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.

On January 20 at 09.00–10.00 you are welcome to room 6217 in the EDIT house (Johanneberg), with questions 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 09.00 on January 20, I will not meet with you.

Richard will be at the written exam twice (first time after approximately one hour).

Menschen vergehen, aber ihre Taten bleiben.

— Baron Augustin-Louis Cauchy

Grade 3: 36 points; ∼50%
Grade 4: 40 points; ∼70%
Grade 5: 66 points; ∼90%
Maximum: 73 points

Question 1 :

(**8p**) In *your* opinion, which **steps** are compulsory when conducting Bayesian data analysis? Please **explain** what steps one take when designing models, so that we ultimately 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...)

Solution: There is no standard. However, quite recently Gelman *et al.* (arXiv:2011.01808) published a manuscript on arXiv explaining what they believed were the key parts in Bayesian data analysis. In short, a lot of the things they talk about in that paper are things we have covered in this course.

Start with a null model, do prior checks, check diagnostics, do posterior checks, then conduct inferential statistics if needed. Also, it would be good if you mention **comparisons** of models and that it is an **iterative** approach. Of course one could always add more...

Question 2 :

(**9p**) In Bayesian data analysis we have three principled ways of avoiding overfitting.

The **first** way makes sure that the model doesn't get too excited by the data. The **second** way is to use some type of scoring device to model the prediction task and estimate predictive accuracy. The **third**, and final way is to, if suitable data is at hand, design models that actually do something about overfitting.

Which are the three ways? (3p)

Explain and provide examples for each of the three ways (2p+2p+2p)

Solution: In the first way we use regularizing priors to capture the regular features of the data. In particular it might be wise to contrast this with sample size!

In the second way we rely on information theory (entropy and max ent distributions) and the concept of information criteria, e.g., BIC, DIC, AIC, LOO, and WAIC.

For the final, third way, the below is a good answer:

complete pooling

$$y_i \sim \text{Binomial}(n, p_i)$$
$$\text{logit}(p_i) = \alpha$$
$$\alpha \sim \text{Normal}(0, 2.5)$$

no pooling

$$y_i \sim \text{Binomial}(n, p_i)$$
$$\text{logit}(p_i) = \alpha_{\text{CLUSTER}[j]}$$
$$\alpha_j \sim \text{Normal}(0, 2)$$

partial pooling using hyper-parameters and hyper-priors

$$y_i \sim \text{Binomial}(n, p_i)$$
$$\text{logit}(p_i) = \alpha_{\text{CLUSTER}[j]}$$
$$\alpha_j \sim \text{Normal}(\bar{\alpha}, \sigma)$$
$$\bar{\alpha} \sim \text{Normal}(0, 1)$$
$$\sigma \sim \text{Exponential}(1)$$

Question 3 :

(**3p**) It is possible to view the Normal likelihood as a purely epistemological assumption, rather than an ontological assumption. How would you argue such a statement?

Solution: See *Statistical Rethinking*, 2nd ed., pp. 81

Question 4 :

(**8p**) **Name** at least four distributions in the exponential family (4p). **Provide examples** of when one can use each distribution when designing statistical models and **explain** what is so special about each distribution you've picked, i.e., their assumptions (4p).

Solution: Take your pick, Gamma, Normal, Exponential, Poisson, Binomial, are some of the distributions...

Question 5 :

(**5p**) Below you see a Generalized Linear Model

$$y_i \sim \text{Binomial}(n, p)$$
$$f(p) = \alpha + \beta x_i$$

What is $f()$ called **and** why is it needed? (2p)

What should $f()$ be in this case? (1p)

Does $f()$ somehow affect your priors? If yes, **provide an example**. (2p)

Solution: Generalized linear models need a *link function*, because rarely is there a "μ", a parameter describing the average outcome, and rarely are parameters unbounded in both directions, like μ is. For example, the shape of the binomial distribution is determined, like the Gaussian, by two parameters. But unlike the Gaussian, neither of these parameters is the mean. Instead, the mean outcome is np, which is a function of both parameters. For Poisson we have λ which is both the mean and the variance.

The link function f provides a solution to this common problem. A link function's job is to map the linear space of a model like $\alpha + \beta x_i$ onto the non-linear space of a parameter!

In the above case the link function is the logit(). If you set priors and use a link function the often concentrate probability mass in unexpected ways.

