Online self assessment for applicants of the Master programme Data Science

TU Dortmund University
Departments of Mathematics, Computer Science and Statistics

Dear interested Data Science students,

many thanks for your interest in our Master program.

- We would like to bring to your notice that our Master in Data Science is a consecutive Master program. This means that our department offers both a Bachelor and a Master in Data Science, where the Master builds upon all the knowledge of our Bachelor. Please note that the Bachelor is taught in German, while the Master is taught in English.
- Our Master program is open for career changers with different backgrounds than our own Bachelor. However, we expect from you that you are familiar with most of the topics that are taught in our Bachelor program. This test is meant give you an overview of the most important contents of our Bachelor: Hence, a student who successfully finished our own Bachelor in Data Science should be able to answer most of the question with ease.
- You are not expected to be able to answer all the questions. You should assess whether you have a sufficient mathematical background that you believe that with the corresponding effort you will be able to answer such questions, whether you are motivated to learn how to answer such questions and new concepts building upon the concepts addressed in these questions.
- While we require you to **do** the self-assessment, we do not consider the answers you actually gave when assessing your application. This is not a knowledge test. There will be no score, grade or points at the end. And your answers are **not** used to evaluate your application (you don't even have to send us your answers).
- This test is meant for you, for your own protection: If you join our Master courses, we expect you to know everything from our Bachelor. if you don't, you will not be able to follow our Master's courses appropriately, and it is rather likely the you will fail our program after investing a lot of time and money.
- Hence, we recommend that you to take this test seriously. It shows you what we expect from you. However, we are not interested in how good you perform in this test. There will be no score, and instead we'll tell you the correct solutions in the end.
- The test is designed in a multiple-choice way. For all 4 areas (Mathematics, Computer Science, Statistics and Data Science), there will be several questions, a total of 50. For each question, there are multiple answers (mostly, 4 different answers). You have to decide for each answer whether it is right or wrong.
- After you finished the test, please fill out our self-disclosure document. In this document you have to sign that you did the test. You have to upload this document on the *uni-assist* platform. Please double-check to sign it twice, once for the self-test and once for the report.

We wish you good luck!

1 Mathematics

- The symbol ln denotes the natural logarithm, that is, the logarithm with base e.
- The symbols $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ denote the sets of integers, rational numbers, real numbers, and complex numbers, respectively.

1.1 Calculus

Question 1. For $a \in \mathbb{R}$, which statements do hold for the one-sided limit $\lim\limits_{x \to a+} \dfrac{\ln(x-a)}{\ln(e^x - e^a)}$?

a) The limit exists.

b) Its value is 0.

c) Its value is 1.

d) Its value is e^a.

Question 2. Which statements do hold for the definite integral $\displaystyle\int_0^{\pi} e^{\cos x} \sin x \, dx$?

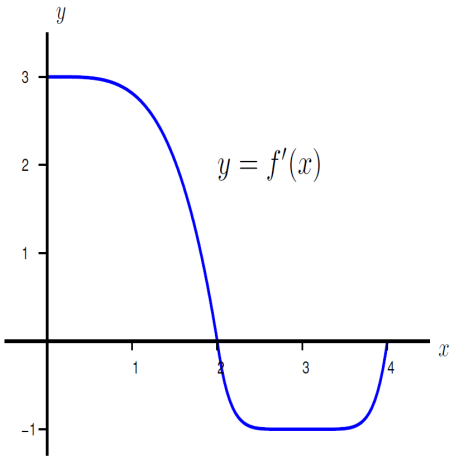
a) The antiderivative is not explicitly calculable.

b) The definite integral has a finite value.

c) Its value is $e - \dfrac{1}{e}$.

d) Its value is 0.

Question 3. The following figure shows the graph of the derivative f' of a function f, where f is continuous on the interval $[0, 4]$ and differentiable on the interval $(0, 4)$. Which of the following statements give the correct ordering of the function values $f(0)$, $f(2)$, and $f(4)$?



a) $f(0) < f(4)$

b) $f(4) < f(2)$

c) $f(2) \leq f(0)$

d) $f(4) = f(2)$

Question 4. Let f be the function defined by the series $f(x) = \displaystyle\sum_{n=1}^{\infty} \frac{x^n}{n}$ for all x such that $-1 < x < 1$. Which statements do hold?

a) The series converges absolutely.

b) The derivative is $f'(x) = \sum_{n=1}^{\infty} x^n$.

c) The derivative is $f'(x) = \sum_{n=0}^{\infty} x^n$.

d) The derivative equals $f'(x) = \dfrac{1}{1-x}$.

1.2 Linear Algebra

Question 5. Consider the following system of linear equations

$$\begin{array}{rcrcrcl} 3x & + & 2y & + & z & = & 0 \\ x & + & y & + & z & = & 0 \\ x & & & - & z & = & 0 \end{array}$$

with solutions of the form (x, y, z) where x, y, z are real numbers. Which of the following statements are correct?

a) The system is consistent.

b) The sum of any two solutions is a solution.

c) The system has a unique solution.

d) The system has infinitely many solutions.

