Exercises

1	2	3	4	5	6	7	8	9

Surname, First name

Modelling, Uncertainty and Data for Engineers EXAM (CEGM1000)
Resit Q2

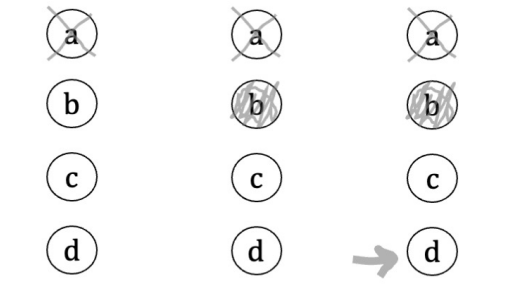
1	1	1	1	1	1	1
2	2	2	2	2	2	2
3	3	3	3	3	3	3
4	4	4	4	4	4	4
5	5	5	5	5	5	5
6	6	6	6	6	6	6
7	7	7	7	7	7	7
8	8	8	8	8	8	8
9	9	9	9	9	9	9
0	0	0	0	0	0	0

Do not open the exam or turn to the back page until given permission by the instructor!

You may write your name and student ID before the exam starts.

Before you start the exam, a few remarks:

- Write your first and last name in the field on the top left corner of this page.
- Write your student number **and** fill in the corresponding circle in the top right corner of this page.
- You may use pen or pencil and scientific (non-graphing) calculator. Any other tools and sources of information are not allowed.
- Exam should remain stapled. Scratch paper is available, but will not be collected or graded.
- On the following pages, some questions have a specific box for you to answer: anything written outside the boxes will not be graded.
- Answer space size does not indicate expected length of your answer! Shorter is generally better.
- A summary of points and questions is provided on the last page.
- Here is the order of priority in selecting multiple choice answers:



Answer: A Answer: B Answer: D

In case you need to correct your answers:

- Ask for a white sticker to cover your undesired written answers.
- For multiple choice questions, see guidance on the last page.

Good luck!

Exercise 1: Programming

2p **1a** Which of the following are correct statements about pandas (the Python library)?

Multiple answers can be selected. If you need to correct a mistake, indicate your final answer clearly (see last page of exam for guidance)

- ☐ It is useful for processing time series data
- ☐ It is useful for computing discrete fourier transforms (DFT)
- ☐ It uses a dictionary-like syntax
- ☐ It has marsupial properties
- ☐ It can integrate well with numpy

2p **1b** A colleague approaches you with a Python problem, telling you that they have two packages that are not compatible. Which of the following would be a reasonable approach to address this?

Select only one answer. If you need to correct a mistake, indicate your final answer clearly (see last page of exam for guidance).

- (a) Reinstall Anaconda
- (b) Restart your computer
- (c) Delete your changes and re-clone your Gitlab repository
- (d) Reactivate your mude environment
- (e) Create a new environment called mude2

3p **1c** A snippet of code:

```
rng = np.random.default_rng()
print(type(rng))
print(rng.random())
```

produces output:

```
<class 'numpy.random._generator.Generator'>
0.01628535642909701
```

Write the output of the following snippet of code, along with a brief explanation (the syntax and value do not need to be perfect):

```
rng.integers(5)
```

Exercise 2: Finite Volume Method

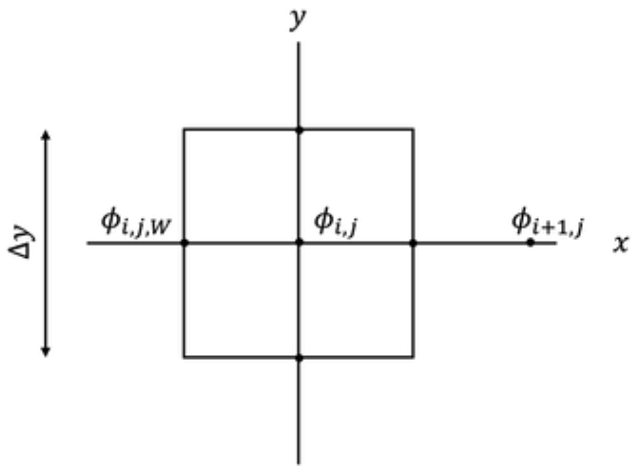
Consider the following discretized form of a partial differential equation (PDE), which includes only the component in the x-direction:

$$\phi_{i,j}^{n+1} = \phi_{i,j}^{n} + \frac{C\Delta t}{\Delta x}\left[\frac{\partial \phi_{i,j,E}^{n}}{\partial x} - \frac{\partial \phi_{i,j,W}^{n}}{\partial x}\right]$$

where:

