Basic introduction to Bayesian statistics

Aleksandra Kaszowska

What are we doing here?

- Take-home summary:
  - why do frequentists and Bayesians disagree?
  - why are Bayesians cool?
  - what is Bayesian inference?
  - what happens when I have two hypotheses?
  - Bayes Factor (and posterior probabilities)
  - how do I set my priors?
- real research example: do valenced odors and trait body odor disgust affect evaluation of emotion in dynamic faces?
- another real research example: wearing a bicycle helmet can increase risk taking and sensation seeking in adults.

Are you frustrated?

- We teach our students that probability is really just expected frequency of events occurring
- And that it is counterintuitive and abstract, and therefore difficult → but if it’s difficult it means it must be true and important!
- And that “classical” statistics are the objective (and therefore the best) way of testing our predictions (hypotheses) about the world

- …and then in 45 minutes we throw at them an entirely new and different way of thinking about probability and research, and expect them to just… understand it and remember it and apply it? While this stuff could easily take a full 10 ECTS course to study?


YEAH..
NO

Your take-home message summary: Bayesian inference edition

Why do frequentists and Bayesians disagree?

- There are two kinds of probability:
  - Aleatory probability: a statement of an expected frequency over many repetitions of a procedure
    - Expresses properties of the long-run behavior of well-defined processes
    - Assumes physical repeatability and independence among repetitions
    - We can make statements about tendencies, but not about single outcomes
  - Epistemic probability: a degree of belief (number between 1 and 0) that quantifies how strongly I think something is true
    - Subjective: depends entirely on the information available to me, so you can have different information and therefore ascribe a different epistemic probability to an event
    - If rational people update their beliefs in a rational way, they will gradually approach agreement – which implies I can improve my thinking if I look for more information
    - We can make statements about both singular and repetitive events

Why do frequentists and Bayesians disagree?

- Frequentists use aleatory probability
  - $p < 0.05$ means that if I repeat the same experiment 100 times, 95 times I will correctly reject H0 and 5 times I will commit Type I error (incorrect rejection of H0, e.g. telling an old man he's pregnant)
    - Null hypothesis is either false or not false, and if I keep repeating my experiment, I will see how many times H0 comes out as false and how many times as not false.
    - The coin has two sides, either heads or tails, and if I keep flipping it, I will see how many times I get heads and how many times I get tails.
- But frequentists ask epistemic questions, because they (mostly) ask about singular events:
  - Is this theory true or not? (will standing in a superhero pose increase your confidence?) → is this *one theory* true?
  - Is this effect positive or negative? (is playing with your cellphone in bed making your sleep worse?) → is this *one effect* true?

Why are Bayesians cool?

- Bayesian use epistemic probability
  - Probability of a proposition is a *degree of belief* in the truth of that proposition given available evidence
    - Proposition: a statement that is either true or false (for example: grass is green)
  - Bayesians use their prior knowledge to estimate the degree of belief in the truth of a hypothesis/proposition, then they test it, and then they adjust their posterior estimate in the truth of proposition
    - For example, a proposition could be: “this event will occur”
    - But can also refer to a statement with epistemic uncertainty

If you are going to remember only ONE slide from this lecture, remember this one!

- Orthodox statistics tell us **P(data|theory)** – how likely is it that the observed value (data) is produced by the hypothetical distribution of H0 (theory)
  - Frequentists rely on long-term probabilities
  - But this tells you nothing about the probability of outcome of *this particula test you just run*
- Bayesian statistics tell us **P(theory|data) -** how likely is it that the distribution of H (theory) is produced by the observed value (data)
  - Key question: what is the factor by which we should change the probability of different theories given the data?
  - For example, Bayes Factor

A: I will teach statistics today
B: I will drink beer tonight

What is Bayesian inference?

- It is the application of the product and sum rules of probability to real problems of inference
  - Product rule simplified: if I have two statistically independent events, A and B, then:
    - Probability of A happening P(A) = P(A|B) – they are independent so knowing whether A happened tells us nothing about the chance of B happening
    - Probability of both happening P(A,B) = P(A)P(B)
  - Sum rule simplified: the probability of event B happening is a sum of all the joint probabilities that involve both event B and elements from set A
    - If A has two options: {A, ¬A} and B has two options: {B, ¬B}, then P(B) = P(A,B) + P(¬A,B)
- Applications of Bayesian inference are creative ways of looking at a problem through the lens of these two rules.

