Research Methods & Statistics 2021: Bayesian Exercises

Johnny van Doorn

University of Amsterdam

Below you will find various exercises on the intricate world of Bayesian inference, testing your knowledge and intuition about Bayesian parameter estimation and Bayesian hypothesis testing in the context of various statistical tests (binomial test, correlation, t-test). The exercises use statistical software JASP (https://jasp-stats.org). Each exercise lists links to webpages with either only JASP output (so you can also complete the exercises without installing JASP), or with annotated JASP output containing the solutions/explanation to the exercises (and additional ramblings that hopefully provide some extra intuition). The output files can also be found here in case the OSF does not work for you: https://surfdrive.surf.nl/files/index.php/s/OHjXsc79TkeYWgN

If you want to conduct an analysis in JASP, but get stuck, you can consult this playlist with 3 instructional videos (one for each of the analyses below): https://youtube.com/playlist?list=PLWPa8RxHarcOyxhMAk_RMT9h_GYHCrgGQ

Dislclaimer This is the first version of this document, so there could be some mistakes. If you find any, or want to give any kind of other feedback, please let me know by email (j.b.vandoorn@uva.nl).

Contents

1 Binomial Test

1.1 Therapeutic Touch

'Therapeutic Touch' (TT) is a nursing practice rooted in mysticism but alleged to have a scientific basis. Practitioners of TT claim to treat medical conditions by using their hands to manipulate a "human energy field" perceptible above the patients' skin. Being a skeptical mind, 9-year old Emily Rosa ventured to test these claims by designing an experiment where TT practitioners had to correctly identify the location of the experimenter's hand (above their left or right hand) while being blinded by a screen. The results of this experiment were later published in the prestigious Journal of the American Medical Association (Rosa et al., 1998). In the following exercise, you will evaluate this data set with JASP.

- JASP file with only the results: `https://osf.io/snu9r/`
- JASP file with results and interpretation/solutions: `https://osf.io/n4tcu/`

1. Open the "Emily Rosa" dataset from the JASP Data Library (you can find it in the folder "Frequencies"). Alternatively, use the link above to view the results without using JASP.
2. Get a descriptive overview of the data by producing a frequency table of the variable Outcome (Go to "Frequencies", then "Bayesian Binomial Test" and select the variable "Outcome"). How many practitioners guessed correctly? What is the proportion?

3. Produce a prior and posterior plot for a two-sided test and interpret the 95% credible interval displayed in it.

4. Conduct a Bayesian Binomial Test to test if the practitioners' ability (i.e., proportion of correct responses) is better than random (i.e., $\theta > 0.5$). Use the default prior for the alternative hypothesis. What is the Bayes factor in favor of the null hypothesis (BF_{0+})?

5. How would you interpret the Bayes factor?

6. Take a look at the sequential analysis plot. What was (approximately) the largest Bayes factor in favor of the alternative hypothesis that was ever reached in the sequential analysis of the data?

7. Imagine a previous experiment with 10 practitioners, where 8 therapists gave the correct answer. Construct an informed prior distribution based on their results. How do the results of the hypothesis test change when compared to a default prior?

1.2 Psychologists Tasting Beer

Friday afternoon, May 12th 2017, an informal beer tasting experiment took place at the Department of Psychology, University of Amsterdam. The experiment was preregistered, the OSF directory with background information is https://osf.io/m3ju9/, and the corresponding article can be found at https://psyarxiv.com/d8bvn/.
Short description: Each participant was (blinded) given two cups of Weihestephan Hefeweissbier, one cup with the regular variety, one cup with the non-alcoholic version. Participants had to indicate which cup contained the alcohol, the confidence in their judgment, and an assessment of how much they liked the beer. The data set contains the following variables:

- AlcBeerFist: 1 = alcoholic version was tasted first; 0 = alcoholic version was tasted second. This is not likely to matter, as participants were allowed to sample the cups at will.
- CorrectIdentify: 1 = correct, 0 = incorrect.
- ConfidenceRating: Assessment of confidence in the identification judgment on a scale from 0 (complete guess) to 100 (certainty).
- AlcRating: Taste rating for the regular, alcoholic version, on a scale from 0 (worst beer ever) to 100 (best beer ever).
- NonAlcRating: Taste rating for the non-alcoholic version, on a scale from 0 (worst beer ever) to 100 (best beer ever).

