given in service is equal to:

0.02 ☑

$9.3 \cdot 10^{-17}$ ☐

1.22 The cost associated with the occurrence of an event A is a function of a random variable X with mean value $\mu_X = 100$ CHF and standard deviation $\sigma_X = 10$ CHF. Their relation can be written as: $C_A = a + bX + cX^2$, where a, b and c are constants and equal to 10, 0.5 and 1 respectively.

The expected cost of event A is equal to:

$E[C_A] = 10060$ CHF. ☐

$E[C_A] = 10160$ CHF. ☑

$E[C_A] = 160$ CHF. ☐

1.23 The random variables X and Y are correlated with a correlation coefficient equal to $\rho_{XY} = 0.25$. The mean values of the random variables are $\mu_X = 100$ and $\mu_Y = 80$, while their standard deviations are $\sigma_X = 10$ and $\sigma_Y = 8$. The variance of the function $D = 2X - Y$ is equal to:

$Var[D] = 424$. ☐

$Var[D] = 464$. ☐

$Var[D] = 374$. ☑

Part 1: Full Solution

1.1 Mode = most frequently occurring value, we have 55 times occurrence of the the value 10'000 CHF.

1.2 The median is equal to the 0.5 quantile; check the section "Quantile plots", page C-12.

1.3 The coefficient of variation is defined as $v = \frac{s}{\bar{x}}$, Equation (C.3) - With this ratio, the dispersion can be measured, so it is a measure of comparison of the dispersion of a data set.

The sample covariance is a measure of the correlation of two data sets - Check Equation (C.6).

1.4 The sample variances of the two given data sets are the same:

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

$$\overline{x_A} = (2+4+6+8)/4 = 5$$
$$\overline{x_B} = (8+10+12+14)/4 = 11$$

$$s_A^2 = \frac{1}{4}\left[(2-5)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2\right] = \frac{1}{4}(9+1+1+9) = 5$$
$$s_B^2 = \frac{1}{4}\left[(8-11)^2 + (10-11)^2 + (12-11)^2 + (14-11)^2\right] = \frac{1}{4}(9+1+1+9) = 5$$

As the sample covariance s_{XY} is a measure of correlation between two data sets –the sentence "The sample covariance of data set A is higher than the one of data set B" cannot be correct.

The coefficients of variation of the two data sets are not the same:

Coefficient of variation $v = \frac{s}{\bar{x}}$ *(C.3)*

$$s_A = \sqrt{s_A^2} = \sqrt{5};\ \ v = \frac{\sqrt{5}}{5}$$
$$s_B = \sqrt{s_B^2} = \sqrt{5};\ \ v = \frac{\sqrt{5}}{11}$$

1.5 The distribution is skewed to the right: can graphically be seen by the tail to the right. Since it is skewed, it cannot be symmetric.
The mode, the more frequently occurring value, is about $x = 0.9$. The mean would be found to its right, as the distribution is skewed to the right; so, it

would be larger and lying around 1-1.5. See also illustrations in the Exercise tutorials handouts.

1.6 First, obviously outside values cannot lie inside an interquartile range, as they are found outside the box plot. In this case however there are NO outside values at all.
The interquartile range goes from the 0.25 quantile to the 0.75 quantile, represented by the closed box.
In this Tukey Box Plot, the 0.25 quantile lies at 29 MPa, and the lower adjacent value lies at 25 MPa.
50% of the data lie between 29 MPa and 37 MPa. The limits at 25 and 39 MPa represent the lower and upper adjacent value.

1.7 Check definition in the chapter “Quantile plots” on page C-12.

1.8 The sample correlation coefficient r_{xy} is defined in the interval $-1 \le r_{xy} \le 1$. If r_{xy} is equal to 0, it means that there is absolutely no correlation between the observed data pairs. A perfect correlation between two data exists if r_{xy} is equal to 1 or equal to -1.

1.9 In the Bayesian interpretation of probability, the probability of occurrence of an event is formulated as a degree of belief that the event would occur (see Script section B.2). In this case, the decision of the engineer to set the prediction probability at a value of 0.6 which is lower than the value of 0.75 obtained from the laboratory experiments is based on his belief that this reduction would account for differences between laboratory and actual on-site conditions.

1.10 (This exercise has not been counted in the marking process)
Consider the event that there is a shortage of water in the city C. This can occur in **one** of the following ways:

i) failure of both pipeline **1** and pipeline **2** - $E_1 \cap E_2$ (Since it is given “If required, either one of the two sources, by itself, is sufficient to supply the water for the city”, hence both pipelines 1 and 2 need to fail for water shortage)
ii) failure of pipeline **3** - E_3
iii) failure of pipeline **4** - E_4

Combining the above listed 3 ways, the event that there is a shortage of water in the city C can be written as $E_1 \cap E_2 \cup E_3 \cup E_4$.