Bayesian probability of a hypothesis being true given the observed data is a ratio of two terms:

probability that the hypothesis true
AND
probability of observing data X given the hypothesis is true

$$P(\mathcal{M}|X) = \frac{P(\mathcal{M})P(X|\mathcal{M})}{P(\mathcal{M})P(X|\mathcal{M}) + P(\neg\mathcal{M})P(X|\neg\mathcal{M})}$$

Sum of all probabilities of observing data X
(data X if hypothesis true plus data X if hypothesis not true)

Research question: does teaching statistics make you drink beer?
Hypothesis M: people drink beer after they teach statistics
Data X: Aleks drank beer tonight

Bayesian Probability that
(people drink beer after teaching
given that Aleks drank beer tonight)

Product rule:
Probability that (people drink beer after teaching stats)
multiplied by
Probability that (Aleks drank beer given that people drink beer after teaching)

$$P(\mathcal{M}|X) = \frac{P(\mathcal{M})P(X|\mathcal{M})}{P(\mathcal{M})P(X|\mathcal{M}) + P(\neg\mathcal{M})P(X|\neg\mathcal{M})}$$

Sum rule:
Sum of probabilities of all reasons Aleks could have drunk beer tonight
(because she was teaching) + (not because she was teaching)

And if I have only two hypotheses…

- After collecting data, I am left with the posterior probability of two hypotheses: P(M|X) and P(¬M|X)
- Ratio of those two hypotheses is a quantification of our relative belief in one hypothesis vis-à-vis another: *posterior odds* P(M|X)/P(¬M|X)
  - If P(M|X) = .75 and P(¬M|X) = .25, the posterior odds are .75/.25 = 3, or 3:1 ("three to one") in favor of M over ¬M
- The Bayes factor can be interpreted as *the extent to which the data sway our relative belief from one hypothesis to the other*, as determined by comparing our hypotheses' abilities to predict the observed data.

$$\underbrace{\frac{P(\mathcal{M}|X)}{P(\neg\mathcal{M}|X)}}_{\text{Posterior odds}} = \underbrace{\frac{P(\mathcal{M})}{P(\neg\mathcal{M})}}_{\text{Prior odds}} \times \underbrace{\frac{P(X|\mathcal{M})}{P(X|\neg\mathcal{M})}}_{\text{Bayes factor}}$$

Hypothesis M: people drink beer after they teach statistics; Data X: Aleks drank beer tonight

Interpreting the Bayes factor

BF_{10}	interpretation
> 100	Extreme evidence for H_1
30 - 100	Very strong evidence for H_1
10 – 30	Strong evidence for H_1
3 – 10	Moderate evidence for H_1
1 – 3	Anecdotal evidence for H_1
1	Equal evidence for H_1 and H_0
1/3 – 1	Anecdotal evidence for H_0
1/10 – 1/3	Moderate evidence for H_0
1/30 – 1/10	Strong evidence for H_0
1/100 – 1/30	Very strong evidence for H_0
< 1/100	Extreme evidence for H_0

- Model comparison
  - In principle, it should be possible to compute Bayes factor for each statistical model
  - So, in principle, by comparing Bayes factors between models, you can decide which model is strongest (= has highest probability of being correct)

$$BF_{10} = H1\_1 / H1\_0$$

$$BF_{01} = H1\_0 / H1\_1$$

From Lee, M.D., & Wagenmakers, E.J. (2014). *Bayesian Cognitive Modeling, a Practical Course*. Cambridge University Press.

Bayes factor vs posterior probabilities

- Bayes factor $\frac{P(X|M)}{P(X|\neg M)}$ is a learning factor that tells us how much evidence the data have delivered in favor of one hypothesis (M) over the other (¬M)
- Posterior probabilities P(M|X) and P(¬M|X) determine our total belief after taking into account the new data

- It is possible that a Bayes factor favors M over ¬M while at the same time posterior probability of ¬M remains greater than M
  - It is possible that we learned about evidence favoring M over ¬M, but the total cumulative belief in M vs ¬M is still acting in favor of ¬M

How do I set/decide my priors?