- i, j represent the x- and y-coordinate, respectively
- $\phi_{i,j}^{n+1}$ is the value of the quantity of interest at the next time step
- $\phi_{i,j}^{n}$ is the value of the quantity of interest at the current time step
- $\phi_{i,j,W}^{n}$ and $\phi_{i,j,E}^{n}$ (*not indicated in the figure*) are the values at the current time step at the West and East face, respectively
- C is a known constant
- Δt is the time step size
- Δx is the space step size

The domain is divided into square ($\Delta x = \Delta y$) finite volumes, one of which, volume i, j, is represented below.



2p **2a** What partial differential equation has been discretized?

Select only one answer. If you need to correct a mistake, indicate your final answer clearly (see last page of exam for guidance).

(a) Advection
(b) Convection
(c) Diffusion
(d) Confusion

3p **2b** Write an expression to approximate the quantity at the East face of the volume $\frac{\partial \phi_{i,j,E}}{\partial x}$, which is needed in order to solve the discretized form of the PDE.

Exercise 3: Finite Element Method

Consider the following code for the computation of an element stiffness matrix with the finite element method in 1D.

```
def get_element_matrix(a, b, dx)

    locations = [(dx - dx/np.sqrt(3))/2, (dx + dx/np.sqrt(3))/2]
    weights = [dx/2, dx/2]
    Ke = np.zeros((3,3))

    for x, w in zip(locations, weights):
        N = [[ 2*x*x/dx/dx - 3*x/dx + 1, -4*x*x/dx/dx + 4*x/dx, 2*x*x/dx/dx - x/dx ]]
        B = [[ 4*x/dx/dx - 3/dx, -8*x/dx/dx + 4/dx, 4*x/dx/dx - 1/dx ]]
        Ke += w*a*np.matmul(np.transpose(B), B)
        Ke += w*b*np.matmul(np.transpose(N), N)

    return Ke
```

3p **3a** How many nodes does this element have?

3p **3b** How many integration points are used for computing the element matrix?

3p **3c** What is the corresponding PDE?

Select only one answer. If you need to correct a mistake, indicate your final answer clearly (see last page of exam for guidance).

(a) $-a\frac{\partial^2 u}{\partial x^2} + bu = 0$

(b) $-a\frac{\partial^2 u}{\partial x^2} = b$

(c) $-b\frac{\partial^2 u}{\partial x^2} = a$

(d) $au - b\frac{\partial^2 u}{\partial x^2} = 0$

Exercise 4: Signal Processing

10p **4** We start from a cosine signal with unit amplitude and a frequency of $f_c = 3$Hz, and sample it at $f_s = 5$Hz, for a duration of $T = 2$ seconds. The $N = 10$ discrete time samples are input to the Discrete Fourier Transform (DFT) and we directly plot the magnitude (modulus) of the output, hence $|X_k|$ with $k = 0, \ldots, N-1$. Create the resulting plot.

The corresponding Python code is:

```
import numpy as np
from matplotlib import pyplot as plt
fc=3
fs=5
dt=1/fs
T=2
N=T*fs
t=np.arange(0,T,dt)
xt=np.cos(2*np.pi*fc*t)
abs_fft = np.abs(np.fft.fft(xt))
plt.stem(abs_fft)
```

Exercise 5: Time Series Analysis

We intend to simulate a time series of 200 samples at 1-day intervals (so m=200 and the time unit is a day) using a first-order auto-regressive $AR(1)$ random process s_t as follows:

$s_t = \beta s_{t-1} + e_t$

where $t = 1, ..., m$ and a certain value for β are the given $AR(1)$ parameters. We further assume $E(s_t) = 0$ and $D(s_t) = \sigma^2 = 2$.