The event that there is no shortage of water in the city C is the complement of the event $E_1 \cap E_2 \cup E_3 \cup E_4$ and is hence obtained as $\overline{E_1 \cap E_2 \cup E_3 \cup E_4}$. By successive application of the associative law and De Morgan’s laws for set operations (Script equations B.4 and B.5),

$$\begin{aligned}\overline{E_1 \cap E_2 \cup E_3 \cup E_4} &= \overline{\{E_1 \cap E_2\} \cup \{E_3 \cup E_4\}} \\ &= \overline{\{E_1 \cap E_2\}} \cap \overline{\{E_3 \cup E_4\}} \\ &= \{\bar{E}_1 \cup \bar{E}_2\} \cap \{\bar{E}_3 \cap \bar{E}_4\}\end{aligned}$$

1.11 Two events A and B are said to be mutually exclusive if they are disjoint and the occurrence of one event precludes the occurrence of the other event. In other words, the occurrence of both events is impossible and hence

$$P(A \cap B) = 0 \Rightarrow$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = 0$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B)$$

1.12 It is given that the experiment has four possible (meaning any one of the outcomes should occur) and mutually exclusive (meaning the occurrence of more than one outcome is not possible) outcomes.

Hence the probabilities corresponding to the 4 outcomes must add up to 1. The sum of probabilities for the 3 given choices is:

$$P(A) = 0.38, P(B) = 0.19, P(C) = 0.11, P(D) = 0.32 \Rightarrow P(A) + P(B) + P(C) + P(D) = 1$$

$$P(A) = 0.34, P(B) = 0.29, P(C) = 0.28, P(D) = 0.19 \Rightarrow P(A) + P(B) + P(C) + P(D) = 1.1 (> 1)$$

$$P(A) = 0.3, P(B) = 0.3, P(C) = 0.28, P(D) = 0.19 \Rightarrow P(A) + P(B) + P(C) + P(D) = 1.07 (> 1)$$

Hence the second and the third choices are NOT correct assignments of probability values.

1.13 Let A be the set representing the number of students enrolled in the statistics course.

Let B be the set representing the number of students enrolled in the mechanics course.

Then the set $A \cap B$ represents the students enrolled in both the statistics and mechanics courses and the set $A \cup B$ represents the students enrolled in either one of both the courses.

From the given information,
Number of elements in set A = 134
Number of elements in set B = 87
Number of elements in the set $A \cap B$ = 28

The number of elements in the set $A \cup B$ is given by

$$A \cup B = A + B - A \cap B$$
$$= 134 + 87 - 28$$
$$= 193$$

Hence 193 students are enrolled in either one of both the courses.

The total size of the class is 225.

The number of students not enrolled in either of the 2 courses is hence 225 - 193 = 32.

1.14 Two events are said to be independent if the occurrence of one event does not influence or affect the probability of occurrence of the other event. If A and B are 2 independent events, then

$$P(A \cap B) = P(A)P(B) \qquad (Script\ Equation\ B.12)$$

Now

$$P(\bar{A} \cap \bar{B}) = P(\overline{A \cup B})$$
$$= 1 - P(A \cup B) \qquad (second\ choice)$$
$$= 1 - (P(A) + P(B) - P(A \cap B)$$
$$= 1 - (P(A) + P(B) - P(A)P(B))$$
$$= (1 - P(A))(1 - P(B))$$
$$= P(\bar{A})P(\bar{B}) \qquad (first\ choice)$$

1.15 The probability that painter A can do a painting work = $P(A)$ = 0.6

The probability that painter B can do a painting work = $P(B)$ = 0.8

The probability that either one of the 2 painters can do a painting work = $P(A \cup B)$ = 0.9

Then the probability that both the painters can do the painting work or $P(A \cap B)$ can be calculated from:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$
$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$
$$= 0.6 + 0.8 - 0.9$$
$$= 0.5$$

The probability that painter B can do a painting work given that painter A cannot do the same painting work needs to be determined. This can be expressed as:

$$P(B \mid \overline{A}) = \frac{P(B \cap \overline{A})}{P(\overline{A})} = \frac{P(B \cap (\Omega - A))}{1 - P(A)} \qquad \textit{,where } \Omega \textit{ is the sample space}$$

$$= \frac{P(B \cap \Omega - B \cap A)}{1 - P(A)} = \frac{P(B - B \cap A)}{1 - P(A)}$$

$$= \frac{P(B) - P(B \cap A)}{1 - P(A)} = \frac{0.8 - 0.5}{1 - 0.6} = \frac{0.3}{0.4} = 0.75$$

1.16 Let x % of the components produced be defective.

Then if the component is installed without testing, the probability that the component would be defective is $x/100$. The cost incurred to repair the machine based on this probability is hence $1200x/100$ CHF.

