Bayesian Data Analysis

Martial Foegel

2024-07-06

Laboratoire de Linguistique Formelle

What is Bayesian data analysis ?

Table of contents

Definition

It is a statistical method based on a Bayesian interpretation of probability. In this case probability expresses the belief in an event. The initial degree of belief in an event can be based on knowledge or on previous events. Using Bayes' theorem, this belief can be updated with new data to produce a new, updated belief in that event.

Example: the weather for May

A different approach to statistics and data analysis

The frequentist approach

Based on "Long run frequency":

- Take a sample from the population, where the population parameters considered "fixed";
- Get some centrality and dispersion parameters from that sample;
- Use them to construct Confidence Intervals with the interpretation that we have a certain amount of confidence (usually 95%) that the true centrality parameter (*i.e.* the population centrality parameter) is within that interval;
- Accept or reject a null hypothesis H_0 using "*p*-value", leading to a commonly dichotomized interpretation of the results.[1]

[1] Hespanhol et al. (2019)

The Bayesian approach

Based on the Bayesian interpretation of probability:

- Parameters from the population are viewed as random variables with distributions;
- Combine both prior belief about the population and the data to get some updated distribution of the population parameters;
- The interest is in the updated parameter distribution, and the aim of the approach is often to describe said distribution;
- To that end centrality and dispersion statistic can be used. Most prevalent is the Credible Interval (equal tail or HPD CrI), with the interpretation that given the data, there is a certain amount of probability (usually 95%) that the true population parameter lies within that interval.[2]

[2] Hespanhol et al. (2019)

Bayes Theorem

! Bayes Theorem:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

A quick numerical example

	A	$\bar{A}$	Total
B	5	3	8
$\bar{B}$	2	8	10
Total	7	11	18

$$P(A|B) = \frac{5}{8}$$

$$\frac{P(B|A)P(A)}{P(B)} = \frac{\frac{5}{7} \times \frac{7}{18}}{8/18} = \frac{5}{18} \times \frac{18}{8} = \frac{5}{8}$$

$$P(B) = \sum_{j=1}^{J} P(B|A_j)P(A_j)$$

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$$

$$P(B) = \frac{5}{7} \times \frac{7}{18} + \frac{3}{11} \times \frac{11}{18} = \frac{5}{18} + \frac{3}{18} = \frac{8}{18}$$

The categorical version

Let θ be a categorical parameter with j classes, and y taking L different values or being continuous, then the Bayes Theorem for categorical parameter is:

$$p(\theta_j|y) = \frac{p(y|\theta_j)p(\theta_j)}{\sum_{j=1}^{J} p(y|\theta_j)p(\theta_j)}$$

The continuous version

Let θ be a continuous parameter, and $\mathbf{y} = y_1, \ldots, y_n$ be an i.i.d. sample, then the Bayes theorem for continuous parameters is:

$$p(\theta|\mathbf{y}) = \frac{L(\mathbf{y}|\theta)p(\theta)}{\int L(\mathbf{y}|\theta)p(\theta)d\theta}$$

Terminology :

$$Posterior = \frac{Likelihood \cdot Prior}{Evidence}$$

And all of this can be generalized for multiple parameters $\theta = \theta_1, \ldots, \theta_k$.

Priors

A prior ($p(\theta)$) is the prior probability distribution of a parameter before taking into account new information. The parameter can be a direct parameter of a model (*e.g.* the mean or the proportion of successes) or a latent one. In that sense, priors can have priors.

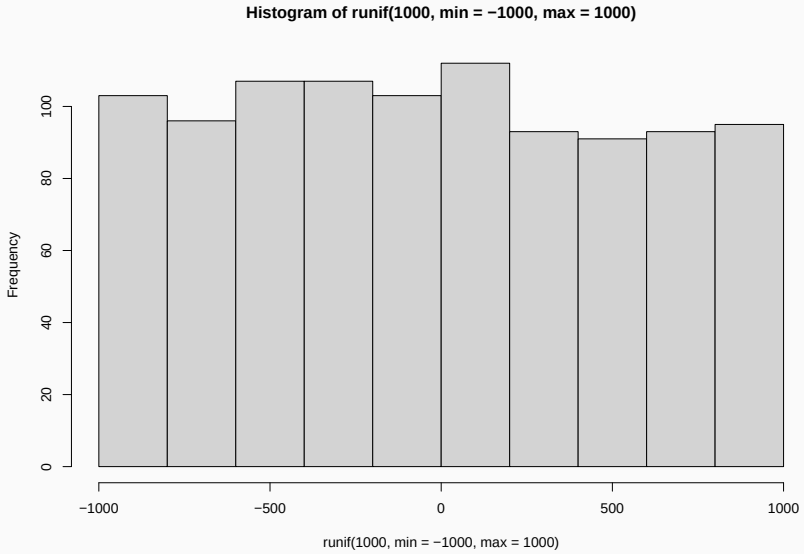
When the prior distribution is determined using historical data, prior knowledge or from experts opinion, those priors are called informative. Priors can be more or less informative, and this informativeness is relative to the question asked.

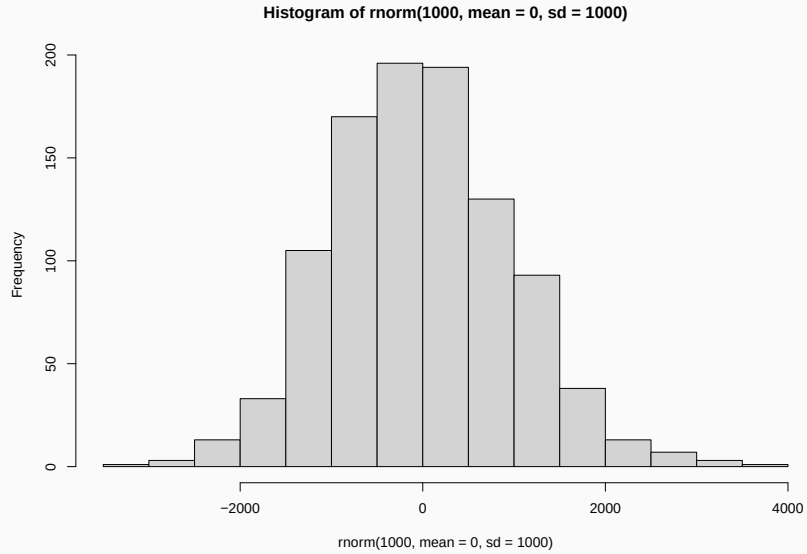
Recommendations about prior choice are available online.[3]

