Session 4: Intro to Bayesian Methods

Thomas J. Faulkenberry, Ph.D.

Tarleton State University

TLU R Workshop – 2018

Plan for the afternoon

- convince you that p-values are problematic
- introduce Bayesian inference
- show you some free software for conducting Bayesian inference (JASP)
- do some examples together

An exercise

Suppose you have a treatment that you suspect may alter performance on a certain task. You find that the experimental group **significantly** outperforms the control group, $t(18) = 2.7$, $p = 0.01$.

How many of the following are True statements?

- The probability that the null hypothesis is true is 1%
- The probability that the research hypothesis is true is 99%
- If you reject the null, the probability that you are making the wrong decision is 1%
- If you repeated the experiment over and over, then you would get a significant result 99% of the time.

An exercise

Suppose you have a treatment that you suspect may alter performance on a certain task. You find that the experimental group **significantly** outperforms the control group, $t(18) = 2.7$, $p = 0.01$.

They're all false!

So what is a p-value?

The **p-value** is the probability of obtaining a test statistic at least as extreme as the one that was observed, given that the null hypothesis is true. . . i.e., $P(D \mid \mathcal{H}_0)$.

Core logic:

- suppose the opposite of what we want (i.e., that there is no effect)
- show that this *under this null hypothesis*, our data is very unlikely (our p-value is very small)
- conclude that there must be an effect

Formal logic

This argument is similar to the *modus tollens* syllogism from formal logic:

- **Premise**: If A, then B;
- **Premise**: not B;
- **Conclusion**: therefore, not A.

Formal logic

Example of modus tollens:

- **Premise**: If Anabel is happy, then she is smiling;
- **Premise**: Anabel is not smiling;
- **Conclusion**: therefore, Anabel is not happy.

Formal logic

Fisher's disjunction (logic of hypothesis testing):

- **Premise**: If $\mathcal{H}_0$, then not `data`;
- **Premise**: `data`;
- **Conclusion**: therefore, not $\mathcal{H}_0$.

Hypothesis testing

In hypothesis testing, we use a *probabilistic* version of Fisher's disjunction:

- **Premise**: If $\mathcal{H}_0$, then `data` very unlikely;
- **Premise**: `data`;
- **Conclusion**: therefore, $\mathcal{H}_0$ very unlikely.

Hypothesis testing

In hypothesis testing, we use a *probabilistic* version of Fisher's disjunction:

- **Premise**: If $\mathcal{H}_0$, then `data` very unlikely;
- **Premise**: `data`;
- **Conclusion**: therefore, $\mathcal{H}_0$ very unlikely.

Problem: this is not logically valid!

Hypothesis testing

- **Premise**: If an individual is a man, he is unlikely to be Pope;

- **Premise**: Francis is the Pope;

- **Conclusion**: therefore Francis is probably not a man.

"Marginal" p-values

You conduct a high powered experiment and find $p = 0.045$.

- What do you do?

- What do you conclude?

"Marginal" p-values

Some context: suppose we want to test whether a *brain training* program had a significant effect on IQ scores. Let μ equal the population mean IQ score **after** training. Recall that IQ scores are normally distributed with mean 100 and standard deviation 15.

We can define two competing hypotheses:

1. $\mathcal{H}_0$: $\mu = 100$

2. $\mathcal{H}_1$: $\mu \neq 100$

Suppose further that the training had a moderate effect (i.e., Cohen's $d = 0.6$). This would imply that under $\mathcal{H}_1$, μ would be equal to ______

What p-values should we *expect*?

We'll run a simulation in R. On each run, we'll:

- take a sample of $n = 50$ IQ scores
- compare the sample mean to hypothesized mean of 100
- save the p-value from the test

Let's do this 50,000 times and plot the **distribution of p-values**:

- assuming $\mathcal{H}_0$ is true
- assuming $\mathcal{H}_1$ is true

Simulation in R

Question

You conduct a high powered experiment and find $p = 0.045$.

- How surprised should we be to see $p = 0.045$?

How surprising?

From simulation, we see:

- $p = 0.045$ is surprising when there is no effect
- $p = 0.045$ is EVEN MORE SURPRSING when there IS an effect!

A paradox?

This is known as *Lindley's Paradox*

- do I **reject** $\mathcal{H}_0$ (since $p < 0.05$)?
- or do I **accept** $\mathcal{H}_0$, since $p = 0.045$ is twice as likely under $\mathcal{H}_0$ as it is under $\mathcal{H}_1$?

