INTRODUCTION TO DATA SCIENCE

Example final exam questions 2025/26
(from lectures #3-#4, #6-#7)

Student's name: ....
Student's ID: ....

This exam consists of 25 multiple-choice questions with 3 possible answers. More than one answer can be correct! Also, it can happen that no answers are correct! In every question, for a correct assessment of answers from a) to c), you get 2 points. In case of correct assessment of 2 out of 3 possible answers, you get 0.4 points. For assessing only one out of three answers correctly, you get 0.2 points. A maximum number of points to be obtained is 50. To pass, you need to score at least 25 points.

You are not allowed to use the Internet or any notes or any course materials during the exam. Any attempt of cheating will immediately result in finishing the exam. Additionally, an appropriate information to the Dean and a formal request for disciplinary consequences to the University Disciplinary Commission will be sent. The above also applies in case of writing the exam after the time is over.

Good luck!

Q1. A researcher conducts an experiment to test whether a new dietary supplement affects drivers' reaction times. Thirty participants are measured under two conditions: before taking the supplement and after taking it. Reaction times are recorded in milliseconds.

The hypotheses are:

$$H_0 : \text{the supplement does not change the average reaction time}$$

$$H_1 : \text{the supplement changes the average reaction time}$$

The paired-sample t-test statistic is calculated as:

$$t = \frac{\bar{d}}{s_d/\sqrt{n}}$$

where: $\bar{d} = 12$ ms, $s_d = 30$ ms, $n = 30$.

Critical t-values for a two-tailed test are:

- $\alpha = 0.10 \Rightarrow t_{0.95} = 1.699$
- $\alpha = 0.05 \Rightarrow t_{0.975} = 2.045$
- $\alpha = 0.01 \Rightarrow t_{0.995} = 2.756$

Critical t-values for a right-tailed test are:

- $\alpha = 0.10 \Rightarrow t_{0.90} = 1.311$
- $\alpha = 0.05 \Rightarrow t_{0.95} = 1.699$
- $\alpha = 0.01 \Rightarrow t_{0.99} = 2.462$

Critical t-values for a left-tailed test are:

- $\alpha = 0.10 \Rightarrow t_{0.10} = -1.311$
- $\alpha = 0.05 \Rightarrow t_{0.05} = -1.699$
- $\alpha = 0.01 \Rightarrow t_{0.01} = -2.462$

Which of the following conclusions are correct?

A. At the 10% significance level, H_0 can be rejected.

B. At the 5% significance level, H_0 can be rejected.

C. At the 1% significance level, H_0 can be rejected.

Q2. A regional economic research institute is studying consumer behaviour in the market for a newly introduced household cleaning product. The product has been available for only two years, and the institute has collected a cross-sectional dataset of 150 households who purchased it during the past month.

Researchers believe that the quantity purchased by a household (Q) depends mainly on two factors:

- the price the household faced (P), which varies across stores and discount programs,
- the household's monthly income (I).

Because the true functional form of the demand relationship is unknown, the team estimates three competing OLS models: a log-log model, a semi-log model, and a linear model. They hope that analysing elasticities will help determine which specification best matches economic theory.

To simplify comparison, you are also given the sample means of the key variables:

$$\bar{P} = 20, \quad \bar{I} = 50, \quad \bar{Q} = 80.$$

The estimated models are as follows:

Model A (log-log):

$$\ln(Q_i) = \beta_0 + \beta_1 \ln(P_i) + \beta_2 \ln(I_i) + u_i,$$

with estimated coefficients:

$$\hat{\beta}_0 = 0.4, \qquad \hat{\beta}_1 = -0.6, \qquad \hat{\beta}_2 = 0.3.$$

Model B (semi-log, log-linear in Q):

$$\ln(Q_i) = \gamma_0 + \gamma_1 P_i + v_i,$$

with estimated coefficients:

$$\hat{\gamma}_0 = 0.3, \qquad \hat{\gamma}_1 = -0.1.$$

Model C (linear):

$$Q_i = \delta_0 + \delta_1 P_i + w_i,$$

with estimated coefficients:

$$\hat{\delta}_0 = 0.9, \qquad \hat{\delta}_1 = -1.5.$$

The research team wants to interpret price elasticities implied by the three models. Using the provided estimates and sample means, determine which of the following statements are correct.

A. In Model A, the price elasticity of demand for Q with respect to P equals -0.6.

B. In Model B, the point elasticity of demand with respect to price, evaluated at $\bar{P} = 20$, is approximately -0.30.

C. In Model C, the point elasticity of demand evaluated at $(\bar{P}, \bar{Q}) = (20, 80)$ equals -0.375.

Q3. A technology startup is developing a system that predicts the energy consumption of households (single value per observation, expressed in kWh) based on features such as house size, outdoor temperature, time of day, and number of occupants. As a baseline, the data science team decides to implement a simple feedforward neural network.

After preliminary experiments, they choose the following architecture:

- input layer: 4 features,
- hidden Layer 1: 8 neurons with ReLU activation,
- hidden Layer 2: 4 neurons with ReLU activation,
- output layer: 1 neuron with linear activation.

The team wants to ensure they properly understand the structure and implications of this architecture. Evaluate the correctness of the following statements.

A. The total number of trainable parameters between the input layer and hidden layer 1 is $4 \times 8 = 32$.

B. Using ReLU activation ensures that the outputs of both hidden layers are always strictly positive (greater than zero).

C. Increasing the number of neurons in output layer would allow the network to approximate more complex functions of the input features, and therefore predict the target more accurately.

Q4. A media analytics company wants to analyze customer reviews for a popular smartphone to identify the main topics discussed by users. The dataset contains 5,000 reviews collected from multiple online stores.

The data science team decides to apply standard text mining preprocessing steps before building a topic model. They perform the following:

- clean the texts by removing punctuation and numbers, and converting all words to lowercase,
- perform tokenization to split the text into individual words or terms,
- remove stop words (such as "the", "and", "of") from all reviews,
- apply stemming to reduce words to their root forms (e.g., "running" → "run"),
- create a document-term matrix including unigrams, bigrams, and trigrams to capture multi-word expressions,
- fit a Latent Dirichlet Allocation (LDA) model to discover k latent topics in the reviews, with k optimized based on topic coherence measures.

Based on these preprocessing steps, evaluate the correctness of the following statements.

A. Stemming is an approach that successfully enables more efficient text analysis and information retrieval across different languages, including English, German, Polish, and Mandarin.

B. Incorporating unigrams, bigrams, and trigrams into the document-term matrix means that, only tha sequences of three consecutive tokens are considered.

C. In the procedure described above, a mistake was made because no topic modeling algorithm actually takes a document-term matrix as input; all of them rely on word embeddings obtained from large language models.