Question 6 :

(**6p**) What is the **purpose** and **limitations** of using *Judgement Studies* and *Sample Studies* as a research strategy? Provide **examples**, i.e., methods for each of the two categories, and **clarify** if one use mostly qualitative or quantitative approaches (or both).

Solution:

Judgement Studies: Neutral setting

Judge or rate behaviors

Respond to stimulus

Discuss a topic of interest

Setting: Neutral setting

Actors: Systematic sampling

Goal: Seek generalizability over the responses not over the population of actors.

Responses are not specific to any context.

Unobtrusive and no control.

Examples of research methods:

Delphi study

Panel of experts

Focus group

Evaluation study

Interview studies

Qualitative and/or quantitative data

Sample Studies: Study: the distribution of a characteristic of a population or the correlation between two or more characteristics. Setting: Neutral setting Actors: Systematic sampling

Goal: Generalizability over a population of actors Responses are not specific to any context Unobtrusive and no control

It does not need to be sample of humans. It can be of artifacts such as projects, models, etc...

Examples of research methods:

Surveys

Systematic literature reviews

Software repository mining

Interviews

Mostly quantitative data but can include qualitative data

Question 7 :

(**8p**) Below follows an abstract from a research paper. Answer the questions,

- Which of the eight research strategies presented in the ABC framework does this paper likely fit? **Justify and argue!**
- Can you argue the main validity threats of the paper, based on the research strategy you picked?
  - It would be very good if you can **list threats in the four common categories** we usually work with in software engineering.

Requirement prioritization is recognized as an important decision-making activity in requirements engineering and software development. Requirement prioritization is applied to determine which requirements should be implemented and released. In order to prioritize requirements, there are several approaches/techniques/tools that use different requirements prioritization criteria, which are often identified by gut feeling instead of an in-depth analysis of which criteria are most important to use.

In this study we investigate which requirements prioritization criteria are most important to use in industry when determining which requirements are implemented and released, and if the importance of the criteria change depending on how far a requirement has reached in the development process. We conducted a quantitative study of one completed project from one software developing company by extracting 32,139 requirements prioritization decisions based on eight requirements prioritization criteria for 11,110 requirements. The results show that not all requirements prioritization criteria are equally important, and this change depending on how far a requirement has reached in the development process.

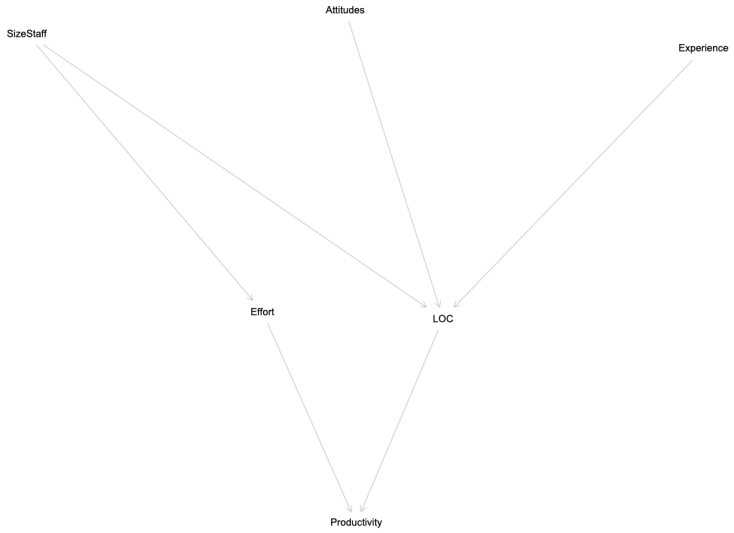
Solution: A reply from one of the authors of the ABC paper:

> This is a good one—and quite a common one as well. It's not straightforward, and I don't think there is a crystal-clear 'right' answer, and I've struggled a bit with that in the past—but I've come to the realization that it's "OK" that there is no right answer *per se*, but rather that at least we have some anchor points to have the discussion; the root question of course is:
>
> Is this a quantitative case study, i.e., a case study using a sample study as strategy, or vice versa, a sample study with data from a single case study?
>
> It very much depends I think on the nature of the sample, and whether that is a very specific sample to the company. If the sampling allows generalizing to beyond the company, then one could argue it's a **sample study**.