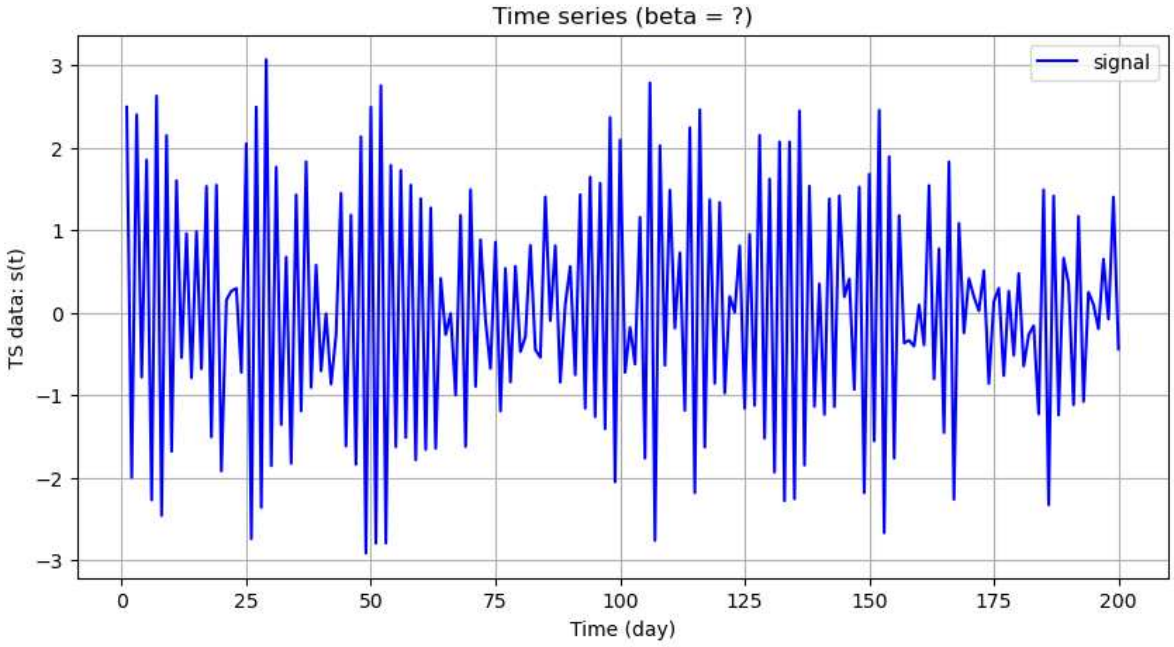
We simulate data using a normal distribution. To do so, we use the above recursive form, which needs initialization. To initialize the first data, we use:

$s_1 = s(t = 1) = randn(1)$

Further, the standard deviation of the random noise e_t can be obtained from:

$\sigma_e^2 = \sigma^2(1 - \beta^2)$.

A time series simulation versus time is shown in the plot below. The horizontal axis shows the time in units of day.



The following is a printed array of 10 consecutive samples from the time series:

```
10 consecutive sample values from time series:
~[ 2.49474675 −1.99859861 2.40207336 −0.78048668 1.85367808
−2.27074481 2.62934417 −2.4597126 2.15011297 −1.68199175]
```

Given the information and the plot above, answer the following questions.

2p **5a** What can you say about the type of noise?

Select only one answer. If you need to correct a mistake, indicate your final answer clearly (see last page of exam for guidance).