Question 6. Which are eigenvalues of the matrix

$$\begin{pmatrix} 3 & 2 & 5 \\ 0 & 2 & 3 \\ 0 & 1 & 4 \end{pmatrix} \quad ?$$

a) 2

b) 3

c) 5

d) 0

1.3 Analytic Geometry

Question 7. Consider the solid in xyz-space, which contains all points (x, y, z) whose z-coordinate satisfies

$$0 \leq z \leq 4 - x^2 - y^2.$$

Which statements do hold?

a) The solid is a sphere.

b) The solid is a pyramid.

c) Its volume is 8π.

d) Its volume is $\dfrac{16\pi}{3}$.

Question 8. Consider the function g defined by $g(x, y) = e^y(y - x^2)$ for all real x, y. Which of the following terms are needed to represent the length of the gradient $\nabla g(1, -1)$?

a) $\sqrt{10}$

b) $\sqrt{5}$

c) e

d) π

Question 9. A circular helix in xyz-space has the following parametric equations, where $\theta \in \mathbb{R}$.

$$\begin{aligned} x(\theta) &= 4\cos\theta \\ y(\theta) &= 4\sin\theta \\ z(\theta) &= 3\theta \end{aligned}$$

Let $L(\theta)$ be the arclength of the helix from the point $P(\theta) = (x(\theta), y(\theta), z(\theta))$ to the point $P(0) = (4, 0, 0)$, and let $D(\theta)$ be the distance between $P(\theta)$ and the origin $(0, 0, 0)$. Let $L(\theta) = 10$. Which statements do hold?

a) $\theta = 4$

b) $\theta = 2$

c) To calculate the value of D for a given θ, $x(\theta)$ and $y(\theta)$ have to be evaluated explicitely.

d) $D(\theta) = \sqrt{52}$

1.4 Differential Equations

Question 10. Let $y : \mathbb{R} \to \mathbb{R}$ be the real-valued function defined on the real line, which is the solution of the initial value problem

$$y' = -xy + x, \qquad y(0) = 2.$$

Which statements are correct?

a) The problem is not uniquely solvable.

b) The solution $y(x)$ contains an exponential function.

c) $\lim\limits_{x\to\infty} y(x) = 1$

d) $\lim\limits_{x\to\infty} y(x) = 0$

2 Computer Science

2.1 Data Structures

Question 11. The number of steps taken for searching the value x in a binary tree with n nodes ...

a) depends not on x.

b) depends on n.

c) is $O(\log_2 n)$.

d) is $O(\log_x n)$.

Question 12. The average-case performance when looking up a single search key ...

a) is better with a Linked List than with a Hash Table.

b) is better with a Hash Table than with an Array.

c) is better with a Binary Search Tree than with a Hash Table.

d) is the same with a Linked List, an Array, and a Hash Table.

Question 13. Given 100 000 numbers, the minimum height of a binary search tree that can store all these numbers ...

a) depends on the numbers.

b) is larger than 20 levels.

c) is smaller than 19 levels.

d) can be calculated as $\log_{10}(100\,000)$.

Question 14. Which of the following statements are correct for a max-heap?

a) The root always contains the largest key.

b) All keys in the left subtree are always smaller than any key in the corresponding right subtree.

c) All leaves are located on the same level.

d) Each subtree is also a max-heap.

Question 15. Which of the following statements are correct for a binary search tree?

a) The root always contains the largest key.

b) All keys in the left subtree are always smaller than any key in the corresponding right subtree.

c) All leaves are located on the same level.

d) Each subtree is also a binary search tree.

Question 16. The following operations are applied to an empty stack `s`:

```
s.push(1)
s.push(2)
s.push(3)
s.pop()
s.push(4)
s.pop()
```

The result of a further `s.pop()` is ...

a) a number

b) undefined

c) 4

d) 2

2.2 Algorithms and Programming

Question 17. Sorting a data set is an important sub-problem in data science. Given the size n of a data set, which statements are correct?

a) Bubble Sort has worst-case run-time complexity $O(n)$.

b) Bubble Sort has worst-case run-time complexity $O(n \log(n))$.

c) Bubble Sort has worst-case run-time complexity $O(n^2)$.

d) Merge Sort has worst-case run-time complexity $O(n)$.

e) Merge Sort has worst-case run-time complexity $O(n \log(n))$.

f) Merge Sort has worst-case run-time complexity $O(n^2)$.

g) Quick Sort has worst-case run-time complexity $O(n)$.

h) Quick Sort has worst-case run-time complexity $O(n \log(n))$.

i) Quick Sort has worst-case run-time complexity $O(n^2)$.

Question 18. C1 and C2 are classes written in an object-oriented programming language (such as Java, C#, or C++). Which of the following statements are correct if C1 is a superclass of C2?

a) C1 is always an abstract class.

b) C2 contains all public features defined by C1.

c) Each C2 object may be replaced by a C1 object.

d) C2 is a subclass of C1.