> *Example using Response/Reading Time (mean of 300ms, heavy right tail)*[4]

[3] Vehtari (n.d.)
[4] Lindeløv (2024)

Non-informative prior



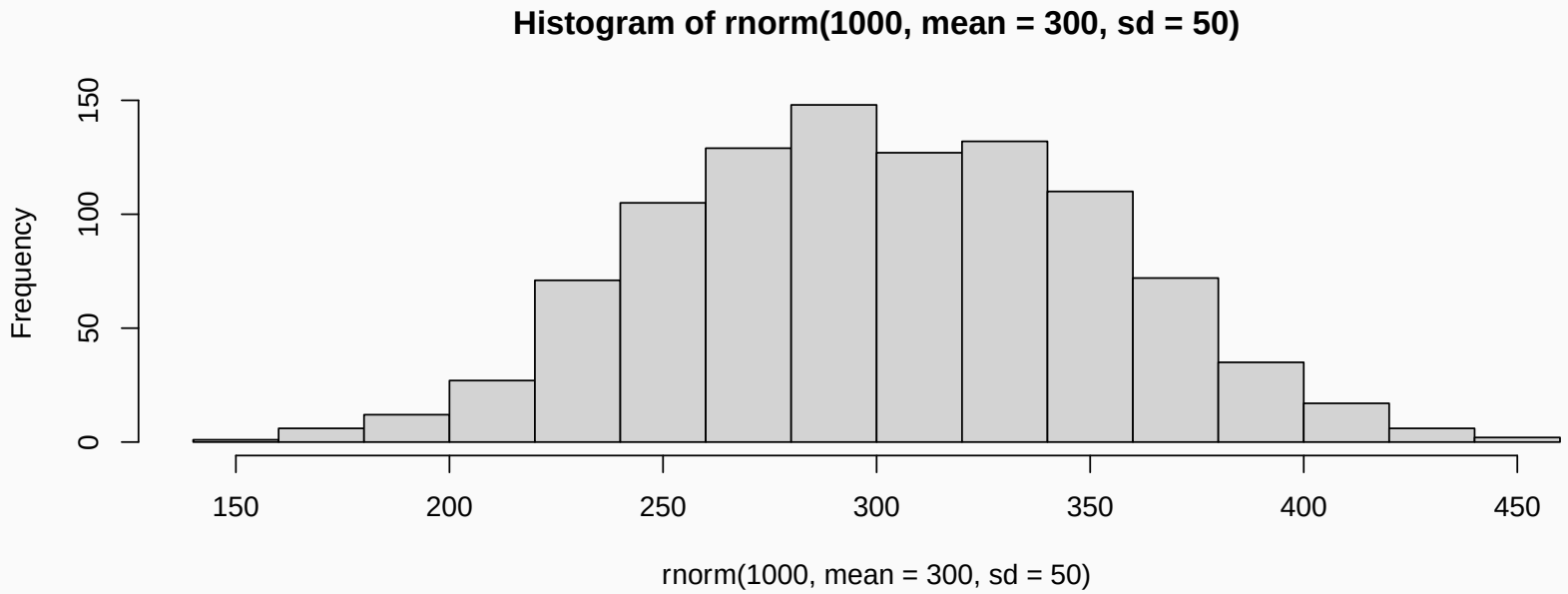


- If the software allows for a default prior, it will usually be a non-informative one;
- A flat prior doesn't necessarily mean a non-informative prior;
- Provide the same results as simpler frequentist methods;
- Increased type 1 and type M error rates.[5]

[5] Lemoine (2019)

Weakly-informative prior

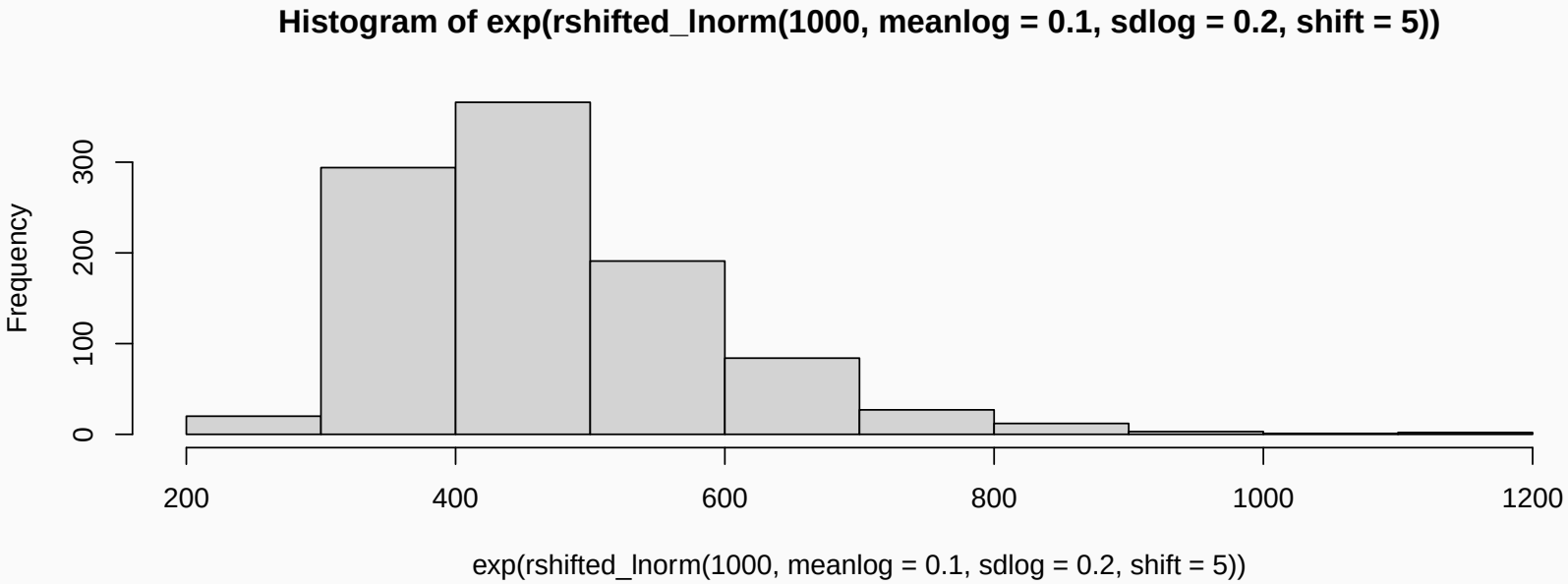
Allow for an analysis that is a compromise between the information available about the parameters, and the actual data. The aim is to align the results with existing knowledge and prevent extreme estimates.