Criticisms

Nelder (1999)

> "The most important task before us in developing statistical science is to demolish the p-value culture, which has taken root to a frightening extent in many areas of both pure and applied science, and technology. (p. 261)"

Nelder, J. A. (1999). From statistics to statistical science. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *48*, 257-269. `doi: 10.1111/1467-9884.00187`

Criticisms

Lindley (1999)

> "My personal view is that p-values should be relegated to the scrap heap and not considered by those who wish to think and act coherently. (p. 75)"

Lindley, D.V. (1999). Comment on Bayarri and Berger. In J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (Eds.), *Bayesian Statistics* (Vol. 6, p. 75). Oxford: Clarendon.

Criticisms

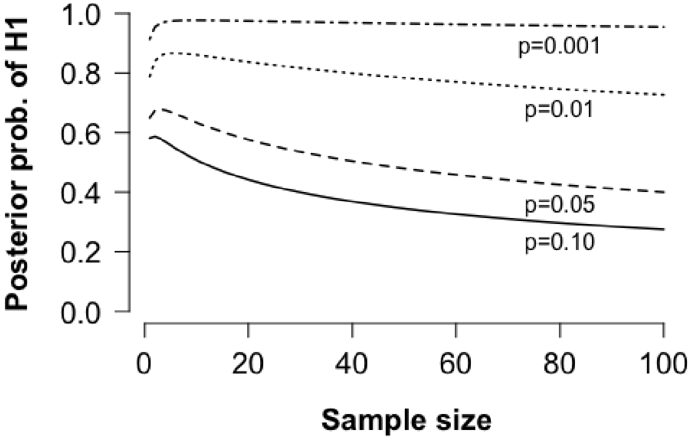
Townsend (2008)

> "[R]ecent years have seen a lively debate about the value of null hypothesis testing, in addition to various means of improving the strategy and avoiding its threatening sloughs of despond. (p. 271)"

Townsend, J. T. (2008). Mathematical Psychology: Prospects For The 21st Century: A Guest Editorial. *Journal of Mathematical Psychology*, *52*, 269–280. `doi:10.1016/j.jmp.2008.05.001`

Criticisms

Berger & Sellke (1987) - p-values **overstate** evidence against the null



Note: for a given p-value, that p-value becomes **less evidential** against $\mathcal{H}_0$ as sample size increases

An alternative

Use *Bayesian* inference

Bayesian inference

The machinery underlying Bayesian inference is *Bayes Theorem*:

$$p(\mathcal{H} \mid \text{data}) = \frac{p(\text{data} \mid \mathcal{H}) \cdot p(\mathcal{H})}{p(\text{data})}$$

Bayesian inference

Perhaps this is a more useful presentation:

$$\underbrace{p(\mathcal{H} \mid \text{data})}_{\substack{\text{Posterior beliefs}\\\text{about hypothesis}}} = \underbrace{p(\mathcal{H})}_{\substack{\text{Prior beliefs}\\\text{about hypothesis}}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H})}{p(\text{data})}}_{\text{predictive updating factor}}$$

Bayesian inference

Natural action in our discipline is to *compare* two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$.

- Bayes theorem gives us a natural way to do this by computing **relative** likelihoods

$$\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})} = \frac{\frac{p(\text{data}|\mathcal{H}_1)\cdot p(\mathcal{H}_1)}{p(\text{data})}}{\frac{p(\text{data}|\mathcal{H}_0)\cdot p(\mathcal{H}_0)}{p(\text{data})}}$$

Bayesian inference

Natural action in our discipline is to *compare* two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$.

- Bayes theorem gives us a natural way to do this by computing **relative** likelihoods

$$\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})} = \frac{p(\text{data} \mid \mathcal{H}_1) \cdot p(\mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0) \cdot p(\mathcal{H}_0)}$$

Bayesian inference

Natural action in our discipline is to *compare* two hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$.

- Bayes theorem gives us a natural way to do this by computing **relative** likelihoods

$$\underbrace{\frac{p(\mathcal{H}_1 \mid \text{data})}{p(\mathcal{H}_0 \mid \text{data})}}_{\substack{\text{posterior beliefs} \\ \text{about hypotheses}}} = \underbrace{\frac{p(\mathcal{H}_1)}{p(\mathcal{H}_0)}}_{\substack{\text{prior beliefs} \\ \text{about hypotheses}}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}}_{\text{predictive updating factor}}$$

Bayesian inference

The predictive updating factor

$$B_{10} = \frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}$$

tells us how much better $\mathcal{H}_1$ predicts our observed data than $\mathcal{H}_0$.

This ratio is called the **Bayes factor** (Jeffreys, 1961)

Bayes factors

Example: suppose $B_{10} = 5$.

Interpretation: the observed data are 5 times more likely under the alternative hypothesis $\mathcal{H}_1$ than the null hypothesis $\mathcal{H}_0$.