Question 8 :

(**12p**) In software engineering research, which has taken inspiration from the social sciences, we often classify threats to validity in four categories. Which are the four categories (4p)? Please **explain** what question(s) each category tries to answer and provide an **example** for each category (8p).

Solution: See `Validity threats.pptx`



Question 9 :

(**6p**) See the DAG above. We want to estimate the total causal effect of *SizeStaff* on *Productivity*. What should we **condition on**? (1p)

Design a **complete** model, in math notation, where the outcome *Productivity* is a count $0, \ldots, \infty$, and **include the variable(s)** needed to answer the above question. Also add, what you believe to be, **suitable priors on all parameters**. State any assumptions! (5p)

Solution: 4p for appropriate model as further down below.

```
d <- dagitty('dag{
      bb="0,0,1,1"
    Attitudes [pos="0.474,0.305"]
    Effort [exposure,pos="0.414,0.522"]
    Experience [pos="0.734,0.333"]
    LOC [pos="0.545,0.527"]
    Productivity [outcome,pos="0.482,0.674"]
    SizeStaff [pos="0.243,0.322"]
    Attitudes -> LOC
    Effort -> Productivity
    Experience -> LOC
    LOC -> Productivity
    SizeStaff -> Effort
    SizeStaff -> LOC
    }')
r$> adjustmentSets(d, exposure = "SizeStaff", outcome = "Productivity")
 {}
```

and the answer is `{}`, i.e., nothing (well except for SizeStaff obviously)! In short, check what elementary confounds we have in the DAG (DAGs can only contain four), follow the paths, decide what to condition on.

$$\begin{aligned}
\text{Productivity}_i &\sim \text{Poisson}(\lambda_i) \\
\log(\lambda_i) &= \alpha + \beta_{\text{ss}}\text{SizeStaff}_i \\
\alpha &\sim \text{Normal}(0, 2.5) \\
\beta_{\text{ss}} &\sim \text{Normal}(0, 0.2)
\end{aligned}$$

We'd like to see a slightly broader prior on α (what could productivity be, well it's a count and I assume that it's probably not in the millions), and a significantly more narrow prior on β since we have a log link, i.e., a Normal(0,0.2) still allow values as high as ~ 2.5, which is perfect for a β (that would be a large effect after all!)

Question 10 :

(**8p**) Below follows a number of multiple choice questions. There can be more than one correct answer! Mark the correct answer by crossing the answer. In the case of DAGs, X is the treatment and Y is the outcome.

Q1 What is this construct called: $Y \leftarrow Z \leftarrow X$?
{Collider} {Pipe} {Fork} {Descendant}

Q2 What is this construct called: $Y \rightarrow Z \leftarrow X$?
{Collider} {Pipe} {Fork} {Descendant}

Q3 We should never condition on Z: $Y \leftarrow Z \rightarrow X$?
{True} {False}

Q4 If we condition on Z we close the path $X \rightarrow Z \leftarrow Y$.
{True} {False}

Q5 What is this construct called: $Y \leftarrow Z \rightarrow X$?
{Collider} {Pipe} {Confounder} {Descendant}

Q6 An interaction is an influence of predictor...
{on a parameter} {conditional on other predictor} {conditional on parameter}

Q7 What distribution maximizes this? $H(p) = -\sum_{i=1}^{n} p_i \log(pi)$
{Most complex} {Most structured} {Flattest} {Distribution that can happen the most ways}

Q8 What distribution to pick if it's a real value in an interval between 0 and 1, but where neither 0 nor 1 are allowed?
{Exponential} {Log-Normal} {Beta}

Q9 What distribution to pick if it's a real value with finite variance?
{Binomial} {Negative-Binomial} {Normal}

Q10 Dichotomous variables, varying probability?
{Binomial} {Beta-Binomial} {Negative-Binomial/Gamma-Poisson}

Q11 A 2-parameter distribution with a shape parameter k and a scale parameter θ for positive real values $0, \ldots, \infty$?
{Binomial} {Beta} {Gamma}

Q12 Natural value, positive, mean and variance are equal?
{Gamma-Poisson} {Poisson} {Cumulative}

Q13 Unordered (labeled) values?
{Categorical} {Cumulative} {Dinominal}

Q14 On which effect scale are parameters?
{Absolute} {Relative} {None}

Q15 On which effect scale are predictions?
{None} {Absolute} {Relative}

Q16 In *zi* models we assume the data generation process consist of two disparate parts?
{Yes} {No}