Doesn't have to come from a very elaborate standpoint, just taking into account that the data cannot be negative is already a form of information that will help with the choice of a prior distribution.

Strongly-informative prior

With this type of prior, the prior distribution will most likely overshadow the data acquired.

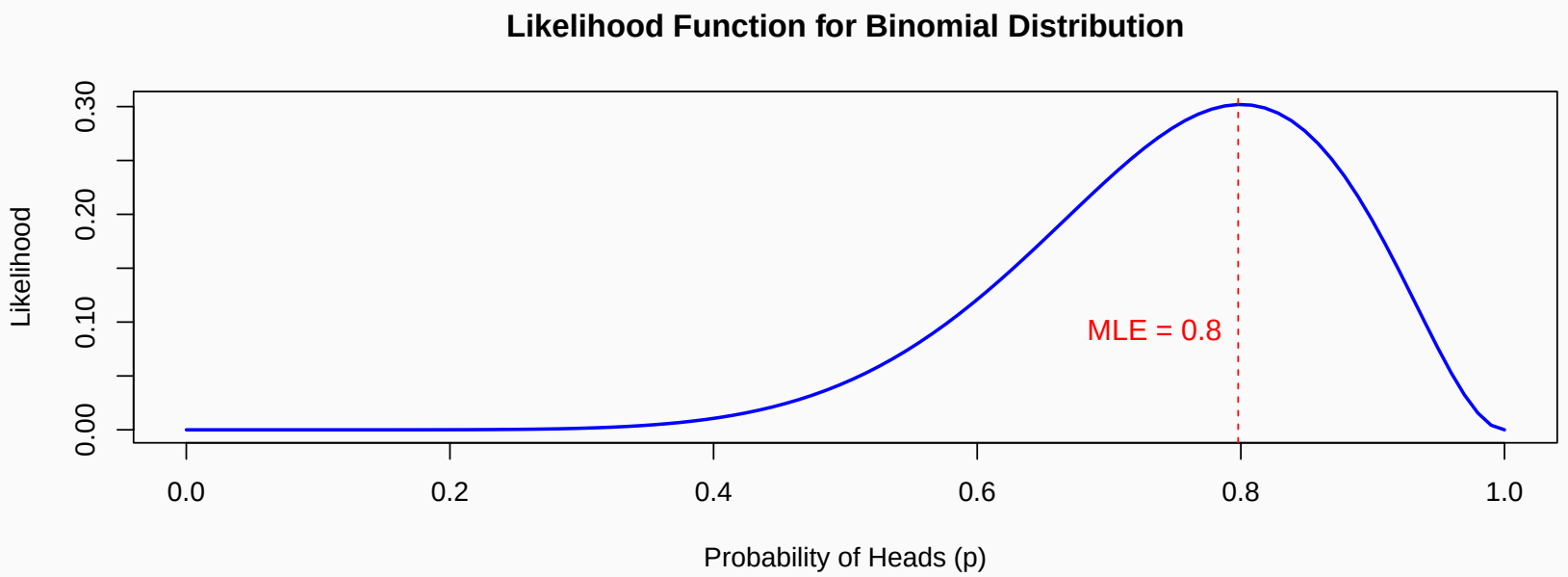


- Not generally recommended;
- Useful if access to data is complicated.

The Likelihood

The likelihood, $L(\theta|\mathbf{y})$, is the probability of observing the gathered data under different parameter values[6].

```
#8 heads out of 10 flips
k <- 8; n <- 10
#likelihood funtion
likelihood <- function(p){choose(n, k)*p^k*(1-p)^(n-k)}
#sequence of all possible probabilities
prob_values <- seq(0, 1, length.out = 100)
# Calculate likelihood for each p value
likelihood_values <- sapply(prob_values, likelihood)
```



[6] Etz (2018)

Intermezzo : what we have seen until now

In the literature you will often see this shorthand to illustrate Bayes' theorem:

$$p(\theta|\mathbf{y}) \propto L(\mathbf{y}|\theta)p(\theta)$$

The posterior $p(\theta|\mathbf{y})$ is proportional to the likelihood, $L(\theta|\mathbf{y})$ (the probability of observing the gathered data under different parameter values) times the prior $p(\theta)$ (the prior probability distribution of a parameter). Sometimes, we can rather easily compute the posterior *up to a proportionality*, but we need to re-scale this to a get probability distribution.

This is where the evidence $p(\mathbf{y})$, also known as the marginal likelihood $\int L(\mathbf{y}|\theta)p(\theta)d\theta$, comes in to play the troublemaker.

The Evidence or Marginal Likelihood

The marginal likelihood, $\int L(\mathbf{y}|\theta)p(\theta)d\theta$, is a likelihood function integrated over the parameter space θ, *i.e.*, the probability of generating the observed data for all possible values of the parameters. It is a normalizing constant to ensure that the posterior is a probability, unless used in the case of model comparison (like the Bayes factor).

However, calculating the marginal likelihood is not a trivial task... . We can know it analytically for some simple distributions, notably in the case of conjugate distributions, otherwise some more complex methods will be needed.

How to get to the posterior...

... Using conjugate distributions

If, given a likelihood function, the prior and the posterior distribution stem from the same probability distribution family, then they are **conjugate distributions** and the prior is a **conjugate prior**.