In this exercise, you will conduct a Bayesian binomial test to analyze the proportion of correct responses.

- csv file with the data: https://osf.io/zh9md/
- JASP file with only the results: https://osf.io/azkvc/
- JASP file with results and interpretation/solutions: https://osf.io/8ec4g/

1. Download the csv file from the OSF and open the data in JASP. Alternatively, use the link above to view the results without using JASP.
2. Get a descriptive overview of the data by producing a frequency table of the variable Outcome (go to "Frequencies", then "Bayesian Binomial Test" and select the variable "CorrectIdentify"). What percentage of participants had a correct response ("1")?
3. Which settings can we use to specify our alternative hypothesis (see Figure 1 below for the available settings)? What do these settings imply?
4. Using a uniform prior distribution, what are the two-sided Bayes factors (BF_{10}) for the proportion of incorrect responses ("0") and for the proportion of correct responses ("1")? Are they equal? Explain why (not).
5. In order to test whether people are able to taste the difference, conduct a one-sided hypothesis test to test if $\theta > 0.5$. What is the Bayes factor (BF_{+0})? How do you interpret this? What is your verdict on the claim that psychologists can taste the difference between alcoholic and non-alcoholic beer?
6. Conduct a Sequential analysis to inspect how the Bayes factor developed as the data accumulated. Did it ever provide strong evidence for the null hypothesis? Around which value did the Bayes factor cross the threshold of 100?
7. Imagine being someone who is very confident in psychologists beer tasting ability, because they once did their own experiment and saw that 10 out of 11 could taste the difference. What would their prior distribution look like? How does this affect the one-sided Bayes factor (BF_{+0}) for the proportion of correct responses?