This is taken as **positive evidence** for the alternative $\mathcal{H}_1$

Bayes factors

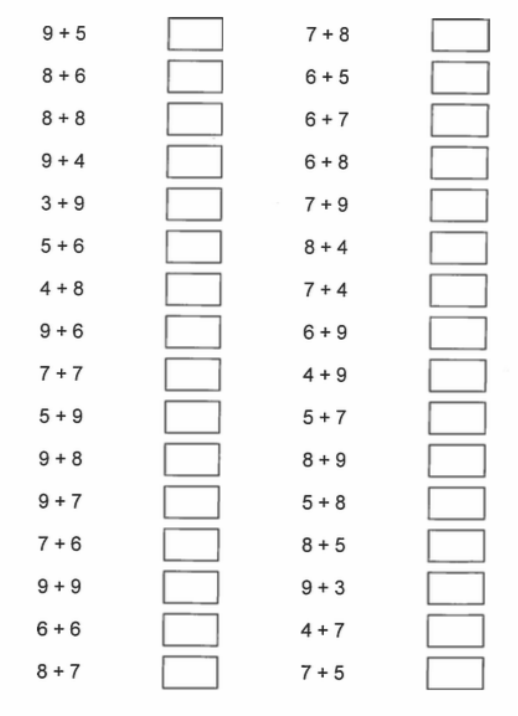
Jeffreys (1961) proposed the following thresholds for evidence:

Bayes factor	Evidence
1-3	anecdotal
3-10	moderate
10-30	strong
30-100	very strong
>100	extreme

www.jasp-stats.org

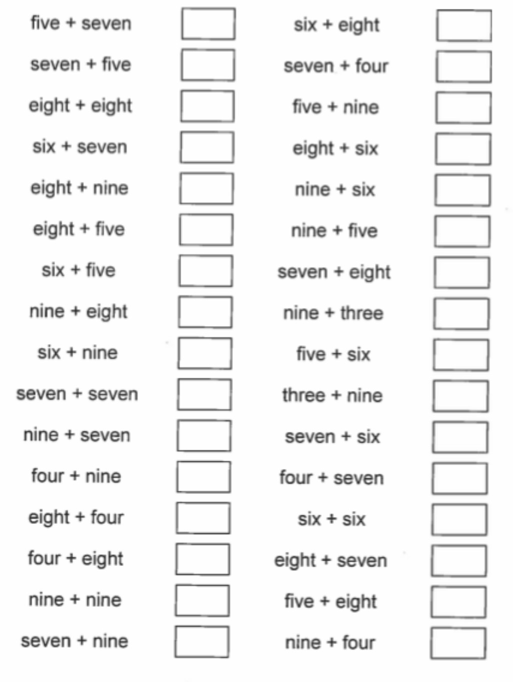
JASP demo

Ex 1 - Does problem format affect arithmetic performance?



9 + 5 □
8 + 6 □
8 + 8 □
9 + 4 □
3 + 9 □
5 + 6 □
4 + 8 □
9 + 6 □
7 + 7 □
5 + 9 □
9 + 8 □
9 + 7 □
7 + 6 □
9 + 9 □
6 + 6 □
8 + 7 □

7 + 8 □
6 + 5 □
6 + 7 □
6 + 8 □
7 + 9 □
8 + 4 □
7 + 4 □
6 + 9 □
4 + 9 □
5 + 7 □
8 + 9 □
5 + 8 □
8 + 5 □
9 + 3 □
4 + 7 □
7 + 5 □



five + seven □
seven + five □
eight + eight □
six + seven □
eight + nine □
eight + five □
six + five □
nine + eight □
six + nine □
seven + seven □
nine + seven □
four + nine □
eight + four □
four + eight □
nine + nine □
seven + nine □

six + eight □
seven + four □
five + nine □
eight + six □
nine + six □
nine + five □
seven + eight □
nine + three □
five + six □
three + nine □
seven + six □
four + seven □
six + six □
eight + seven □
five + eight □
nine + four □

Ex 1 - Does problem format affect arithmetic performance?

Your task:

- load `mental_arithmetic.csv` into JASP
- conduct an independent samples t-test
  - DV = `completed`
  - Grouping variable = `format`

Ex 1 - Does problem format affect arithmetic performance?

Now, let's do the Bayesian version

Ex 1 - Does problem format affect arithmetic performance?

Example writeup:

"We computed a Bayesian independent samples t-test to quantify the evidence for the hypothesis that participants can complete more problems in digit format than word format. We found a Bayes factor of $B_{10} = 19.2$, indicating that the observed data are approximately 19 times more likely under the alternative hypothesis than the null hypothesis. According to the classification scheme of Jeffreys (1961), this constitutes strong evidence for a format effect in arithemtic performance"

Ex 2 - Do horizontal saccades improve recall?