- My prior is my knowledge about the distribution of a variable from other studies – so my prior is a probability density of a variable from another study
  - For example: combine a sample mean with a 95% interval and distribution shape
  - Credible interval is a Bayesian analog of confidence interval in frequentist statistics
- Ideally, you would do a meta-analysis; realistically, use published confidence intervals from leading studies
- But keep in mind:
  - Your prior must be acceptable to a skeptical audience
  - You can also perform a robustness check: conduct your analysis with different priors, for example pessimistic, realistic, and optimistic prior. Often you will get similar results.

Real research example time!

There are many different approaches to using Bayesian thinking in research (for a comprehensive introduction see, for example, Krushke, J.K. (2015). *Doing Bayesian Analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Amsterdam, The Netherlands: Academic Press); one of them is using the Bayes Factor for hypothesis testing

- We compute the Bayes Factor for two competing hypotheses and use that to decide which one of the two is more likely

Syrjänen, E., Liuzza, M. T., Fischer, H., & Olofsson, J. K. (2017). Do valenced odors and trait body odor disgust affect evaluation of emotion in dynamic faces? *Perception*, *46*(12), 1414-1426.

- Simplified for the sake of presentation

Previous research

(Syrjänen, Liuzza, Fischer & Olofsson, 2017, simplified for presentation)

- Contextual information can inform how we interpret facial expression
- Unpleasant odor contexts enhance recognition accuracy for disgusted faces – *congruency* effect
- Valenced odors (pleasant / unpleasant) result in unspecific enhanced recognition of happy or disgusted expressions – *general emotion* effect
- Methodological problem: previous research has been done on static photographs, while "dynamic faces that change from neutral to emotionally expressive might be regarded as more ecologically valid and more affective" (Syrjänen et al., p. 1414)

Bayes Factor Example: valenced odors hypotheses

(Syrjänen, Liuzza, Fischer & Olofsson, 2017, simplified for presentation)

- H1: “General arousing properties of odors affect recognition speed, which would lead to faster emotion recognition in the presence of an odor (vs. no odor condition), irrespective of odor-face emotional congruency”
- H2: “Emotions that are congruent with the valence of the odor are recognized faster (e.g., disgusted faces are recognized faster in an unpleasant odor context, whereas happy faces are recognized faster in a pleasant odor context)”

MAIN EFFECT (H1)

	HAPPY	DISGUSTED
PLEASANT ODOR	++	+
NO ODOR	0	0
DISGUSTING ODOR	+	++

INTERACTION (H2)

Bayes Factor Example: valenced odors method

(Syrjänen, Liuzza, Fischer & Olofsson, 2017, simplified for presentation)

- Within-subject design: participants performed emotion recognition task, categorizing stimuli as either "disgusted" or "happy" (2AFC: two-alternative forced choice)
- 16 stimuli presented in three blocks (pleasant odor—lilac essence, smells like flower-scented bath soap; unpleasant odor—valeric acid, smells like bad human sweat; neutral odor), each stimulus three times per block
    - Odor manipulation by placing odorized cotton pads in an elastic cotton tube attached under a participant's nose
- IVs: odor condition, stimulus emotion; DV: reaction time
- Stimuli: 16 unique video clips of 4 male and 4 female faces, lasting 3 seconds each, morphing from neutral expression to either happy or disgusted expression (thus (4 + 4) x 2 = 16 clips)
- Sample: 21 adults (eight male, M_{age} = 31.86, SD_{age} = 10.93), sampling from minimal sample size n = 20 until either enough evidence for H0 (no effect) or strong evidence for H1 (effect present), maximum sample n = 70 from NHST a priori power analysis for medium effect size r = .3

Bayes Factor Example: H1

(Syrjänen, Liuzza, Fischer & Olofsson, 2017, simplified for presentation)

H1: does not mean what the odor is. If there is an odor, then people will recognize emotions faster. If there is no odor, then people will recognize emotions slower (compared to presence of odor)

- H1: “General arousing properties of odors affect recognition speed, which would lead to faster emotion recognition in the presence of an odor (vs. no odor condition), irrespective of odor-face emotional congruency”
  - H1_0: there is no effect of odor on recognition speed of emotional faces
  - H1_1: there is an effect of odor on recognition speed of emotional faces

- Result: Bayes Factor BF_{01} = 4.69 in favor of H1_0: odor in general did not affect emotion recognition speed

$$\underbrace{\frac{P(X|\mathcal{M})}{P(X|\neg\mathcal{M})}}_{\text{Bayes factor}}$$

$$\frac{\text{P(data | H1\_0)}}{\text{P(data | H1\_1)}} = 4.69$$

The Bayes factor can be interpreted as *the extent to which the data sway our relative belief from one hypothesis to the other*, as determined by comparing our hypotheses’ abilities to predict the observed data.