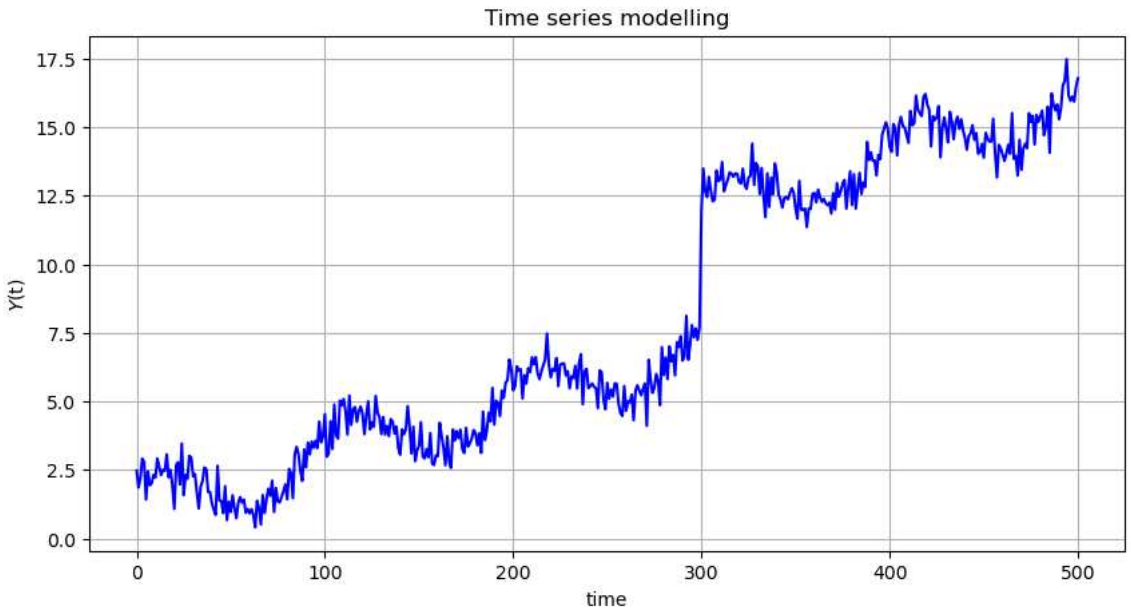
(a) White
(b) Colored
(c) Greyscale

3p **5b** Sketch the normalized auto-covariance function (ACF) of the generated time series above. Include the following in your plot:

- clearly labelled axes (ACF and time lag)
- a standard stem plot for each time lag (include at least 5 stems)
- besides being within appropriate numerical bounds, the absolute value of the ACF is not important; however, make sure it is possible to clearly distinguish whether the value is positive, negative or zero
- you do not need to include the confidence interval

Below we have a new time series of 500 observations versus time.

We want to model this time series using the Best linear Unbiased Estimation (BLUE).
$E(y) = Ax$



2p **5c** What would the parameters of x be? Stated in a different way: what components are clearly visible in the time series?

Multiple answers can be selected. If you need to correct a mistake, indicate your final answer clearly (see last page of exam for guidance)

- [ ] Noise
- [ ] Correlation
- [ ] Trend
- [ ] Drift
- [ ] Bias
- [ ] Seasonality
- [ ] Irregularities/outliers
- [ ] Offset
- [ ] Derivative

Exercise 6: Optimization

10p **6** Consider the following optimization problem:

Maximize $Z = 3x_1 + 2x_2$

Subject to:
$2x_1 + x_2 \leq 18$
$-4x_1 + 5x_2 \leq 10$
$x_1, x_2 \geq 0$

Convert the problem into its augmented form and then, if needed, find a next basic solution. Explain the process and the steps taken.
After finding a new basic solution (if applicable), determine the current values of the objective function and all decision variables.
State if this new solution is the optimal solution or not.

Exercise 7: Machine Learning

3p **7a** Consider the following kNN model trained on 100 noisy observations of a target function:

where thick/bold x's indicate the observations used (the neighborhood) to make a prediction at $x = 3$.
From classical decision theory, we are interested in minimizing the expected loss

$$\mathbb{E}[L] = \int \int (y(x) - t)^2 p(x, t) \, \mathrm{d}x\mathrm{d}t$$

which from variational calculus yields the result:

$$y(x) = \mathbb{E}_t[t|x]$$

Regarding the aforementioned decision theory concepts and the kNN model above, which **ONE** of the following statements is true?

Select only one answer. If you need to correct a mistake, indicate your final answer clearly (see last page of exam for guidance).

(a) Looking at how $\mathbb{E}[L]$ changes as a function of k we see that using higher values of k will in general lead to lower losses, as that corresponds to more flexible models

(b) When using kNN to predict at the edges of the dataset (around $x = 0$ or $x = 6.3$), we expect the function $\mathbb{E}_t[t|x]$ to become much more sensitive to small changes in x because there will be less samples in their neighborhood

(c) We can decompose $\mathbb{E}[L]$ into a bias term, a variance term and a noise term. For kNN, the value of $k = 1$ corresponds to the model with the highest possible bias, and therefore also the highest possible variance

(d) The function $y(x)$ obtained by kNN is an approximation of $y(x) = \mathbb{E}_t[t|x]$: the distribution $p(t|x = 3)$ is approximated by the thick/bold samples, while the expectation is approximated by averaging over these thick/bold samples

3p **7b** Regarding the use of feedforward neural networks for regression, which **ONE** of the following statements is true?