Question 19. The following function f uses recursion:

```
def f(n):
    if n <= 1
        return n
    else
        return f(n-1) + f(n-2)
```

Let `n` be a valid input, i.e., a natural number. Which of the following functions returns the same result but without recursion?

a)
```
def f(n):
    a <- 0
    b <- 1
    if n = 0
        return a
    elsif n = 1
        return b
    else
        for i in 1..n
            c <- a + b
            a <- b
            b <- c
        return b
```

b)
```
def f(n):
    a <- 0
    i <- n
    while i > 0
        a <- a + i + (i-1)
    return a
```

c)
```
def f(n):
    arr[0] <- 0
    arr[1] <- 1
    if n <= 1
        return arr[n]
    else
        for i in 2..n
            arr[i] <- arr[i-1] + arr[i-2]
        return arr[n]
```

d)
```
def f(n):
    arr[0..n] <- [0, ..., n]
    if n <= 1
        return arr[n]
    else
        a <- 0
```

```
for i in 0..n
    a <- a + arr[i]
return a
```

2.3 Logic and Databases

Question 20. If A, B, and C are Boolean variables, which of the following statements are correct?

a) $A \wedge (B \vee C) = (A \wedge B) \vee (A \wedge C)$

b) $A \vee (B \wedge C) = (A \vee B) \wedge (A \vee C)$

c) $(A \wedge B) \vee C = C \vee (B \wedge A)$

Question 21. A large retail company keeps sales data local to the individual branches where sales transactions were performed. To compute overall sales statistics, the company wants to avoid sending the full sales data set to a central server. Instead, only aggregated sales information (sum, average, minimum, variance, median, maximum) is sent from each branch to the central site. Which of the following statements are correct?

a) The overall sum can be derived from the sums per branch.

b) The overall average can be derived from the averages per branch.

c) The overall minimum can be derived from the minimums per branch.

d) The overall variance can be derived from the variances per branch.

e) The overall median can be derived from the medians per branch.

f) The overall maximum can be derived from the maximums per branch.

Question 22. Consider the following table in a relational database.

Last Name	Rank	Room	Shift
Smith	Manager	234	Morning
Jones	Custodian	33	Afternoon
Smith	Custodian	33	Evening
Doe	Clerical	222	Morning

According to the data shown in the table, which of the following could be candidate keys of the table?

a) {Last Name}

b) {Room}

c) {Shift}

d) {Rank, Room}

e) {Room, Shift}

Question 23. The database interface of a library allows searching only for a single attribute (such as _Title_ or _Author_) in each query. Your friend decided to extend it's functionality and wrote an algorithm that allows searching for books that satisfy multiple predicates over single attributes in conjunction. He tells you the algorithm reuses the already implemented query functionality and works by intersecting the results (_book id's_) of queries over single attributes.

Which of the following assumptions on your friend's algorithm are plausible?

a) Its worst-case run-time necessarily increases exponentially with respect to the number of attributes in the query.

b) Its worst-case run-time depends on the length of the longest result of the single-attribute queries.

c) It might be implemented using an join.

d) It might be implemented using sorting.

2.4 Fundamentals of theoretical computer science

Question 24. Given an implementation of an algorithm, you want to check formally its run-time performance before you apply the algorithm to big data sets, in order to prevent endless runs of algorithms on your computer. The check if your algorithm runs endlessly on this data is depending on...

a) the length of the source code, it is a coding problem.

b) function calls in the algorithm, it is a call-graph problem.

c) recursion in the algorithm, it is a software design problem.

d) the size of your data, it is a big data problem.

Question 25. Which of the following languages are regular?

a) Words that consist of only vowels ('a', 'e', 'i', 'o', 'u').

b) Words where the 6th-last character is a vowel.

c) Words that contain as many vowels as consonants (non-vowels).

d) Palindroms (reading the word backwords yields the same word).

2.5 Computer Architecture

Question 26. In computer architecture, SIMD may refer to the situation where...

a) multiple CPU cores can access the same memory concurrently.

b) the same operation can be applied to multiple operands with only a single instruction.

c) multiple independent instructions can be executed at the same time in the same CPU core.

d) multiple independent memory banks show up as a single address space.

3 Statistics

3.1 Descriptive Statistics

Question 27. Which of the following sets have an arithmetic mean of 100, but a median smaller than 100?