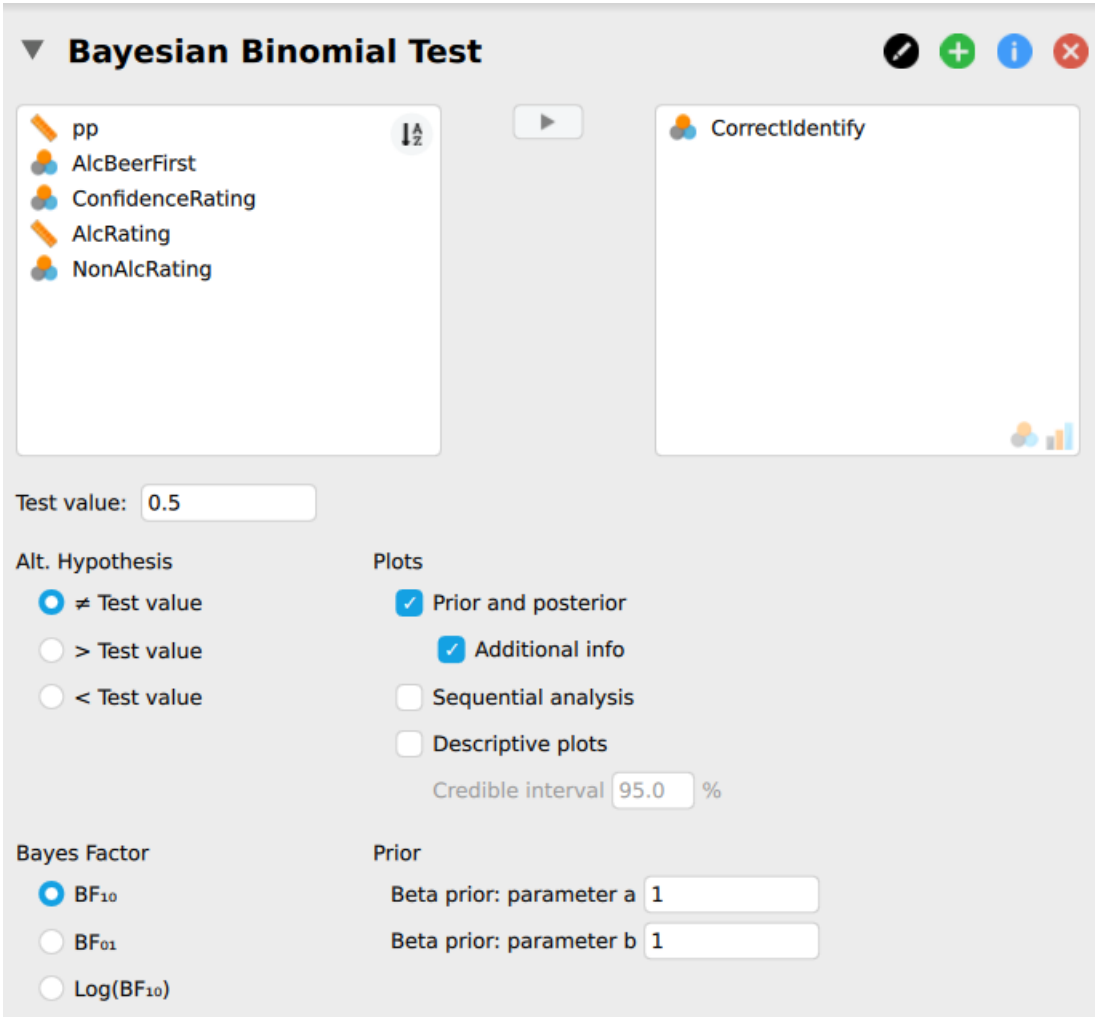

Figure 1: The options available in the Bayesian binomial test - which settings influence the alternative hypothesis?

2 Correlation

2.1 Correlation: A.W.E.S.O.M.-O 4000

In South Park episode 116, one of the series' main protagonists, Eric Cartman, pretends to be a robot from Japan, the "A.W.E.S.O.M.-O 4000". When kidnapped by Hollywood movie producers and put under pressure to generate profitable movie concepts, Cartman manages to generate thousands of silly ideas, 800 of which feature Adam Sandler. We conjecture that the makers of South Park believe that Adam Sandler movies are profitable regardless of their quality. In this exercise, we put forward the following South Park hypothesis: "For Adam Sandler movies, there is no correlation between box-office success and movie quality (i.e., 'freshness' ratings on Rotten Tomatoes; www.rottentomatoes.com)."

- JASP file with only the results: https://osf.io/xs9dq/
- JASP file with results and interpretation/solutions: https://osf.io/7fxpr/

1. Open the "Adam Sandler" dataset from the JASP Data Library (you can find it in the folder "Regression").
2. Produce a boxplot for the variables "Freshness" and "Box Office ($M)". Are there any outliers in the data?
3. Conduct a Bayesian Correlation Pairs test to test if there is any correlation is absent or present ($\rho = 0$ vs. $\rho \neq 0$). Use the default prior ("Stretched beta prior width" $= 1$) for the alternative hypothesis. What is the correlation coefficient r?
4. Produce a prior and posterior plot and interpret the median and 95% credible interval displayed in it.
5. How much did the posterior probability at $\rho = 0$ increase with regards to the prior probability? How are these points displayed in JASP? And how is this reflected in the result of the analysis?
6. What is the Bayes factor in favor of the null hypothesis (BF_{01})? How would you interpret the Bayes factor?
7. Now perform a one-sided test to test whether the correlation is positive ($\rho = 0$ vs. $\rho > 0$). How does this Bayes factor (BF_{+0}) differ from the two-sided Bayes factor(BF_{10}) – is it higher or lower?
8. Produce a Bayes Factor Robustness Plot. What is the effect of the prior width on the Bayes factor (BF_{0+})? What does this mean for our result?