Examples :

- The Bernoulli likelihood has a Beta conjugate prior with parameters α, β. In this case we know that the posterior parameters will be $\alpha + k$, $\beta + n - k$;
- A categorical and a multinomial likelihood both have a Dirichlet conjugate prior;
- A normal likelihood with unknown mean μ and variance σ^2 has a Normal-inverse-gamma conjugate prior.

... Using conjugate distributions

Historically Bayesian analysis was mostly constrained to conjugate distributions, because for them we have a closed form expression[7] of the posterior. Conjugate distributions are convenient mathematically (and computationally), and for the sake of interpretation. Otherwise numerical integration (computing the integral through an algorithm) was another way calculate the marginal likelihood and get to the posterior.

[7] "Formed with constants, variables and a finite set of basic functions connected by arithmetic operations ($+$, $-$, $\times$, $/$, and integer powers) and function composition" ("Closed-Form Expression" 2024).

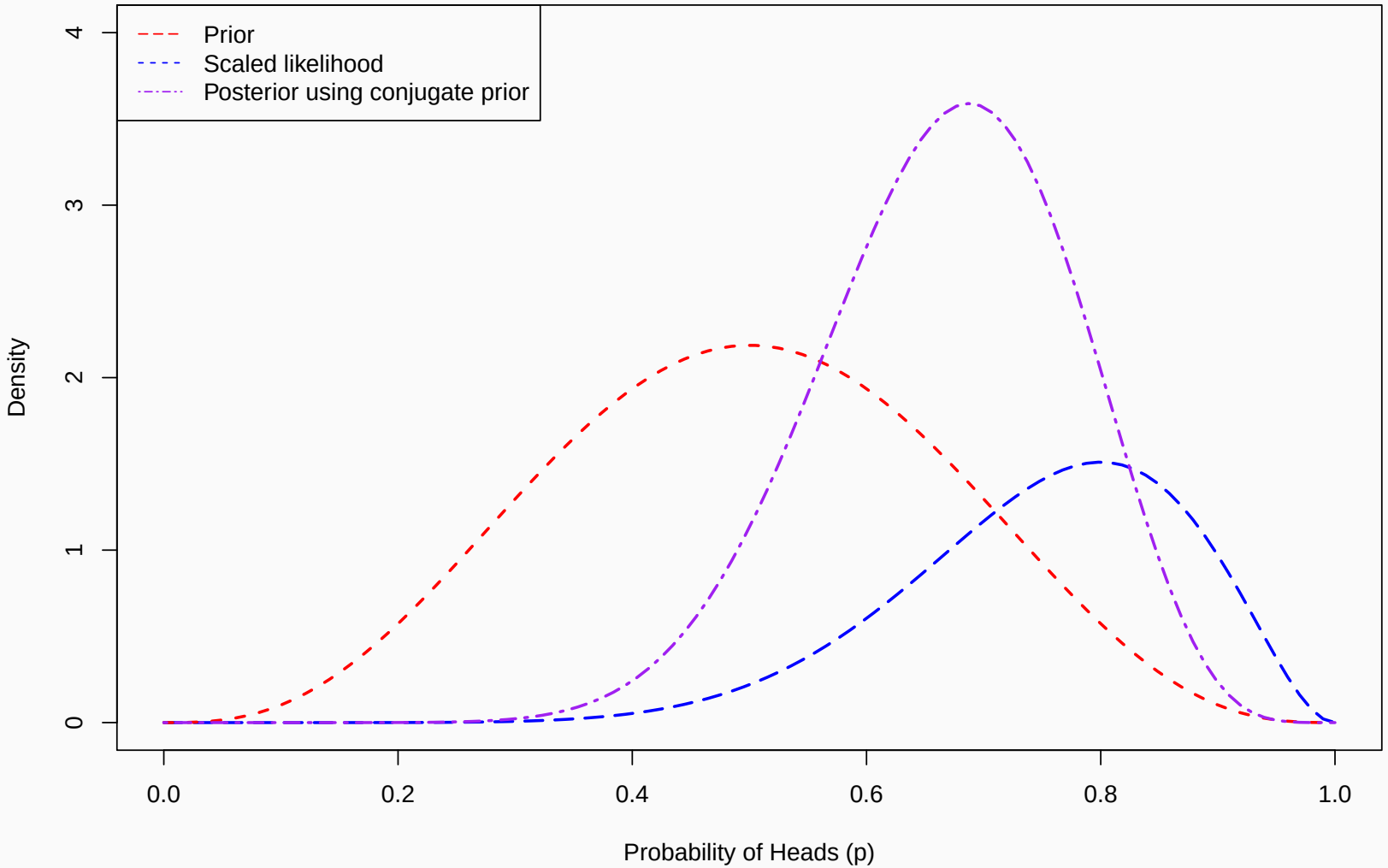
Conjugate distributions in practice i

The initial elements :

```
#8 heads out of 10 coin flips
k <- 8
n <- 10
#likelihood funtion
likelihood <- function(p){
  choose(n, k)*p^k*(1-p)^(n-k)
  }
#sequence of all possible probabilities
prob_values <- seq(0, 1, length.out = 100)
# Calculate likelihood for each p value
likelihood_values <- sapply(prob_values, likelihood)
# prior density Beta (4, 4) since we assume the coin is fair
prior <- function(p){
  dbeta(p, 4, 4)
  }
prior_values <- sapply(prob_values, prior)
posterior <- function(p){
  dbeta(p, 4+k, 4+n-k)
  }
posterior_values <- sapply(prob_values, posterior)
```

Conjugate distributions in practice ii

Using the conjugate prior:



... Using sampling

Markov chain Monte Carlo sampling

Making a first appearance in the 1950s the Markov Chain Monte Carlo (MCMC) methods and have since completely revolutionized Bayesian analysis.[8] Those methods completely sidestep the issues concerning the marginal likelihood while also solving another problem : the *curse of dimensionality*. With high number of parameters $\theta = \theta_1, \ldots, \theta_k$, compared to a low number of data points, the data becomes sparse and reliable results become difficult to obtain.

