# Empirical Software Engineering

Write your answers directly on these pages (there's always a risk that loose papers disappear)—use the back also if possible. I'll be at the written exam twice (first time after approximately one hour).

On September 11 at 10.15 you are welcome to Richard's office (4th floor in the Jupiter building at Campus Lindholmen) to complain about the grading. Before the meeting you *must* send Richard an email clearly pointing out where you think the error is, what you wrote, and why you believe the grading was not correct. If I don't receive such an email before 10.15 on September 11, then I will not meet with you.

Repetition is the mother of all knowledge.

— Richard Torkar

Grade 3: 42 points; ~50%
Grade 4: 57 points; ~70%
Grade 5: 74 points; ~90%
Maximum: 82 points

**Question 1 :**

(**8p**) Over the years Bayesian data analysis has evolved and spread to all natural sciences. There are several reasons for this, e.g., a principled way to incorporate prior knowledge, increase in computational power making Markov chain Monte Carlo a viable option for sampling, and probabilistic programming languages gaining ground. However, not until lately have statisticians developed guidelines researchers can follow to systematically design Bayesian models.

In *your* opinion, what key steps are compulsory when conducting Bayesian data analysis? Please explain each step one takes when designing models so that we can place *some* confidence in the results.

You can either draw a flowchart and explain each step, or write a numbered list explaining each step.

(It's ok to write on the backside, if they haven't printed on the backside again...)

**Question 2 :**

(**4p**) Underfitting and overfitting are two concepts not always stressed a lot with black-box machine learning approaches. In this course, however, you've probably heard me talk about these concepts a hundred times...

What happens when you underfit and overfit, i.e., **what would the results be**? What are some principled **ways to deal with** under- and overfitting?

**Question 3 :**

(**10p**) To understand **how team size affects psychological safety** the following data was collected:

| Team | Team size | SPI | Psychological safety |
|---|---|---|---|
| 1 | 5 | 67% | High |
| 2 | 15 | 33% | Low |
| 3 | 11 | 49% | Low |
| 4 | 7 | 90% | High |
| ⋮ | ⋮ | ⋮ | ⋮ |

The experiment started with assuming that planning effectiveness and psychological safety has a very strong association. For planning effectiveness they used schedule performance indicator (SPI) as a stand-in variable.

Schedule Performance Indicator = (Completed points / Planned points)

Based on the result, if the SPI is more than 50% they are classified as a team with high psychological safety. If less than 50% they are classified as a team with low psychological safety.

With the above data, the firm wants to use your knowledge to understand the association between team size and psychological safety.

Write down the mathematical model definition for this prediction using any variable names and priors of your choice.

State the ontological and epistemological reasons for your likelihood. Remember to clearly state and justify the choices and assumptions regarding your model.

**Question 4 :**

(**3p**) What is **epistemological justification** and how does it differ from **ontological justification**, when we design models and choose likelihoods? Please provide **an example** where you **argue** epistemological and ontological reasons for selecting a likelihood.

**Question 5 :**

(**4p**) When diagnosing Markov chains, we often look at several diagnostics to form an opinion of how well things have gone. Name four diagnostics we commonly use? What do we **look for** in each diagnostics (i.e, what thresholds or signs do we look for)? Finally, **what do they tell us**?

**Question 6 :**

(**4p**) Explain the four main benefits of using multilevel models.

Table 1: Output from running WAIC on three models.

| | WAIC | SE | dWAIC | dSE | pWAIC |
|---|---|---|---|---|---|
| m1 | 127.6 | 14.69 | 0.0 | NA | 4.7 |
| m3 | 129.4 | 15.10 | 1.8 | 0.90 | 5.9 |
| m2 | 140.6 | 11.21 | 13.1 | 10.82 | 3.8 |

**Question 7 :**

(**4p**) As a result of comparing three models, we get the above output. What does each column (WAIC, SE, dWAIC, dSE, and pWAIC) **mean**? **Which** model would you **select** based on the output?

**Question 8 :**

(**5p**) Write an example mathematical model formula for a Poisson regression model with two different kinds of varying intercepts, also known as a cross-classified model.

**Question 9 :**

(**4p**) Explain the terms in your own words:

- prior
- posterior
- information entropy
- instrumental variable

**Question 10 :**

(**2p**) What are the two kinds of varying effects? Explain the effect they have on a statistical model.

**Question 11 :**

(**8p**) What are the four elemental confounds on which any Directed Acyclic Graph can be explained?

Please draw the four confounds. Explain what they mean (preferably by explaining if one should condition or not on certain elements).

**Question 12 :**

(**4p**) What is the **purpose** and **limitations** of using laboratory experiments and experimental simulations as a research strategy?

**Question 13 :**

(**11p**) A common research method in software engineering that is used to complement other methods is survey research. Here follows a number of questions connected to survey research:

1. We often differ between reliability and validity concerning surveys,
   (a) What is the difference between the reliability and validity in survey design? (2p)
   (b) Name and describe at least two types of reliability in survey design. (3p)
   (c) Name and describe at least two types of validity in survey design. (3p)
2. Even if you measure and estimate reliability and validity you still want to *evaluate* the survey instrument. Which are the two (2) common ways of evaluating a survey instrument? Explain their differences. (3p)

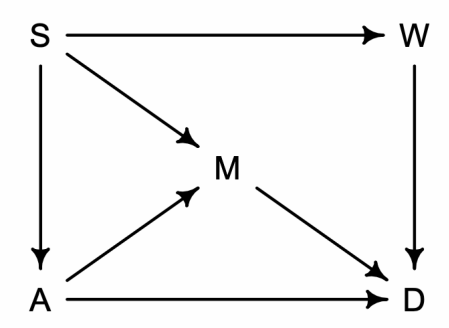


Figure 1: A messy Directed Acyclic Graph.

**Question 14 :**

(**6p**) Look at the DAG above. We want to estimate the total causal effect of $W$ on $D$. Which variable(s) should we condition on and why?

**Question 15 :**

(**5p**) You get one point if you answer a question correctly. Simply write your answer below.

1. We should always start the Bayesian data analysis by designing a ....
2. Adding predictors to a model can lead to several things. Two common things are ...
3. My $\widehat{R}$ value is 1.04. I should ...
4. We can quantify ... using Kullback-Leibler divergence.
5. I am first and foremost always interested in propagating ... while doing BDA.

Certum est.