9. Adam Sandler doesn't approve of our hypothesis, comes out and says that movie critics on Rotten Tomatoes purposely down-rate his awesome movies. He argues that there should be a negative correlation between box-office success and the quality of his movies. Perform a Bayesian Correlation Test with Adam's alternative hypothesis. Would you support him in his claim? Why (not)?

3 T-Test

3.1 The Effect of Directed Reading Exercises

A teacher believes that directed reading activities in the classroom can improve the reading ability of elementary school children. She convinces her colleagues to give her the chance to try out the new method on a random sample of 21 third-graders. After they participated for 8 weeks in the program, the children take the Degree of Reading Power test (DRP). Their scores are compared to a control group of 23 children who took the test on the same day and followed the same curriculum apart from the reading activities. In the following exercise, you will evaluate this dataset with JASP.

- JASP file with only the results: https://osf.io/85fbv/
- JASP file with results and interpretation/solutions: https://osf.io/xjbya/

1. Open the "Directed Reading Activities" dataset from the JASP Data Library (you can find it in the folder "T-Tests").
2. Get a descriptive overview of the data.
   (a) Create a table that shows the means and standard deviations of DRP scores in the control and treatment group (specify "drp" as the "Variable" and specify "group" as the "Split").
   (b) Create a boxplot that shows the distribution of DRP scores for each group.
3. Using the Bayesian independent samples t-test (use the variable "group" as the "Grouping Variable"), with the default prior distribution, obtain a posterior distribution for the standardized effect size δ. What is the 95% credible interval? Is 0 in this interval? What if you used a 99% credible interval?
4. How would you interpret the credible interval?
5. Now conduct a Bayesian independent samples t-test to test whether the control group performs worse than the treatment group (i.e., the mean of the control group is lower than the mean

of the treatment group). Use the default prior for the alternative hypothesis. What is the Bayes factor in favor of the alternative hypothesis (BF_{-0})?

6. How would you interpret the Bayes factor?

7. Create a Bayes factor robustness plot.

   (a) What is the maximum Bayes factor in favor of the alternative hypothesis that you could achieve by tinkering with the scale parameter of the Cauchy distribution?

   (b) Which scale parameter corresponds to the maximum Bayes factor?

   (c) Is the default prior that you used wider or narrower than the prior corresponding to the maximum Bayes factor?

   (d) Would you say that the Bayes factor is robust against different formulations of the prior?

3.2 Psychologists Tasting Beer 2: T-Test Boogaloo

Friday afternoon, May 12th 2017, an informal beer tasting experiment took place at the Department of Psychology, University of Amsterdam. The experiment was preregistered, the OSF directory with background information is https://osf.io/m3ju9/, and the corresponding article can be found at https://psyarxiv.com/d8bvn/.
Short description: Each participant was (blinded) given two cups of Weihestephan Hefeweissbier, one cup with the regular variety, one cup with the non-alcoholic version. Participants had to indicate which cup contained the alcohol, the confidence in their judgment, and an assessment of how much they liked the beer. The data set contains the following variables:

- AlcBeerFist: 1 = alcoholic version was tasted first; 0 = alcoholic version was tasted second. This is not likely to matter, as participants were allowed to sample the cups at will.
- CorrectIdentify: 1 = correct, 0 = incorrect.
- ConfidenceRating: Assessment of confidence in the identification judgment on a scale from 0 (complete guess) to 100 (certainty).
- AlcRating: Taste rating for the regular, alcoholic version, on a scale from 0 (worst beer ever) to 100 (best beer ever).
- NonAlcRating: Taste rating for the non-alcoholic version, on a scale from 0 (worst beer ever) to 100 (best beer ever).