The way around that for MCMC is to do an "intelligent" search: combine a random search with some intelligent jumping around, without having the results depending on the starting position. Let's break it down a bit...

[8] Robert and Casella (2011)

Monte Carlo

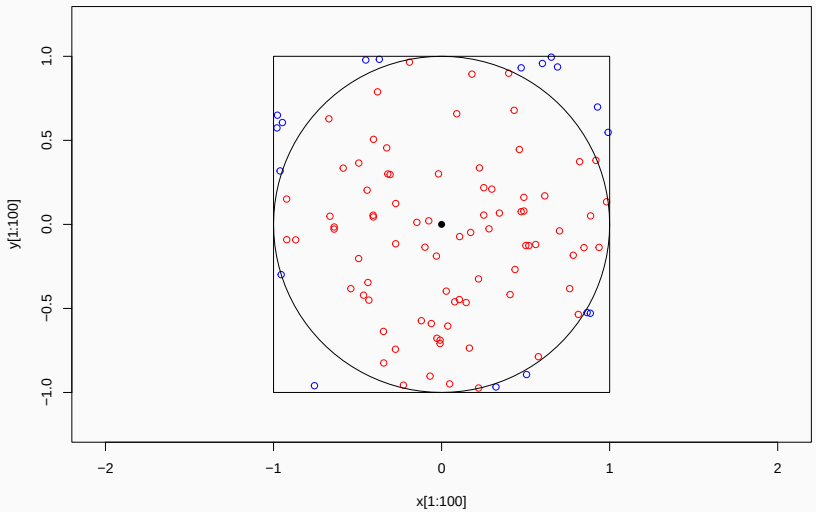
Monte Carlo is a class of algorithm that relies on repeated random sampling to obtain numerical results. Uses random inputs and an accept/reject step before aggregating the data to get a result.

Let us try to evaluate π using a Monte Carlo method.

```
#we take a circle of radius 1 inside a square of size 2 by 2
#first simulate uniform values along x and y axis
N <- 50000
x <- runif(N, -1, 1)
y <- runif(N, -1, 1)

#distance from origin
sum_sq_xy <- sqrt(x^2+y^2)

#those that are less than 1 are within the circle
within_circle <- sum_sq_xy < 1

#multiply by 4 because the proportion of area
#of a circle inside a square is pi/4
4*sum(within_circle)/N
```

```
[1] 3.14136
```

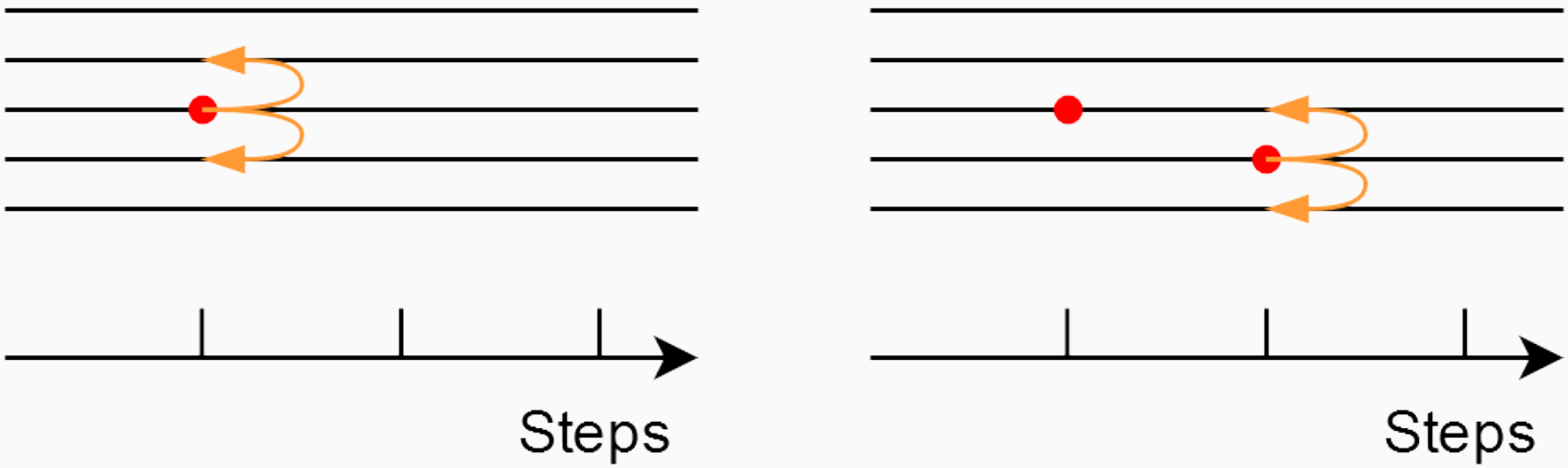
$$\frac{\text{area of a circle}}{\text{area of a square}} = \frac{\pi \times r^2}{s^2} \text{ and here } = \frac{\pi}{4}$$



Markov Chain

Markov Chain is a stochastic model[9] that describe a sequence of event in which the probability of an event depends only on the state of the previous event.

For example, let's take a point that is on a line among other lines. At each step, the point can only jump up or down, one line away, with a probability of 0.5 either way. Those probabilities only depend on where the point is at the current step, so this process is a Markov Chain.



[9] "A stochastic model predicts a set of possible outcomes weighted by their likelihoods, or probabilities" (Taylor and Karlin 2009).

Metropolis algorithm

Metropolis algorithm[10] unfolds as follow :

Let $f(x)$ be function proportional to the desired probability function $P(x)$:

Take an arbitrary number x_0 and choose a symmetrical proposal function $g(x)$ which will be used to propose a new candidate x' depending on the current iteration x_i.

For each iteration i:

1. Propose a candidate x' by using the the distribution $g(x'|x_i)$,
2. Calculate acceptance ratio $\alpha = \frac{f(x')}{f(x_i)} = \frac{P(x')}{P(x_i)}$ (since f is proportional to P),
3. Accept or reject the proposed candidate x':
   - If $\alpha > 1$ (meaning we are moving up the distribution f), $x_{i+1} = x'$,
   - Otherwise, generate a random number form a uniform distribution $u \in [0, 1]$:
     - If $u \leq \alpha$ then $x_{i+1} = x'$,
     - If $u > \alpha$ then $x_{i+1} = x_i$.