Bayes Factor Example: H2

(Syrjänen, Liuzza, Fischer & Olofsson, 2017, simplified for presentation)

- H2: “Emotions that are congruent with the valence of the odor are recognized faster (e.g., disgusted faces are recognized faster in an unpleasant odor context, whereas happy faces are recognized faster in a pleasant odor context)”
  - H2_0: there is no effect of odor valence congruency on recognition speed of emotional faces
  - H2_1: there is an effect of odor valence congruency on recognition speed of emotional faces
- Result: Bayes Factor BF_{01} = 107.55 in favor of H2_0: congruency between odor and facial expression did not decrease the recognition speed

$$\underbrace{\frac{P(X|\mathcal{M})}{P(X|\neg\mathcal{M})}}_{\text{Bayes factor}} \qquad \frac{P(\text{data} \mid \text{H2\_0})}{P(\text{data} \mid \text{H2\_1})} = 107.55$$

The Bayes factor can be interpreted as *the extent to which the data sway our relative belief from one hypothesis to the other*, as determined by comparing our hypotheses’ abilities to predict the observed data.

Valenced odors: what went wrong?

(Syrjänen, Liuzza, Fischer & Olofsson, 2017, simplified for presentation)

- Main conclusion: results from static faces may not generalize to dynamic faces
- Unsurprising result: people react faster to disgusted faces (compared to happy faces)
- Exploratory (post hoc) analysis: weak evidence (Bayes Factor 2.87) in favor of RT advantage in disgusting odor condition against the other two odor contexts

Another real research example

Gamble, T., & Walker, I. (2016). Wearing a bicycle helmet can increase risk taking and sensation seeking in adults. *Psychological Science*, *27*(2), 289-294.

- a “classical” study reanalyzed as a Bayesian study from summary statistics using JASP (https://jasp-stats.org/) – see Ly et al. (2018) for details

Bayes Factor Example 2: Risk-taking

(Gamble & Walker, 2016, simplified for presentation)

- Problem: humans adapt risk-taking behavior on the basis of perception of safety
- Can perception of safety be experimentally induced by an irrelevant manipulation?
- 80 participants (15 male and 24 female in the helmet condition, 19 male and 22 female in the cap condition, $M_{age} = 25.26$, $SD_{age} = 6.59$) wore a helmet or a baseball cap under the pretense of wearing a mobile eye-tracker → two conditions
- DV: Balloon Analogue Risk Task (BART), participants press a button to inflate a virtual balloon, each trial earns fictional currency, at a random point the balloon bursts, the participant can stop a trial and cash in early; risk-taking score is the mean number of pumps made on trials where the balloon did not burst
- Hypothesis: participants will score higher on risk-taking when wearing a helmet

Bayes Factor Example 2: Risk-taking

(data from Gamble & Walker, 2016)

- R Code:
  - ```
    library(BayesFactor)
    setwd("d:/R_Statistical_Exercises/gamble_walker_2016_bicycle")
    d <- read.csv( "GambleWalkerPsychologicalScienceData.csv", sep = "\t" )
    t.test( BART ~ Condition, data = d, var.equal = TRUE)
    ttestBF(formula = BART ~ Condition, data = d)
    ```
- "Wearing a helmet was associated with higher risk-taking scores (M = 40.40, SD = 18.18) than wearing a cap (M = 31.06, SD = 13.29), $t(78)$ = 2.63, p = .01, d = 0.59" (Gamble & Walker, 2016, p. 291)
- Bayes Factor 4.40 in favor of alternative hypothesis
  - r = 0.707 in the output is the scale for Cauchy distribution, a family of random distributions used to specify a diffuse prior that there is some non-zero effect; the same distribution is used in Syrjänen et al. (2017) study
    ```