In this exercise, we revisit this data set but instead apply a series of t-tests in order to answer two questions. First, was there a meaningful difference in confidence between people that were

correct and people that were incorrect? Second, did people like the alcoholic beer better than the non-alcoholic beer?

- csv file with the data: https://osf.io/zh9md/
- JASP file with only the results: https://osf.io/xjdyt/
- JASP file with results and interpretation/solutions: https://osf.io/36rgq/

1. What sort of t-test (e.g., independent samples or paired samples) do we need for the first research question? And for the second research question?
2. Download the csv file from the OSF and open the data in JASP.
3. Conduct a Bayesian independent samples t-test on the confidence ratings, and group by "CorrectIdentify". Tick the boxes to get a table with the descriptive statistics, and a raincloud plot (displayed horizontally or vertically). What can you learn from these statistics?
4. What is the one-sided Bayes factor (BF_{-0}), where the alternative hypothesis postulates that confidence ratings are *higher* for the correct responses than for the incorrect responses?
5. What can we conclude about the confidence ratings?
6. Conduct a paired samples t-test in order to test whether participants preferred the taste of the alcoholic beer over the taste of the non-alcoholic beer. To do so, go to "Bayesian Paired Samples T-Test" and then drag the two variables with the ratings ("AlcRating" and "NonAlcRating") to the "Variable Pairs" box.
7. What is Bayes factor comparing the one-sided alternative hypothesis (which postulates that alcoholic beer is tastier than non-alcoholic beer) to the null hypothesis?
8. Tick the box "Descriptives" under "Additional Statistics" to get a table with some descriptive and inferential statistics for each of the two variables. What is the difference between the two means?
9. What can you conclude about the taste of the two beers?

4 Summary Stats

4.1 T-Test: Flag Priming

In this exercise, we conduct a reanalysis of the article "A Single Exposure to the American Flag Shifts Support Toward Republicanism up to 8 Months Later" (Carter, Ferguson, & Hassin, 2011). Carter et al. (2011):

> "(...) tested whether a single exposure to the American flag would lead participants to shift their attitudes, beliefs, and behavior in the politically conservative direction. We conducted a multisession study during the 2008 U.S. presidential election."

The study featured various outcome measures and analyses, but for the purposes of this exercise we focus on the following result:

> "As predicted, participants in the flag-prime condition ($M = 0.072, SD = 0.47$) reported a greater intention to vote for McCain than did participants in the control condition ($M = -0.070, SD = 0.48$), $t(181) = 2.02, p = .04, d = 0.298$."

We assume that the total sample size was 183, and that there were 92 people in the flag-prime condition and 91 in the control condition. How strong is the evidence that corresponds to this "$p = .04$" result?

- JASP file with only the results: https://osf.io/tm6us/
- JASP file with results and interpretation/solutions: https://osf.io/p59fs/

1. Use the *Summary Stats* module in JASP to conduct a two-sided Bayesian independent samples t-test. Interpret the strength of evidence. Additionally, produce a plot of the prior and posterior distribution for the standardized effect size.
2. Compare the Bayesian result from the previous question to the frequentist p-value. What is the p-value here? Does this p-value lead to a different conclusion? Assume an α level of 0.05.
3. Since the authors' hypothesis is clearly directional, conduct a one-sided Bayesian independent samples t-test (where the alternative hypothesis states that the flag group scored greater than the control group). Interpret the strength of evidence for the authors' claim.
4. Conduct a Bayes factor robustness check analysis to investigate the strength of evidence for the authors' claim across a number of different prior settings. What is the most evidence we could have in favor of the one-sided alternative hypothesis?
5. Discuss your results with your classmate/neighbor/cat – do they agree?

References

Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011). A single exposure to the american flag shifts support toward republicanism up to 8 months later. *Psychological Science*, *22*, 1011–1018.

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General*, *144*, e1–e15.