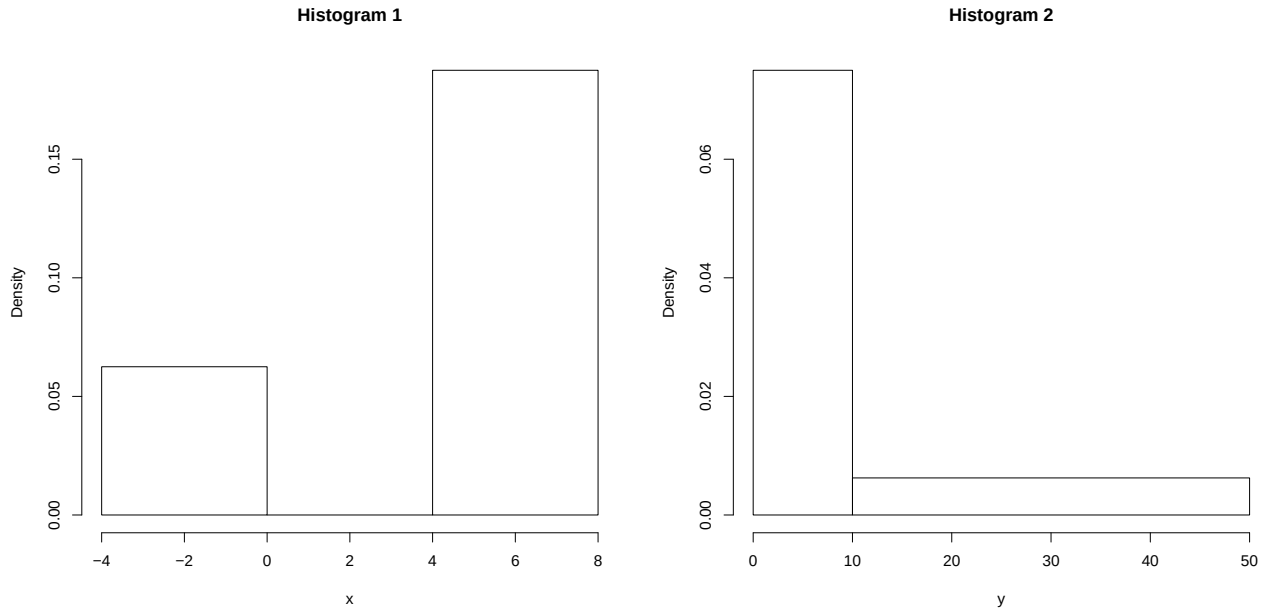
a) {80, 100, 120}

b) {80, 80, 140}

c) {0, 50, 150}

d) {60, 120, 120}

Question 28. Can there be a set of data fitting to both the following histograms? Which of these answers are correct?



a) No, because the right one is calculated from positive data only.

b) Yes, the right one includes all possible data from which the left one may be calculated.

c) No, the right one must be calculated with at least one value greater than 8.

d) No, the left one can not have been calculated with a value of 10 or more.

Question 29. Calculate estimates of the standard deviations s_x, s_y of the samples $x = (5, 9, 7)$ and $y = (-1, 2, 5)$ as well as the Pearson coefficient of correlation r_{xy} of x and y. Which of the following answers are correct?

a) $s_x = 4$, $s_y = 9$

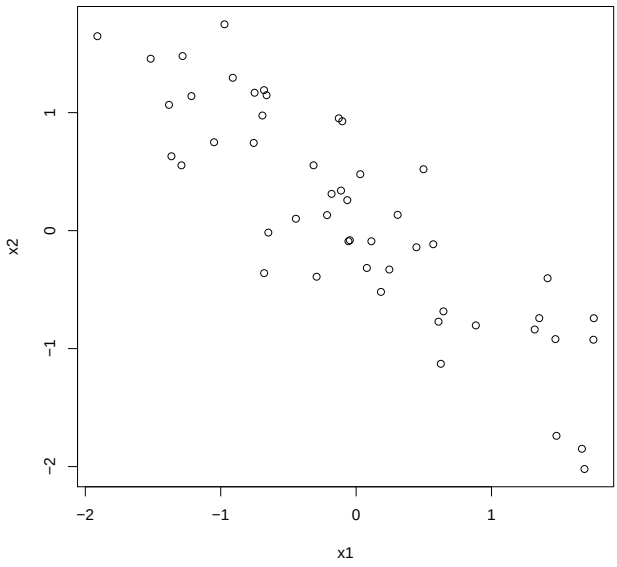
b) $r_{xy} = 0$

c) $s_x = 2$, $s_y = 3$

d) $r_{xy} = \frac{1}{2}$

e) $r_{xy} = \frac{1}{4}$

Question 30. Consider the following scatter plot.



The coefficient of correlation of the two variables ...

a) is negative.

b) is positive.

c) should have an absolute value greater than 0.4.

d) should be close to zero.

Question 31. Let the coefficient of correlation of two variables X and Y be larger than zero. What will be the effect on it, if the data of X are multiplied by the factor of 2?

a) The effect depends on the data of X.

b) It depends on Y.

c) The coefficient will be doubled.

d) It will be increased fourfold.

3.2 Probability

Question 32. There are 8 socks in your drawer: 4 black and 4 red. You take 3 of them with you in the dark. Which statements are correct?

a) It is sure that you get at least two socks (a pair) of the same colour.

b) It is sure that you get a pair of reds.

c) The probability to get 3 of the same colour is $\frac{1}{8}$.

d) The probability to get 3 of the same colour is $\frac{1}{7}$.