[10] Metropolis et al. (1953)

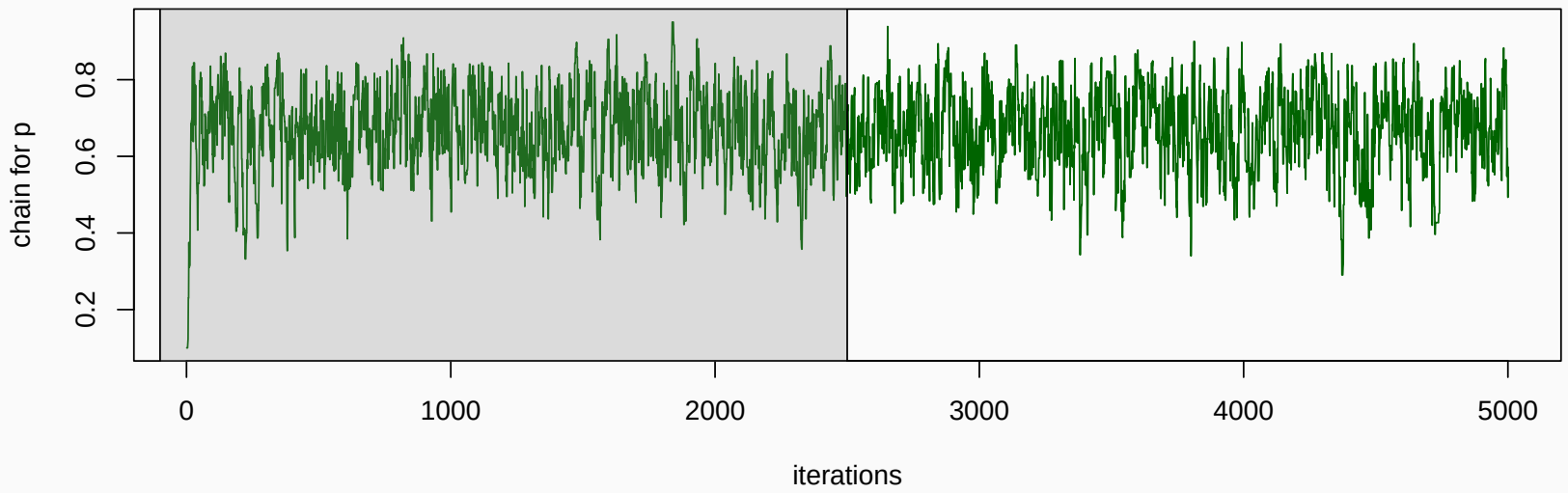
Metropolis algorithm in practice i

The actual algorithm:

```
# the prior times the likelihood is proportional to the posterior
f_x <- function(x){likelihood(x) * prior(x)}
#normal is the proposal function
g_x <- function(x){rnorm(1, mean = x, sd = 0.1)}

metropolis_algo <- function(x_0, nb_iter){
  #one run for a parameter is called a chain
  chain <- vector(length = nb_iter)
  chain[1] <- x_0
  for (i in 1:nb_iter) {
    candidate_x <- g_x(chain[i]) #chain[i] equivalant to x_i
    alpha <- f_x(candidate_x)/f_x(chain[i])
    if(alpha > 1){chain[i+1] = candidate_x}
    else{u <- runif(1)
    if(u <= alpha){chain[i+1] = candidate_x}
    else{chain[i+1] = chain[i]}
    }
  }
  return(chain)
}
```

