DAT246-251027 Exam Information

DAT246/DIT246: Empirical Software Engineering

Date and time: 2025-10-27, 14:00 - 18:00
Place: Johanneberg
Examiner: Richard Torkar
Teaching Assistant: Julian Frattini
Questions will be answered by Julian Frattini who can be reached via telephone at +46703690228.

- The exam can give max. 70 points. Note that some questions give more points than others.
- For multiple choice (MC) questions, you receive the full marks when selecting all correct and none of the incorrect answers, and 0 points otherwise. There are no negative points that carry over.
- Provide brief and concise answers in English.
- All answers need to be provided by you alone and in your own words. Collaboration with other students or consulting additional material is not permitted.
- The answers must be consistent with the course content.

Points	Grade
[0, 35[	U
[35, 49[	3
[49, 65[	4
[65, 70]	5

1 ABC

What do A, B, and C mean in the “ABC of Software Engineering Research” by Klaas-Jan Stol and Brian Fitzgerald [1]?

Write your answer in the box below. Changes are saved automatically.

[1] Stol, K. J., & Fitzgerald, B. (2018). The ABC of software engineering research. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *27*(3), 1-51.

Maximum marks: 3

2 Obtrusiveness (Matching)

Klaas-Jan Stol and Brian Fitzgerald structure eight research strategies along two axes. One axis represents the “level of obtrusiveness”. Rank the following research strategies regarding their typical level of obtrusiveness as presented by Klaas-Jan Stol and Brian Fitzgerald [1].

Please match the values:

	Most obtrusive	Very obtrusive	Slightly obtrusive	Least obtrusive
Laboratory Experiment	○	○	○	○
Sample Study	○	○	○	○
Computer Simulation	○	○	○	○
Field Experiment	○	○	○	○

[1] Stol, K. J., & Fitzgerald, B. (2018). The ABC of software engineering research. *ACM Transactions on Software Engineering and Methodology (TOSEM)*, *27*(3), 1-51.

Maximum marks: 4

3 Case Study Time Frame

Explain the difference between a **retrospective** and **longitudinal** case study. Name one unique advantage and disadvantage for each of them.

Write your answer in the box below. Changes are saved automatically.

Maximum marks: 4

4 Triangulation

Explain what triangulation in case study research is and what it is used for. Describe one type of triangulation.

Write your answer in the box below. Changes are saved automatically.

Maximum marks: 3

5 Experimentation Terminology (Matching)

In a (fictional) study, researchers investigated whether the use of ChatGPT helps university students solve coding assignments faster. Match the following terms relevant to experimentation to their respective entity in the study: treatment, control, outcome, subject, object.

Please match the values:

	control	object	outcome	subject	treatment
duration of the assignment	○	○	○	○	○
not using ChatGPT	○	○	○	○	○
university student	○	○	○	○	○
using ChatGPT	○	○	○	○	○
coding assignment	○	○	○	○	○

Maximum marks: 5

6 Sampling

Explain the difference between probabilistic and non-probabilistic sampling. Give one example of a probabilistic and one of a non-probabilistic sampling method.

Write your answer in the box below. Changes are saved automatically.

Maximum marks: 4

7 Threats to Validity

Name the four types of threats to validity that are typically discussed in software engineering research. Explain, how each of these types threatens the validity of research results.

Write your answer in the box below. Changes are saved automatically.

Maximum marks: 6

8 Compulsory Steps of BDA

Name and describe compulsory steps in a full Bayesian data analysis in the right order.

Write your answer in the box below. Changes are saved automatically.

Maximum marks: 8

9 Types of Relationships

In a directed acyclic graph (DAG), there are three major ways how a variable z could be related to two variables $x \rightarrow y$. For each of the three ways,

- specify how the variable would be related to x and y (using -> as an arrow symbol),
- state how this relationship is commonly called, and
- explain how conditioning on z would affect the estimation of the effect of x on y.

Write your answer in the box below. Changes are saved automatically.

Maximum marks: 9

10 Transformations in Regressions

Some regression models have the following form.

$$y \sim Binomial(n, p)$$
$$f(p) \leftarrow \alpha + \beta_x x + \beta_z z$$

Explain, what $f()$ generally does, and how it is usually called. Then, name the respective function that would be appropriate in the given example above and explain, how this particular function works.

Write your answer in the box below. Changes are saved automatically.

Maximum marks: 5

11 Outcome Distribution: Count (MC)

Which of the following distribution families have maximum entropy for count variables?