Question 33. In the sports injuries unit of a hospital, 40% of the patients are rugby players, 20% are swimmers and the remaining 40% play soccer. For a rugby player, the probability to be released on the first day is 10%; for a swimmer, it is 20%; for a soccer player, it is 80%. Which of the following statements are correct?

a) 40% of all patients are released on the first day.

b) Given a patient is released on the first day, the probability of her/him being a soccer player is 80%.

c) 80% of the non-swimmers have to stay for more than one day.

Question 34. Let X be a random variable with probability density function

$$f(x) = \begin{cases} \frac{1}{9}x^2, & x \in [0,3], \\ 0, & \text{else.} \end{cases}$$

Which of the following statements are correct?

a) The expected value of X is $\frac{9}{4}$.

b) The probability of $X < 1$ is $\frac{1}{27}$.

c) The probability of $X \in [0, 0.5]$ is $\frac{1}{54}$.

d) The probability of $X = 1$ is zero.

3.3 Inference and Linear Models

Question 35. Let X be a random variable defined by the density function

$$f(x) = \begin{cases} \dfrac{\alpha\beta^\alpha}{x^{\alpha+1}} & \text{, if } x \geq \beta \\ 0 & \text{, else} \end{cases}$$

with parameters $\alpha > 0$ and $\beta > 0$. We observe a sample $\{3, 4, 8\}$. Which of the following statements are correct?

a) The expected value of X exists for all combinations of α and β.

b) The expected value does only depend on α, but not on β.

c) Given the sample, β cannot be larger than 3.

d) If we assume $\beta = 2$, the estimate of α derived by the method of moments given the sample is $\dfrac{5}{3}$.

e) $\dfrac{5}{3}$ is also the maximum likelihood estimate of α in this case.

Question 36. We are interested in significant differences (level $\alpha = 0.05$) between the expected values μ_1 and μ_2 of two populations. Which of the following statements on statistical tests are correct?

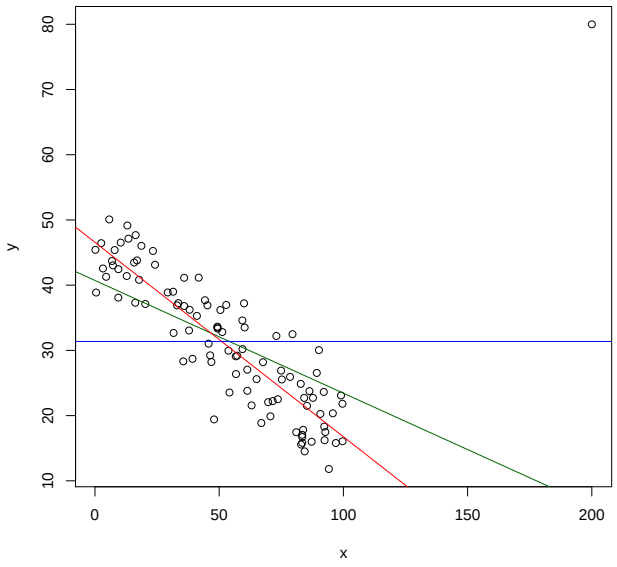
a) We will formulate the null hypothesis as $\mu_1 = \mu_2$.

b) A t-test can always be applied in this situation.

c) A p-value is the probability that the null hypothesis is correct, given the observed data.

d) If we obtain a p-value of 0.04, we will reject (level $\alpha = 0.05$) the null hypothesis.

Question 37. One of the lines in the following scatter plot is the regression line fitted to the data. Which of the statements are correct?



a) The red and green line have the right direction, and, hence, one of them could be the regression line.

b) The blue line seems to represent the mean value of the data with respect to y and thus could be the regression line.

c) The point in the top right corner has a strong influence on the regression line.

d) Leaving aside the point in the corner, the red line seems to fit better to the rest of the data.

Question 38. You have performed a linear regression analysis to explore sunflowers' growth (in meters per month) depending on the watering (in litres per day). You have estimated the regression coefficient to be $\hat{\beta} = 1.6$. What can you conclude?

a) There is a significant correlation between watering and growth.

b) An average sunflower growths 1.6 meters per month.

c) If you give it an additional litre of water per day, there will be an additional average growth of 1.6 meters per month.

d) According to the model assumptions, an additional litre of water per day will result in additional 19.2 meters of growth after one year.

e) You should consider further influencing quantities.

3.4 (Alternative 1) R Programming

In this section, you can opt between `R` (here) and `Python` (below) representations of the same questions.

Question 39. Which of the following `R` commands evaluates to `TRUE`?

a) `5 >= 5`

b) `TRUE & FALSE | FALSE & TRUE`

c) `FALSE & FALSE & FALSE | TRUE`

d) `!(((TRUE > FALSE) > TRUE) & !TRUE)`

Question 40. Consider the following code chunk:

```
x <- 0
while(x < 4) {
    x <- sample(1:3, 1)
    print(x)
}
```

It is not a good idea to run these lines because...

a) `x` is an invalid argument to `print()`.

b) the condition `x < 4` is never violated.

c) the function `sample()` does not exist.

d) `x` is initialised with the wrong type.

