Take-Home Problems

Instructions.

- These problems are due **on Wednesday, May 5, 2021**.
- You may use your notes and homework, the class text, any online materials from our class, and any resources linked from within the test itself. You may not consult other resources or other people. Send me an email if you have questions or difficulties of any sort.
- Prepare your solutions using R Markdown and submit via Gradescope. Also email me your Rmd file. (I usually don't need these, but every once in a while it is handy to have the.)

1 A scientist treated 7 plants of the species *Brassica campestris* with the substance Ancymidol (Ancy) and compared them to 8 plants that were not treated. Height (in cm) after 14 days of growth was measured for each plant. The mean for the Ancy group was 11.0; for the control group it was 15.9. We want to know whether this is evidence of a true population difference and we will analyze the data three ways.

The data are located in `CalvinBayes::Ancy`.

```
     group group_num height
1  Control         1   10.0
2  Control         1   13.2
3  Control         1   19.8
4  Control         1   19.3
5  Control         1   21.2
6  Control         1   13.9
7  Control         1   20.3
8  Control         1    9.6
9     Ancy         2   13.2
10    Ancy         2   19.5
11    Ancy         2   11.0
12    Ancy         2    5.8
13    Ancy         2   12.8
14    Ancy         2    7.1
15    Ancy         2    7.7
```

In a frequentist course, we might analyze this with a two-sample t test. John Kruschke book wrote a paper called *Bayesian Estimation Supersedes the t-test* (a.k.a., BEST) in which he outlines why he prefers Bayesian methods for this situation. To make it easy for people to follow the advice of this paper, Rasmus Bååth coded up an online app to do the work. You can find it at `http://www.sumsar.net/best_online/`.

Part 1

Enter the data from the Ancy study into the app. There is a box for each group. Then hit "Click to restart". The app takes care of the rest.

a) The underlying model has 5 parameters. What are they? (You may want to return to this question after doing some of the questions below and exploring the app a bit.)

b) What is the 95% HDI for the difference in mean heights? How would you interpret that for a biologist interested in the results of this study?

c) What is the posterior probability that the Ancy mean is less than the control mean? How would you interpret that for a biologist interested in the results of this study?

d) Does it appear that the two groups have the same standard deviations? (Don't just say yes or no, explain how you are getting your answer.)

e) The app says there is a "measure is labeled 'Normality' below where a normality estimate ≤ 1.5 indicates that the data isn't normally distributed." But it doesn't specify what that measure is exactly. The paper does (on page 8): It's $\log_{10}(\nu)$, where ν is the degrees of freedom of a t-distribution.

If $\log_{10}(\nu) = 1.5$, what is ν?

f) Make a plot with the densities of 3 distributions: (a) a normal distribution, (b) a t-distribution with ν degrees of freedom where $\log_{10}(\nu) = 1.5$, and (c) a t-distribution with ν degrees of freedome where $\log_{10}(\nu) = 1$. How does this relate to the app's comment about the number 1.5?

g) Does it appear that heights are normally distributed in each group? (Don't just say yes or no, explain how you are getting your answer.)

Part 2

a) The paper claims that it uses "weakly informed priors". What is a "weakly informed prior"?

b) The paper actually gives the details for the prior for ν in appendix A: "Mathematically, it is $p(\nu \mid \lambda) = (1/\lambda)\exp(-(\nu-1)/\lambda)$ for $\nu \geq 1$ and $\lambda = 29$."

Let me translate this for you: The prior for $\nu - 1$ is $\mathsf{Exp}(1/29)$. (Remember exponential distributions only take on positive values, but we want $\nu > 1$; that explains why 1 is subtracted.)

i. What is the mean of an $\mathsf{Exp}(1/29)$ distribution?

ii. What is the mean of the prior for ν if the prior for $\nu - 1$ is $\mathsf{Exp}(1/29)$?

iii. What is the prior probability that $\nu \geq 30$?

Note: he also mentions that "several [other priors] ... were considered, including various uniform distributions, various shifted gamma distributions, and various shifted and folded t distributions. You do not need to try those (but you are welcome to experiment if you are curious to see how much it matters).

Part 3

a) Fit the model to our data using `ulam()`, the prior for ν described above, and weakly informed priors for the other parameters. Show a summary of your model using `precis()` or something similar.

b) Use your model to recreate the following 6 plots from the app: "Trace Plot - Difference of Means", "Distribution - Difference of Means", "Distribution - Mean Group 1", "Distribution - SD Group 1", "Distribution - Difference of SDs", "Distribution - Normality"

Your distribution plots don't need to be exact replicas, but they should (a) show the distribution using a histogram or density curve, (b) show the HDI, and (c) be labeled so I can tell which is what (use `gf_labs()`). You may add as much or little of the other stuff as you like.

Note: I don't need pages and pages of plots. Choose the sizes reasonably. Use patchwork to lay them out nicely so all six plots fit on the same page. (Use `\newpage` to introduce a page break if that helps.)

c) Calculate the posterior probability that the Ancy mean is less than the control mean.

Note: Your results using R should be similar to (but perhaps not identical to) the results from the app. So you could answer all of the questions you answered using the app from your R output, *but don't redo those answers.* Just do the specific things asked for.

The point of Part 3 is not to reanswer the questions in Part 1– the answers should be essentially the same – but to show how to use `ulam()` to train the model and R to get the information you need to answer them.