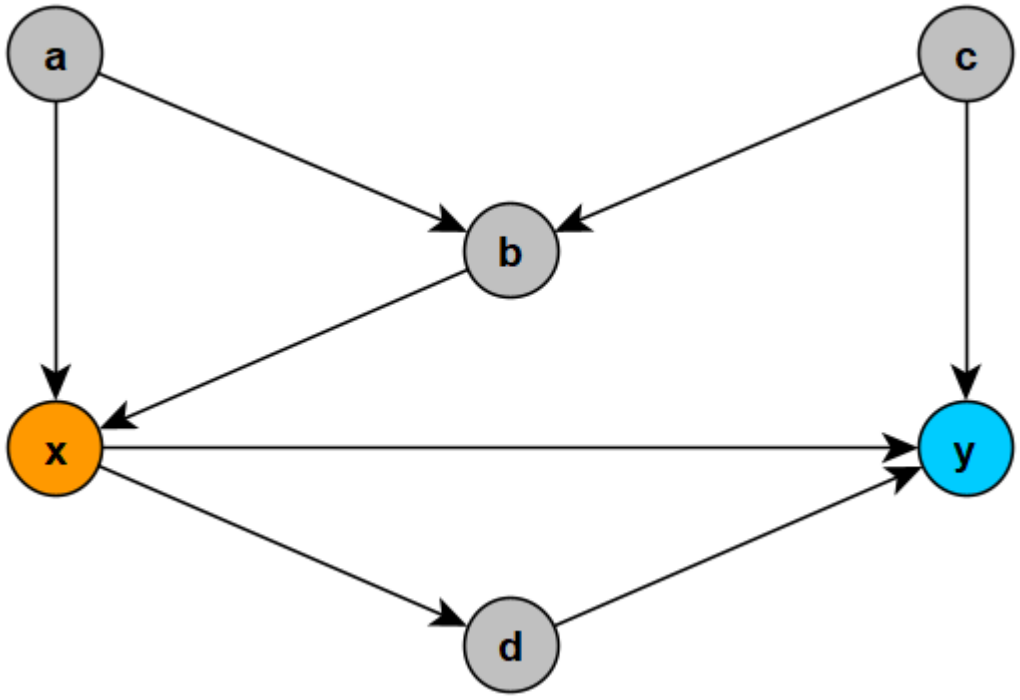
Select one or more alternatives:

- [ ] Negative-Binomial Distribution
- [ ] Poisson Distribution
- [ ] Binomial Distribution
- [ ] Log-Normal Distribution

Maximum marks: 4

12 Adjustment Set Identification

Given the following DAG, select all valid adjustment sets to estimate the unbiased (total or direct) effect of x on y.



Select one or more alternatives:

- [ ] { c }
- [ ] { b ; d }
- [ ] { a; c }
- [ ] { a; b; d }

Maximum marks: 3

13 Model Derivation

Consider a university course with multiple assignments. You want to investigate if **passing an assignment** is affected by **the time spent on it**. You have a data set from with one row per combination of student and assignment containing the following variables:

Variable	Description	Type
$\mathtt{student}$	Unique identifier of the student	ID
$\mathtt{assignment}$	Unique identifier of the assignment	ID
$\mathtt{time.spent}$	Time spent on the respective assignment	Standardized N(0,1)
$\mathtt{assignment.passed}$	Whether or not the student passed the assignment	Boolean
$\mathtt{interest}$	Level of interest of the student in the topic of the respective assignment	Standardized N(0,1)

Through identifying the adjustment set from a DAG, you have already determined that the variable $\mathtt{interest}$ confounds the effect of interest.

Specify two statistical models that estimates the unbiased effect $\mathtt{time.spent} \rightarrow \mathtt{assignment.passed}$ where the outcome distribution has maximum entropy. The models must consist of

1. the outcome variable modeled with a probability distribution that has maximum entropy,
2. a linear model that describes the shape-defining parameter of that distribution where
    1. the linear model contains all predictors necessary to estimate an unbiased effect of interest, and
    2. the shape-defining parameter only accepts valid values, and finally
3. feasible prior distributions for all coefficients in the model.

In the first model, use *complete pooling*. In the second model, use *partial pooling* on $\mathtt{interest}$ based on the fact that some students are generally more interested than others. Finally, explain the potential benefit of partial pooling.

Write your answer in the box below. Changes are saved automatically.

You can either write the model in mathematical notation (using characters like _, ~ and <-) or as you would specify the *alist* object in the *rethinking* package.

Maximum marks: 12