Question 41. Which of the following code lines return `TRUE`?

a) `max(c(2, 3, 4, NA, 1, 5)) == NA`

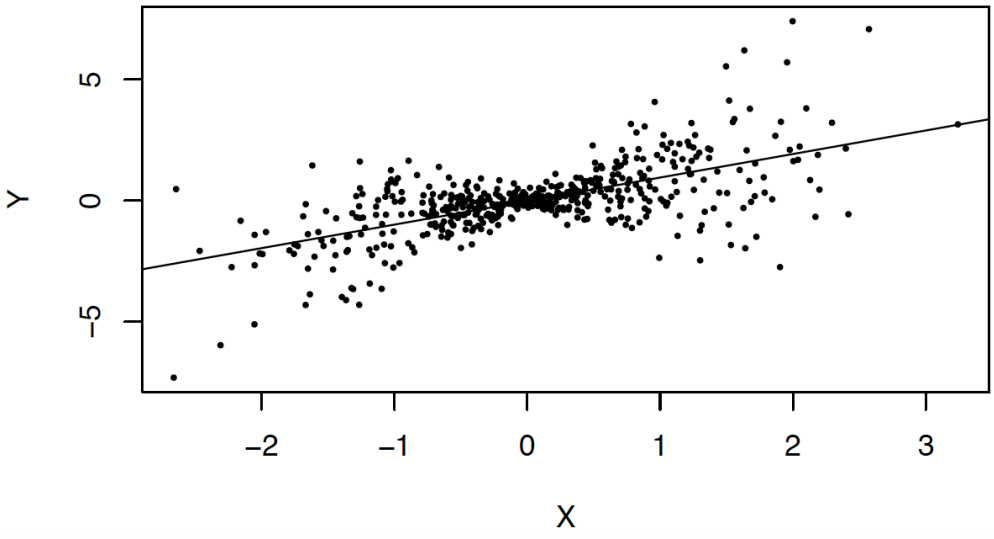
b) `max(c(2, 3, 4, NA, 1, 5), na.rm = TRUE) == 5`

c) `typeof(sum(c(1, 2, 3, 4, NA))) == "double"`

d) `typeof(sum(1:4)) == "integer"`

e) `typeof(sum(c(1L, 2L, 3L, 4L, NA_real_), na.rm = TRUE)) == "integer"`

Question 42. Which functions may have been used to generate the following plot and its underlying data?



a) `lm()`

b) `points()`

c) `abline()`

d) `integrate()`

Question 43. Consider the following code chunk and output and note that `NA` appears in the output of `lm()`.

```
X1 <- rnorm(1e2)
X2 <- X1 + 3
Y <- X1 + X2 + rnorm(1e2)
lm(Y ~ X1 + X2)
```

```
Call:
lm(formula = Y ~ X1 + X2)

Coefficients:
(Intercept)           X1           X2
      2.979        2.019           NA
```

Which of the following statements are correct?

a) Perfectly correlated regressors `X1` and `X2` are used.

b) `lm()` excludes `X2` from the regression so that there is a least squares solution.

c) `NA` indicates that the model fit to the data is perfect.

3.4 (Alternative 2) Python Programming

In this section, you can opt between `R` (above) and `Python` (here) representations of the same questions.

Question 39. Which of the following `Python` commands evaluates to `True`?

a) `5 >= 5`

b) `True & False | False & True`

c) `False & False & False | True`

d) `not(((True > False) > True) & (not(True)))`

Question 40. Consider the following code chunk:

```
import random
x = 0
while(x < 4):
    x = random.choice([1, 2, 3])
    print(x)
```

It is not a good idea to run these lines because...

a) `x` is an invalid argument to `print()`.

b) the condition `x < 4` is never violated.

c) the function `random.choice()` does not exist.

d) `x` is initialised with the wrong type.

Question 41. Which of the following codelines return `True`.