Select only one answer. If you need to correct a mistake, indicate your final answer clearly (see last page of exam for guidance).

(a) L_2 regularization can be used to alleviate overfitting, consisting in the inclusion of a $\frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$ term to the loss function. The lower λ is, the more rigid the model will become (higher bias)

(b) The choice of activation function can make a neural network model more or less flexible. At the high-bias limit, setting the activations of all layers to linear will make the network behave exactly like a simple linear regression model

(c) Neural networks can be seen as extensions of linear basis function models: in basis function models the mappings $\phi(\mathbf{x})$ are linear functions of $\mathbf{x}$ while in neural networks those can become nonlinear

(d) Neural networks are universal approximators and can therefore fit any continuum function to any arbitrary level of accuracy. This means that by introducing enough flexibility to the network we can always bring $\mathbb{E}[L]$ as close to zero as we want

3p **7c** Regarding the use of Stochastic Gradient Descent (SGD) for training regression models, which **ONE** of the following statements is true?

Select only one answer. If you need to correct a mistake, indicate your final answer clearly (see last page of exam for guidance).

(a) With minibatch SGD, each model update is performed with a number of samples larger than the original size of the dataset, which means some samples will appear more than once in the batch

(b) The stochastic nature of SGDs comes from the fact that updates are done with random subsets of the dataset. Nevertheless, the algorithm guarantees the full dataset will always be seen by the model, at which point we say that an epoch has passed

(c) An important parameter in SGD is the **learning rate** η. High values of η lead to more gradual (slower) changes in $\mathbf{w}$ after every update. Lower values are therefore preferred because that leads to faster convergence

(d) When using SGD, it is interesting to keep track of the training progress by computing the loss function on all training samples, even though model weights are updated exclusively on gradients computed using the validation dataset

Exercise 8: Extreme Value Analysis

General information about extreme value distributions is provided below. Think of it like a formula sheet. You may not need this information for all questions, and may not need all of it.

Recall that the Generalized Extreme Value distribution can be written as follows:

$$P[X < x] = \exp\left(-\left[1 + \xi\frac{x-\mu}{\sigma}\right]^{-1/\xi}\right) \quad \left(1 + \xi\frac{x-\mu}{\sigma}\right) > 0$$

whereby the random variable X is defined by the location, scale and shape parameters μ, σ and ξ, respectively. The design value x for annual (yearly) probability of non-exceedance p_y is:

$$x = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - (-\ln(1-p_y))^{-\xi}\right] & \xi \neq 0 \\ \mu - \sigma\ln(1-p_y) & \xi = 0 \end{cases}$$

with the design life probability over DL years given as:

$$p_{DL} = 1 - (1-p_y)^{DL}$$

Recall that the Generalized Pareto distribution can be written as follows for random variable X with parameters th (threshold), shape ξ and scale σ_{th}:

$$P[X < x|X > th] = \begin{cases} 1 - \left(1 + \xi\frac{x-th}{\sigma_{th}}\right)^{-1/\xi} & \xi \neq 0 \\ 1 - \exp\left(-\frac{x-th}{\sigma_{th}}\right) & \xi = 0 \end{cases}$$