Several studies have shown that making horizontal eye movements after list learning improves episodic memory, possibly due to *interhemispheric interaction* (e.g., Christman & Propper, 2010).

We will look at some data from Matzke et al. (2015, JEP:G) and test two competing hypotheses:

- $\mathcal{H}_1$: horizontal $>$ fixation
- $\mathcal{H}_0$: horizontal $\leq$ fixation

Ex 2 - Do horizontal saccades improve recall?

First, let's do the traditional t-test

- load `data_eye_movements.csv` into JASP
- conduct an independent samples t-test
  - DV = `recall`
  - Grouping variable = `condition`
  - be sure to select "Group 1 < Group 2" for hypothesis

Ex 2 - Do horizontal saccades improve recall?

Now lets do the Bayesian version

Ex 2 - Do horizontal saccades improve recall?

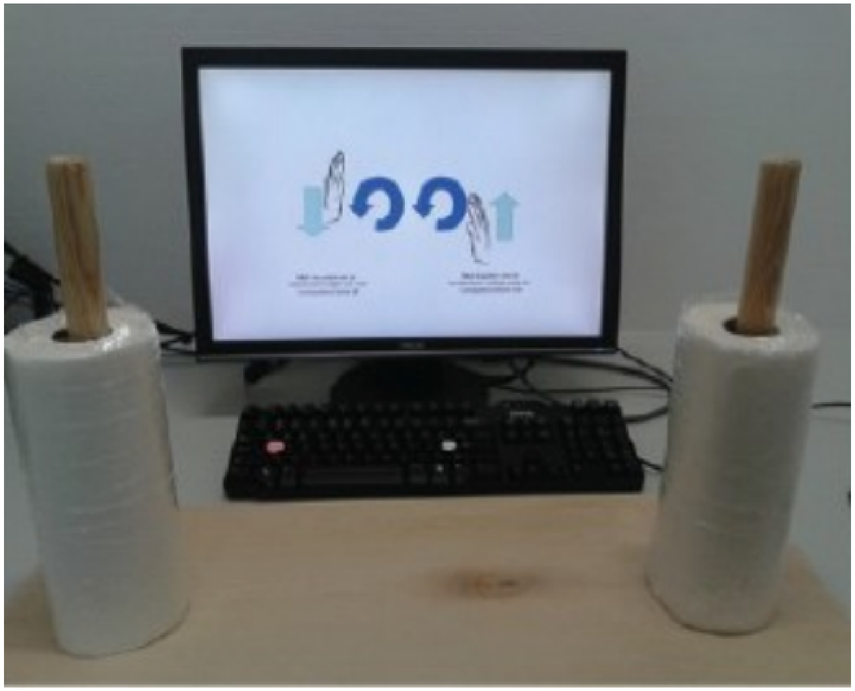
Example writeup:

"We computed a Bayesian independent samples t-test to quantify the evidence for thehypothesis that making horizontal saccades after initial study will increase memory recall. We found a Bayes factor of $B_{0+} = 11.7$, indicating that the observed data are approximately 12 times more likely under the null hypothesis than the alternative hypothesis. According to the classification scheme of Jeffreys (1961), this constitutes strong evidence for the null hypothesis that horizontal saccades have no effect on memory recall"

Ex 3 - Turning the hands of time

Topolinski and Sparenberg (2012): clockwise movements induce psychological states of temporal progression and an orientation toward the future and novelty.

Ex 3 - Turning the hands of time



Participants who turned kitchen rolls *clockwise* reported more "openness to experience" than participants who turned the kitchen rolls *counterclockwise*

Ex 3 - Turning the hands of time

Wagenmakers et al. (2015) performed a confirmatory replication of the original experiment. Let's look at their data:

- load `kitchen_rolls.csv` into JASP
- conduct an independent samples t-test
  - DV = `mean_NEO`
  - Grouping variable = `Rotation`
  - be sure to select "Group 1 > Group 2" for hypothesis

Ex 3 - Turning the hands of time

Now lets do the Bayesian version

Ex 3 - Turning the hands of time

Example writeup:

"We computed a Bayesian independent samples t-test to quantify the evidence for the hypothesis that participants who turn kitchen rolls clockwise will report increased openness. We found a Bayes factor of $B_{0+} = 7.8$, indicating that the observed data are approximately 8 times more likely under the null hypothesis than the alternative hypothesis. According to the classification scheme of Jeffreys (1961), this constitutes strong evidence for the null hypothesis that there is no difference in openness between people rotating clockwise versus counterclockwise"

Questions

- Email: `faulkenberry@tarleton.edu`
- Twitter: `@tomfaulkenberry`