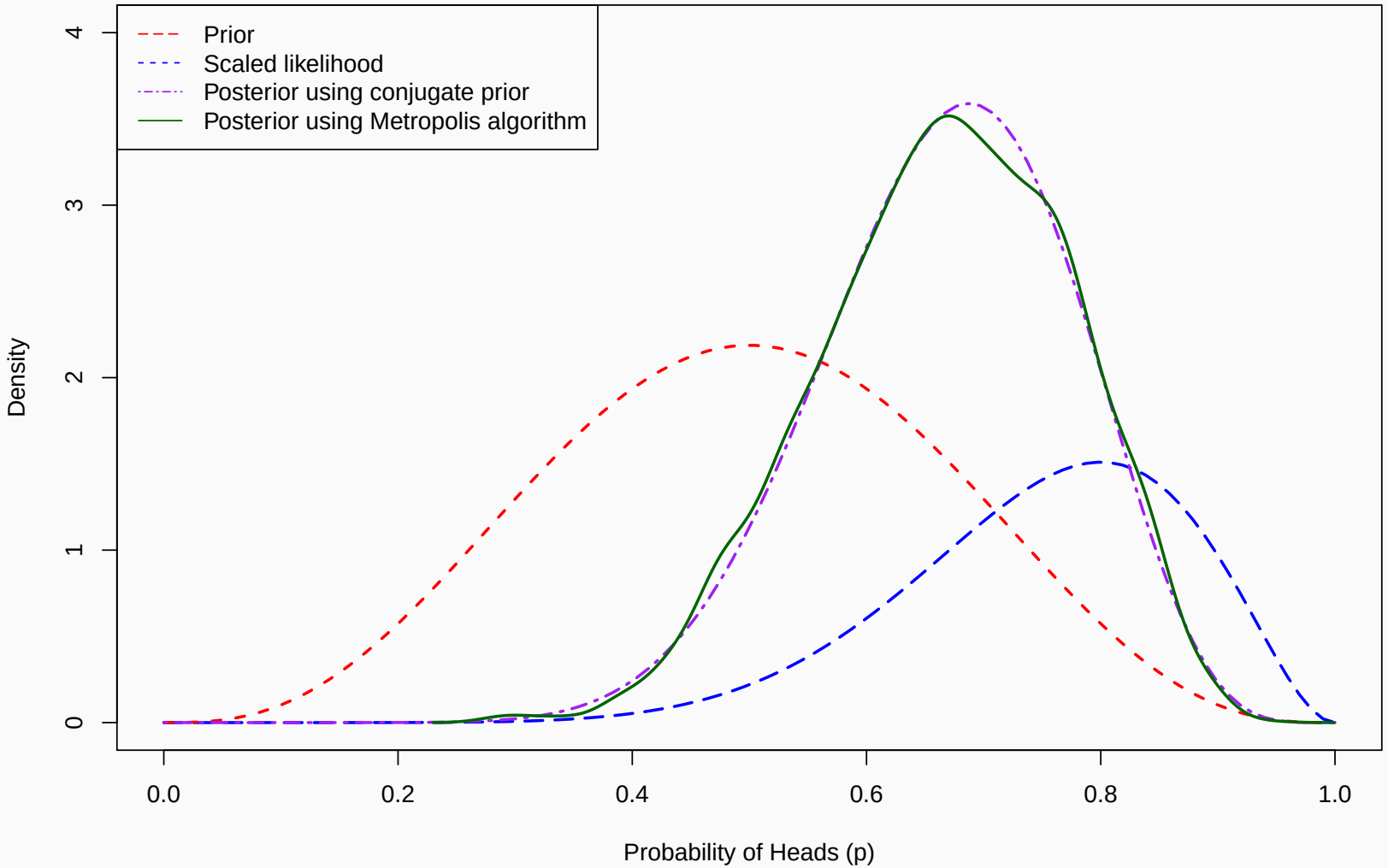
Metropolis algorithm in practice ii

A (Markov) chain is one run of a particular parameter through the algorithm. We usually run multiple chain with different starting values. The first part of the chains are usually discarded, as they are biased by the starting values. This is called the burn-in period.



Metropolis algorithm in practice iii

Adding the MCMC results from the Metropolis algorithm:



Other MCMC algorithms

Metropolis algorithm is simple but suffers serious drawbacks from complicated models. But it is the building block for more advanced algorithm like:

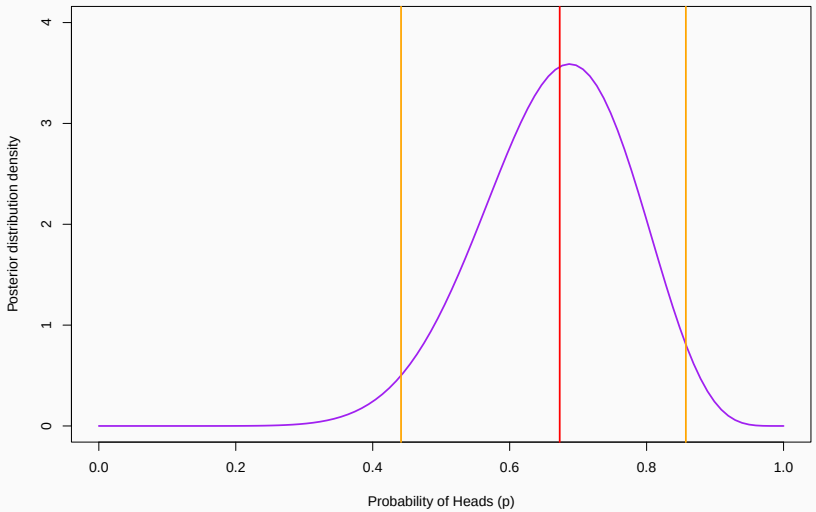
- Metropolis-Hastings algorithm[11] : generalization of the Metropolis algorithm where the proposal function doesn't have to be symmetric;
- Gibbs sampler[12]: useful when the joint distribution is not known explicitly but the conditional distribution of each parameter is known;
- Hamiltonian Monte-Carlo[13]: utilize the local geometry of the target density to move to distant states while maintaining high probability of acceptance.

[11] Hastings (1970)
[12] Geman and Geman (1984)
[13] Duane et al. (1987)

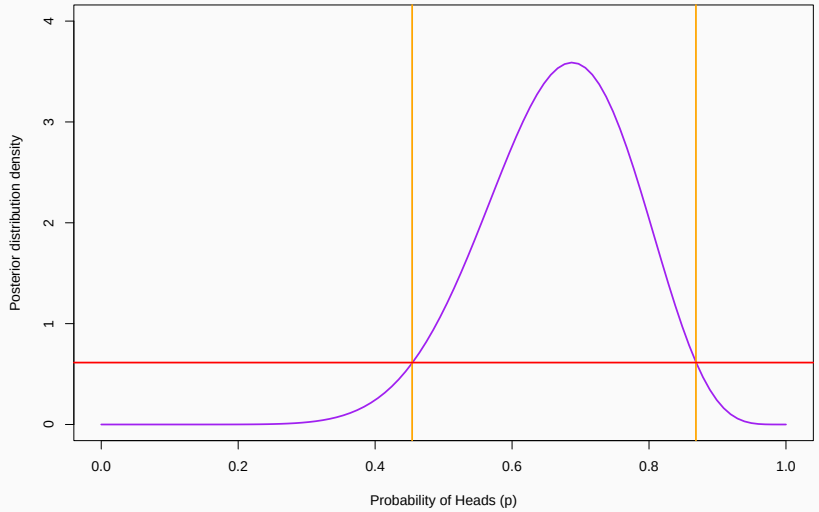
The Posterior

The result of updating the prior probability with the data gathered. We can use this distribution a number of way, the most used one are the credible interval (here we use a 95% CrI):

Equal tail interval, with the probability of being below the interval equal to the probability of being above it:



Highest posterior density interval, which is the narrowest interval:



Conclusion

Advantages and Disadvantages

Advantages:

- Reasoning is more intuitive;
- Incorporate prior information;
- Can use the posterior distribution to directly calculate the probability of different hypothesis;
- Can work with smaller amounts of data.

Disadvantages:

- For a Bayesian analysis to be reproducible, you will need to mention all elements of a Bayesian analysis;
- Subjectivity linked to the choice of priors;
- With small amounts of data, the choice of prior will have a big influence;
- It can be difficult to build a complex Bayesian model and come up with the appropriate priors.

Where to get started ?

Summer school in Potsdam, Germany ("The Eighth Summer School on Statistical Methods for Linguistics and Psychology," n.d.), targeted toward cognitive science. You can find some material from the previous years here linked on the website. The same people organizing the summer school also have an online book (Vasishth, n.d.).