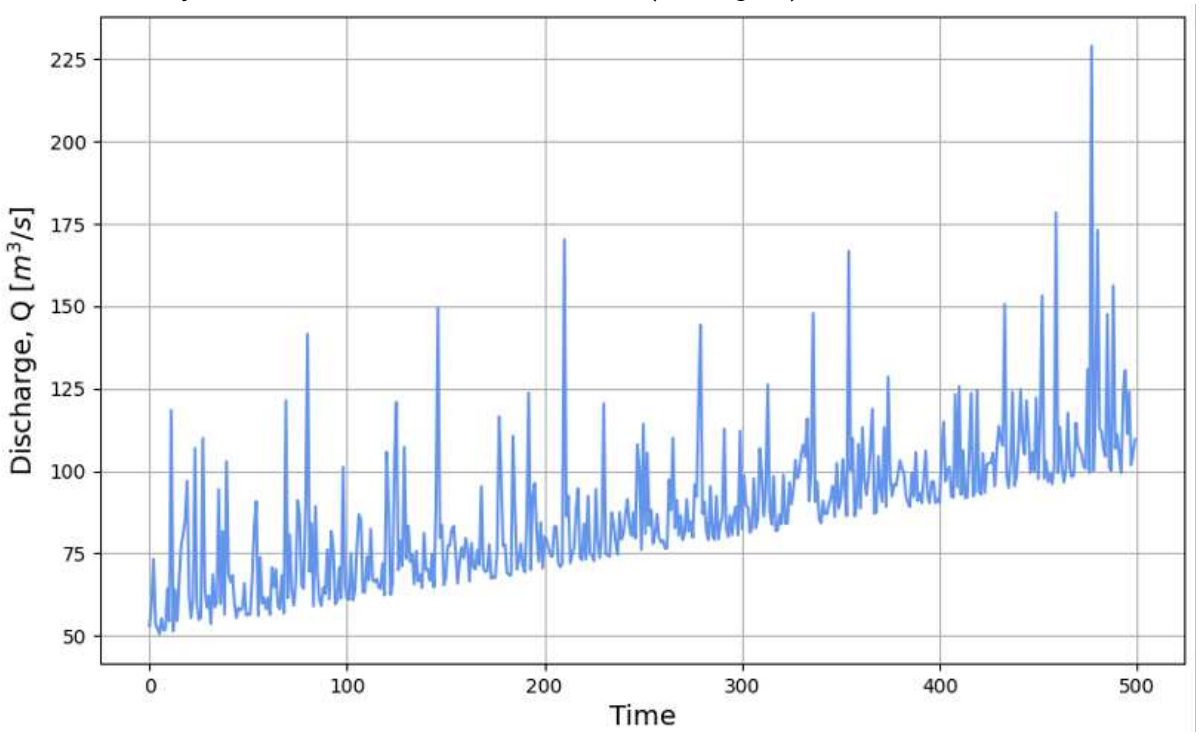
and the design value x_N for an N-years return level.

$$x_N = \begin{cases} th + \frac{\sigma_{th}}{\xi}\left[(\lambda_N)^{\xi} - 1\right] & \xi \neq 0 \\ th + \sigma_{th}\ln(\lambda_N) & \xi = 0 \end{cases}$$

where λ is the average number of excesses during the observation period, and N is the N-years return level.

As a consultant, you need to design a water treatment plant. There are different variables that need to be considered to this end: (1) the incoming water discharge to treat denoted as Q, (2) the concentration of organic matter denoted as C, and (3) the concentration of inorganic matter denoted as M.

You start studying the discharge Q, and you want to model the maximum yearly discharge (our random variable of interest). You managed to get a time series and when you start processing the sampled observations, you detect a trend around the mean (see Figure).



2p **8a** Which of the following properties is clearly not satisfied by the data:

Select only one answer. If you need to correct a mistake, indicate your final answer clearly (see last page of exam for guidance).

(a) Independent
(b) Identically distributed
(c) Asymmetry
(d) Mutually exclusive
(e) Collectively exhaustive

2p **8b** Can you apply Extreme Value Analysis to those observations of Q as they are?

Select only one answer. If you need to correct a mistake, indicate your final answer clearly (see last page of exam for guidance).

(a) Yes (b) No

Since you are not sure how to proceed with the discharge Q you move to the next variable, the concentration C.

2p **8c** You have been studying C for a long time (many years) and you know the distribution it follows, $C \sim F(c)$. What is the distribution of monthly maxima of C? Assume that C are daily observations and N is the number of observations per year.

Select only one answer. If you need to correct a mistake, indicate your final answer clearly (see last page of exam for guidance).

(a) $F(c)^N$
(b) $F(c)^{N/12}$
(c) $F(c)^{12N}$
(d) $F(c)^{N/30}$

Finally, you move to the last variable, the concentration M. A colleague has already done a study for you and characterized the monthly maxima of M using a Generalized Extreme Value distribution (GEV).