a) `numpy.argmax(numpy.array([2,3,4,numpy.NAN,1,5])) == 4`

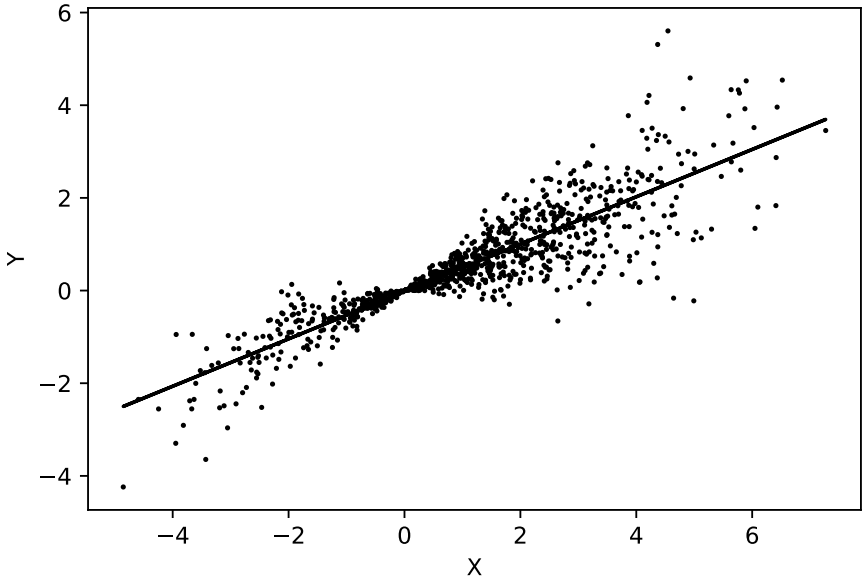
b) `numpy.nanargmax(numpy.array([2,3,4,numpy.NAN,1,5])) == 5`

c) `type(numpy.array([1,2,3,4,numpy.NAN]).sum()) is numpy.float64`

d) `type(numpy.array([1,2,3,4],dtype=object).sum()) is int`

e) `type(numpy.array([1,2,3,4]).sum()) is int`

Question 42. Which packages may have been used to generate the following plot and its underlying data?



a) `numpy`

b) `matplotlib`

c) `statsmodels`

d) `math`

Question 43. Consider the following code chunk and output and note that there are two warnings.

```
import numpy as np
import pandas as pd
import statsmodels.formula.api as sm

X1 = np.random.normal(0, 1, 100)
X2 = X1 + 3
Y = X1 + X2 + np.random.normal(0, 1, 100)

df = pd.DataFrame({"Y": Y, "X1": X1, "X2": X2})

linmodel = sm.ols(formula = "Y ~ X1 + X2", data = df).fit()
linmodel.summary()

Output:
Intercept   -0.0272
X1           1.1074
X2           1.0257

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The smallest eigenvalue is 3.27e-31. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Which of the following statements are correct?

a) Perfectly correlated regressors `X1` and `X2` are used.

b) Either `X1` or `X2` should be excluded, as the second regressor does not add any information to the model.

c) The second warning indicates that the model fit to the data is perfect.

4 Data Science

Question 44. Consider a data set `data` containing all German inhabitants, which is subsetted in the following process:

```
data = subset(data, Gender == "female")
data = subset(data, Status == "married")
data = subset(data, Haircolor == "brown")
```

Select which statements are true after all the subsets have applied.

a) The data set contains all brown haired and married females worldwide.

b) The data set contains all brown haired German inhabitants.

c) The data set contains all brown haired and married female German inhabitants.

d) The data set contains all brown haired, married, female German inhabitants with at least 2 children.

Question 45. For what ultimate purposes may algorithms like *Nelder-Mead*, *Newton-Raphson* or *gradient-descent* be used for?

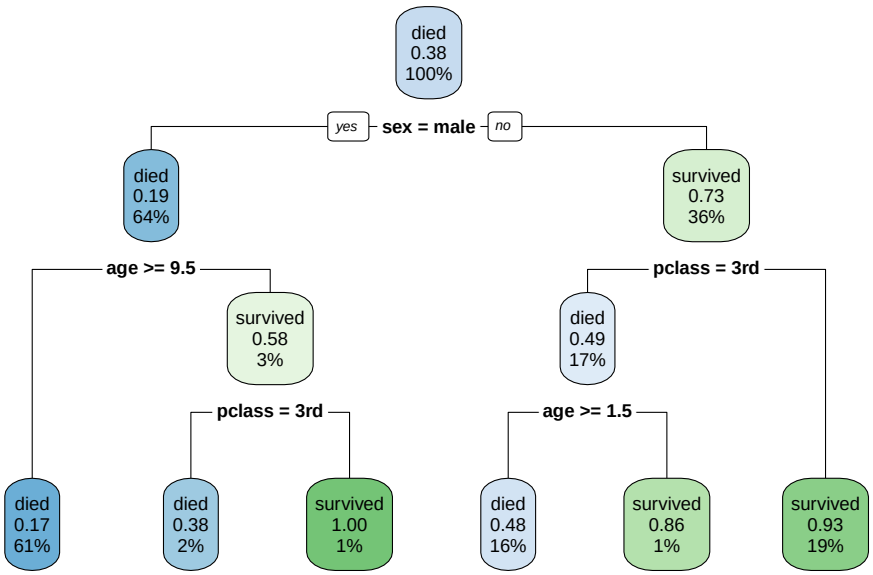
a) To find the minimum of a function.

b) To find all zeros of a function.

c) To evaluate the derivative of a function.

d) To solve a generalised regression problem.

Question 46. The Titanic data set contains information, whether passengers of the Titanic survived the shipwreck, based on their gender, age and passenger class. The following decision tree has been learned on this data. Which of the statements are true?



a) Overall, 62% of the passengers in the data set died.

b) A ‘new‘ passenger (female, 3rd class, 30 years old) is predicted to die in the shipwreck.

c) 62% of the passengers in the data set are female.

d) All male 3rd class passengers in the data set died.

Question 47. Random forests are one of the most famous machine learning methods. They are easy to understand, easy to implement and reach good prediction performances even without a hyper-parameter tuning. Which of the following statements on random forest are correct?

a) The prediction of a classification forest is made by a majority vote of the trees' predictions.

b) The prediction of a regression forest is the median of the tree predictions.

c) Each single tree in the forest uses only a part of the data available.

d) The training time of a random forest scales linear with the number of trees used.

Question 48. Let us return to the Titanic data set. We now have learned several models and want to choose the best one. We used three different methods to validate these models: The training error rate (apparent error rate), the error rate on an external test set and the error rate estimated by a 10-fold cross validation.

Learner	Training Error	Error on the test set	Cross Validation Error
Decision Tree	0.18	0.22	0.21
Random Forest	0.01	0.10	0.12
1-Nearest-Neighbour	0	0.18	0.19

Which of the following statements are correct?

a) 1-Nearest-Neighbour has a perfect training error and hence it should be used here.

b) Random Forests outperforms both 1-Nearest-Neighbour and the Decision Tree in terms of prediction error.

c) Not just in this case, but in general, Cross Validation is the better validation strategy and should always be preferred over the error on a single test set.

d) Not just in this case, but in general, Decision Trees always perform worse than Random Forests.

Question 49. We try a last model class to find the perfect model for the Titanic data-set: An SVM. The SVM is a model class that is very sensitive to hyper-parameter tuning. Especially, the cost parameter C and the bandwidth of the RBF kernel λ must be optimally adjusted in order to obtain a sensible model.

We use a nested resampling strategy to perform this hyper-parameter tuning: At first, 33% of the data are laid aside as an external test set, to validate the result of the hyper-parameter tuning itself (the outer resampling strategy). We use a random search as the tuning algorithm with a budget of 100 iterations. As parameter spaces, we use all positive real numbers for both C and λ. The performance of a single hyper-parameter setting is evaluated using a 10-fold cross validation (the inner resampling strategy). Moreover, in order to speed up the entire tuning process, we utilise parallel computing.

Which of the following statements are correct?

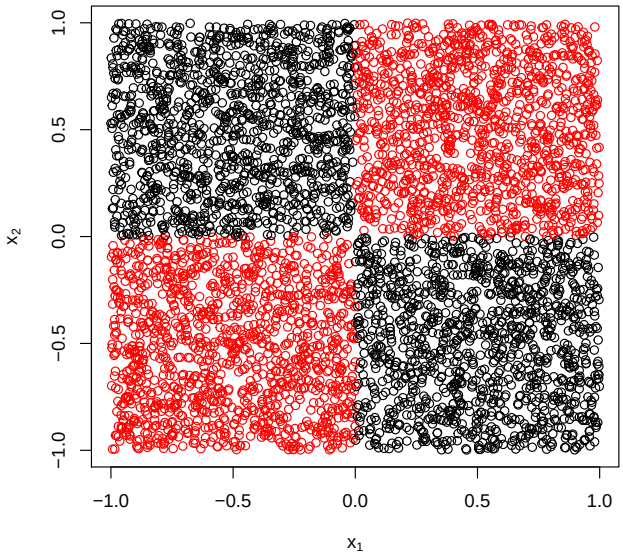
a) Using a nested resampling is necessary in order to detect underfitting.

b) As both C and λ are numeric parameters, any other optimization algorithm could be used instead of random search.

c) The choice of cross-validation as the inner resampling strategy is arbitrary, and a bootstrapping would lead to similar results.

d) The parallelization should take place at the innermost loop, hence, the execution of the inner cross-validation loop should be parallelized.

Question 50. Take a look at the following scatter plot of the so-called XOR data-set:



It is a classification data-set with the goal of separating the red and the black observations. Assume, that the number of red and black observations is approximately equal. Which of the following statements is correct?

a) A Decision Tree can reach a prediction error of (nearly) zero on this data-set.

b) When performing a variable selection using the step-wise forward selection algorithm, neither of the variables x_1, x_2 will be added to the model.

c) A Linear Discriminant Analysis (LDA) can reach a prediction error of (nearly) zero on this data-set.

d) Every model using only one of the two variables x_1, x_2 will have a missclassification error of approximately 50%.

We would like to thank you for taking your time and working through the test until the end. We hope that it helped you to get in insight into the topics of our Bachelor program, and to get an idea about the advanced methods taught in our Master program. If you want to check your answers, and to understand the solutions, please have a look into the solution-pdf under this link:
`https://statistik.tu-dortmund.de/storages/statistik/r/Downloads/Studium/Studiengaenge-Infos/Data_Science/Self_Test_Master_Data_Science_Solutions.pdf`