Statistical rethinking (McElreath, n.d.b) is a good statistics book for applied researchers with freely available lectures on YouTube (McElreath, n.d.a). To check with the help of the bookdown from Kurz (n.d.), in order to have the models refit with `brms` (Bürkner 2021), `ggplot2` (Wickham 2016) and `tidyverse` (Wickham et al. 2019).

In R

Both of the packages bellow use **Stan** a probabilistic programming language written in C++ specifically made for the Bayesian statistical modelling:

- Package `RStan` (Carpenter et al. 2017), a R interface to Stan (need to have a file with R code and one with Stan code). A tutorial for psychologist and linguist is available here : Sorensen and Vasishth (2015).
- Package `brms` (Bürkner 2021) (no need for a Stan file) with a writing format closer to `lme4` (Bates et al. 2015) mixed effects models format.

Some Guidelines

Primer on Bayesian statistics Schoot et al. (2021) also has an updated version of the WAMBS (when to Worry and how to Avoid the Misuse of Bayesian Statistics) checklist Depaoli and Schoot (2017).

Bayesian analysis reporting guidelines (BARG) Kruschke (2021).

References

References i

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. 2015. "Fitting Linear Mixed-Effects Models Using **Lme4**." *Journal of Statistical Software* 67 (1). https://doi.org/10.18637/jss.v067.i01.

Bürkner, Paul-Christian. 2021. "Bayesian Item Response Modeling in R with Brms and Stan." *Journal of Statistical Software* 100 (November): 1–54. https://doi.org/10.18637/jss.v100.i05.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software* 76 (January): 1–32. https://doi.org/10.18637/jss.v076.i01.

References ii

"Closed-Form Expression." 2024. https://en.wikipedia.org/w/index.php?title=Closed-form_expression&oldid=1218849771.

Depaoli, Sarah, and Rens van de Schoot. 2017. "Improving Transparency and Replication in Bayesian Statistics: The WAMBS-Checklist." *Psychological Methods* 22 (2): 240–61. https://doi.org/10.1037/met0000065.

Duane, Simon, A. D. Kennedy, Brian J. Pendleton, and Duncan Roweth. 1987. "Hybrid Monte Carlo." *Physics Letters B* 195 (2): 216–22. https://doi.org/10.1016/0370-2693(87)91197-X.

Etz, Alexander. 2018. "Introduction to the Concept of Likelihood and Its Applications." *Advances in Methods and Practices in Psychological Science* 1 (1): 60–69. https://doi.org/10.1177/2515245917744314.

References iii

Geman, Stuart, and Donald Geman. 1984. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images." *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6 (6): 721–41. https://doi.org/10.1109/TPAMI.1984.4767596.

Hastings, W. K. 1970. "Monte Carlo Sampling Methods Using Markov Chains and Their Applications." *Biometrika* 57 (1): 97–109. https://doi.org/10.2307/2334940.

Hespanhol, Luiz, Caio Sain Vallio, Lucíola Menezes Costa, and Bruno T Saragiotto. 2019. "Understanding and Interpreting Confidence and Credible Intervals Around Effect Estimates." *Brazilian Journal of Physical Therapy* 23 (4): 290–301. https://doi.org/10.1016/j.bjpt.2018.12.006.

References iv

Kruschke, John K. 2021. "Bayesian Analysis Reporting Guidelines." *Nature Human Behaviour* 5 (10): 1282–91. https://doi.org/10.1038/s41562-021-01177-7.

Kurz, A. Solomon. n.d. *Statistical Rethinking with Brms, Ggplot2, and the Tidyverse*. https://bookdown.org/ajkurz/Statistical_Rethinking_recoded/.

Lemoine, Nathan P. 2019. "Moving Beyond Noninformative Priors: Why and How to Choose Weakly Informative Priors in Bayesian Analyses." *Oikos* 128 (7): 912–28. https://doi.org/10.1111/oik.05985.

Lindeløv, Jonas Kristoffer. 2024. *Lindeloev/Shiny-Rt*. https://github.com/lindeloev/shiny-rt.

References v

McElreath, Richard. n.d.a. "Richard McElreath - YouTube." https://www.youtube.com/channel/UCNJK6_DZvcMqNSzQdEkzvzA/playlists.

———. n.d.b. "Statistical Rethinking: A Bayesian Course with Examples in R and STAN." https://www.routledge.com/Statistical-Rethinking-A-Bayesian-Course-with-Examples-in-R-and-STAN/McElreath/p/book/9780367139919.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. 1953. "Equation of State Calculations by Fast Computing Machines." *The Journal of Chemical Physics* 21 (6): 1087–92. https://doi.org/10.1063/1.1699114.

Robert, Christian, and George Casella. 2011. "A Short History of Markov Chain Monte Carlo: Subjective Recollections from Incomplete Data." *Statistical Science* 26 (1). https://doi.org/10.1214/10-STS351.

References vi

Schoot, Rens van de, Sarah Depaoli, Ruth King, Bianca Kramer, Kaspar Märtens, Mahlet G. Tadesse, Marina Vannucci, et al. 2021. "Bayesian Statistics and Modelling." *Nature Reviews Methods Primers* 1 (1): 1–26. https://doi.org/10.1038/s43586-020-00001-2.

Sorensen, Tanner, and S. Vasishth. 2015. "Bayesian Linear Mixed Models Using Stan: A Tutorial for Psychologists, Linguists, and Cognitive Scientists." *arXiv: Methodology*. https://doi.org/10.20982/tqmp.12.3.p175.

Taylor, Howard M., and Samuel Karlin. 2009. *An Introduction to Stochastic Modeling*. 3. ed., [repr.]. San Diego: Academic Press.

"The Eighth Summer School on Statistical Methods for Linguistics and Psychology." n.d. https://vasishth.github.io/smlp2024/.

References vii

Vasishth, Bruno Nicenboim, Daniel Schad, and Shravan. n.d. *An Introduction to Bayesian Data Analysis for Cognitive Science*. https://vasishth.github.io/bayescogsci/book/.

Vehtari, Aki. n.d. "Prior Choice Recommendations." https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the Tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.