Now if the component was tested before installing, the cost incurred would be 60 CHF.

It is more profitable to install the component without testing if

$$1200x/100 \le 60$$

$$or\ x \le 5$$

Hence if 5% or less of components are known to be defective, it is profitable to install the component without testing. The first three choices (1%, 2%, 4%) are therefore the correct choices.

1.17 Check section D.3 in the script under the headline “Moments of random Variables and the Expectation Operator.

1.18 Check section D.3 in the script under the headline “Moments of random Variables and the Expectation Operator. (Equation D.11 and following text).

1.19 From a cumulative distribution plot one can directly read the probability of the random variable being less or equal to a value, i.e. :

$F_X(x) = P(X \le x)$

Hence from the provided plot it is: $P(X > 30) = 1 - P(X \le 30) = 1 - 0.3 = 0.7$

1.20 Any standard Normal distributed random variable has a mean equal to 0 and standard deviation equal to 1 (check Script Figure D.7).

1.21 The probability can be estimated as (script Equation D.57):

$$p_N(n) = p(1-p)^{n-1} = \frac{1}{40}\left(1 - \frac{1}{40}\right)^{11-1} \approx 0.02$$

1.22 Using the properties of the expectation operator (script Equation D.16) it is:

$$E[C_A] = E[a + bX + cX^2] = E\left[a\right] + E\left[bX\right] + E\left[cX^2\right] = a + b\mu_X + cE[X^2]$$

From Equation (D.17) in the script it is:

$$E[X^2] = Var[X] + \mu_X^2 = \sigma_X^2 + \mu_X^2 = 100 + 10000 = 10100$$

And:

$$E[C_A] = E[a + bX + cX^2] = a + b\mu_X + cE[X^2] = 10 + 0.5 \cdot 100 + 1 \cdot 10100 = 10160 \text{ CHF}$$

1.23 (This exercise has not been counted in the marking process)
Using the properties of the variance operator, equation (D.18) it is:

$$Var[D] = Var[2X - Y] = Var[2X] + Var[Y] + 2 \cdot 2 \cdot (-1) \cdot \rho_{XY} \cdot \sigma_X \cdot \sigma_Y =$$

$$4\sigma_X^2 + \sigma_Y^2 - 4 \cdot \rho_{XY} \cdot \sigma_X \cdot \sigma_Y = 4 \cdot 10^2 + 8^2 - 4 \cdot 0.25 \cdot 10 \cdot 8 = 384$$

Part 2: „Exercise - Moments of a random variable “

1) $\mu_X = \sum x_i P_X(x_i) = 1\times 0.3 + 2\times 0.4 + 3\times 0.2 + 4\times 0.1 = 2.1$

$$\begin{aligned}\sigma_X{}^2 &= \sum (x_i - \mu_X)^2 P_X(x_i) \\ &= (1-2.1)^2\times 0.3 + (2-2.1)^2\times 0.4 + (3-2.1)^2\times 0.2 + (4-2.1)^2\times 0.1 \\ &= 0.891 \\ \sigma_X &= \sqrt{0.891} = 0.944\end{aligned}$$

2)

$$\begin{aligned}\mu_Y &= \int y f_Y(y)dy = \int_0^1 y f_Y(y)dy + \int_1^2 y f_Y(y)dy + \int_2^3 y f_Y(y)dy + \int_3^4 y f_Y(y)dy \\ &= \int_0^1 0.3y dy + \int_1^2 0.4y dy + \int_2^3 0.2y dy + \int_3^4 0.1y dy \\ &= 0.3\left[\frac{y^2}{2}\right]_0^1 + 0.4\left[\frac{y^2}{2}\right]_1^2 + 0.2\left[\frac{y^2}{2}\right]_2^3 + 0.1\left[\frac{y^2}{2}\right]_3^4 \\ &= 0.3\cdot\left(\frac{1}{2}-0\right) + 0.4\cdot\left(\frac{2^2}{2}-\frac{1^2}{2}\right) + 0.2\cdot\left(\frac{3^2}{2}-\frac{2^2}{2}\right) + 0.1\cdot\left(\frac{4^2}{2}-\frac{3^2}{2}\right) \\ &= 1.6\end{aligned}$$

$$\begin{aligned}\sigma_Y{}^2 &= \int (y-\mu_Y)^2 f_Y(y)dy \\ &= \int_0^1 (y-1.6)^2\cdot 0.3dy + \int_1^2 (y-1.6)^2\cdot 0.4dy + \int_2^3 (y-1.6)^2\cdot 0.2dy + \int_3^4 (y-1.6)^2\cdot 0.1dy \\ &= 0.973\end{aligned}$$

$$\sigma_Y = \sqrt{0.973} = 0.987$$