Your colleague determined the following values of the parameters: $\mu = 3$, $\sigma = 0.5$ and $\xi = 0.25$, in mg/L.

4p **8d** Compute the design value for a return period of 25 years.

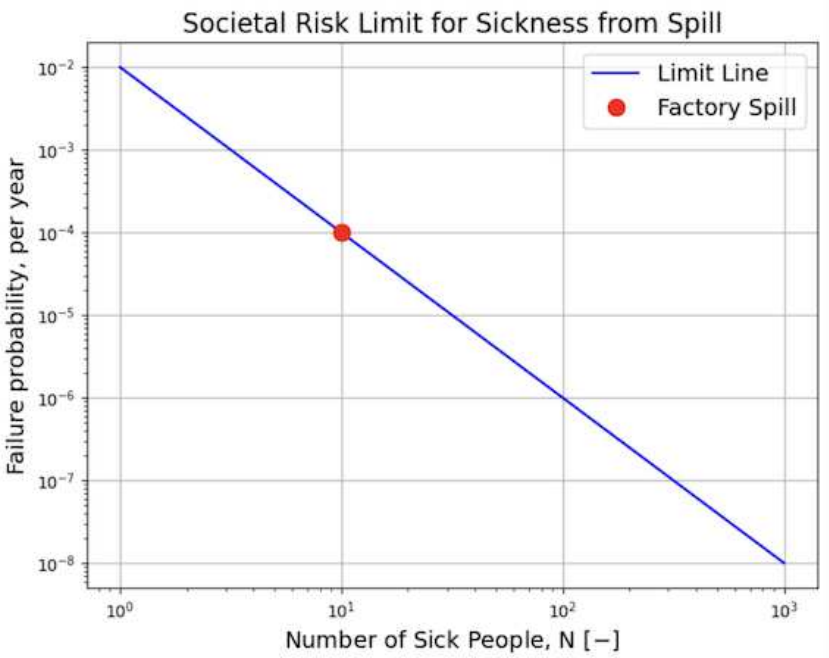
Exercise 9: Risk & Reliability

6p **9** The risk curve illustrated below shows the risk of people getting sick after drinking groundwater that has been contaminated after a "factory spill," a situation that is *barely tolerable*.

Provide an example of what can be done to improve the situation and be sure to include a statement of how one or more of the following should be *quantitatively* changed: consequence, probability, limit line.

Your answer should look like this (one sentence each, max):

- example of what to do
- quantitative effect on consequence, probability or limit line



This page is left blank intentionally

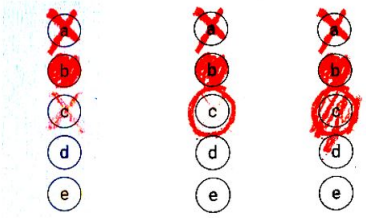
Exam overview

No.	Topic	Number of sub-parts	Points
1	Programming	3	7
2	Finite Volue Method	2	5
3	Finite Element Method	3	9
4	Signal Processing	1	10
5	Time Series Analysis	3	7
6	Optimization	1	10
7	Machine Learning	3	9
8	Extreme Value Analysis	4	10
9	Risk & Reliability	1	6
Total		21	73

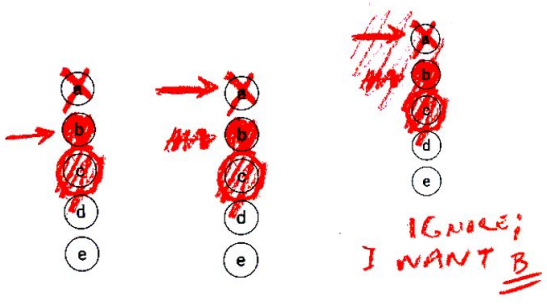
Use the table above to help plan your time during the exam.

In case you want to correct your answer for a multiple choice question put an ARROW in front of your final answer. If you also make a mistake with your arrow, write a clear message on the page. Here are a few examples:

Examples of UNCLEAR multiple choice response:



Examples of CLEAR multiple choice response:



Answer: B Answer: A Answer: B