SOC542 Statistical Methods in Sociology II

Probability and Bayesian Inference

Thomas Davidson

Rutgers University

February 7, 2022

Plan

- Probability review
- Bayes' theorem and its applications
- Comparing Bayesian and Frequentist approaches
- Bayesian estimation
- Lab: Bayesian regression in R

Probability review

Simple probability

- $P(A)$ refers to the probability of an event A
  - e.g. $P(A) = 0.5$ when referring to the probability of receiving a heads on a fair coin toss.
  - e.g. $P(B) = \frac{1}{6}$ is the probability of rolling six with a fair die.
- In each case, we have a *random process* with a set of possible outcomes (e.g. heads or tails) referred to as the *sample space*.

Probability review

Simple probability

- What is the probability of tossing a coin twice and getting two heads?

Probability review

Simple probability

- What is the probability of tossing a coin twice and getting two heads?
  - $P(A)P(A) = P(A) * P(A) = 0.5 * 0.5 = 0.25$

Probability review

Simple probability

- What is the probability of tossing a coin twice and getting two heads?
  - $P(A)P(A) = P(A) * P(A) = 0.5 * 0.5 = 0.25$
- What is the probability of a sequence of N heads?

Probability review

Simple probability

- What is the probability of tossing a coin twice and getting two heads?
  - $P(A)P(A) = P(A) * P(A) = 0.5 * 0.5 = 0.25$
- What is the probability of a sequence of N heads?
  - $P(A)^N$

Probability review

Simple probability

- What is the probability of tossing a coin twice and getting two heads?
  - $P(A)P(A) = P(A) * P(A) = 0.5 * 0.5 = 0.25$
- What is the probability of a sequence of N heads?
  - $P(A)^N$
- In this case, $P(A)$ becomes vanishingly small as $n \to \infty$
  - $0.5^{10} = 0.00098 = \frac{1}{1024}$

Probability review

Simple probability

- We can easily use simulations to verify our calculation. In this case, I use the `rbinom` function to simulate 1024 sequences of 10 tosses of a fair coin.

```
sims <- rbinom(1024, 10, 0.5)
print(length(sims[sims >= 10]))
```

```
## [1] 0
```

Probability review

Independence

- Assume we roll a single die and flip a single coin. What is the probability of rolling a six and getting a tails?

Probability review

Independence

- Assume we roll a single die and flip a single coin. What is the probability of rolling a six and getting a tails?

$$P(A, B) = P(A)P(B) = \frac{1}{2} * \frac{1}{6} = \frac{1}{12}$$

Probability review

Independence

- Assume we roll a single die and flip a single coin. What is the probability of rolling a six and getting a tails?

$$P(A, B) = P(A)P(B) = \frac{1}{2} * \frac{1}{6} = \frac{1}{12}$$

- The two events are independent of one another, so the *joint probability* is simply the product of the probabilities of the two events.

Probability review

Conditional probability and independence

- $P(A)$ and $P(B)$ are independent *if and only if* $P(A|B) = P(A)$.
  - e.g. The number we rolled on the die has no effect on the outcome of the coin toss.

Probability review

Conditional probability and independence

- Consider a deck of 52 standard playing cards. What is the probability of randomly drawing an Ace?[1]

[1] This example is taken from Cunningham 2021, p. 17. It is an example of sampling without replacement.

Probability review

Conditional probability and independence

- Consider a deck of 52 standard playing cards. What is the probability of randomly drawing an Ace?

$$P(Ace) = 4/52 = 1/13$$

- Let's assume we pick an Ace and put it to the side. What's the probability we get another Ace?

Probability review

Conditional probability and independence

- Consider a deck of 52 standard playing cards. What is the probability of randomly drawing an Ace?

$$P(Ace) = \frac{4}{52} = \frac{1}{13}$$

- Let's assume we pick an Ace and put it to the side. What's the probability we get another Ace?
- Wrong answer: $P(Ace_2) = \frac{4}{52} = \frac{1}{13}$.

Probability review

Conditional probability and independence

- Consider a deck of 52 standard playing cards. What is the probability of randomly drawing an Ace?

$$P(Ace) = \frac{4}{52} = \frac{1}{13}$$

- Let's assume we pick an Ace and put it to the side. What's the probability we get another Ace?
- Wrong answer: $P(Ace_2) = \frac{4}{52} = \frac{1}{13}$.
- Correct answer: $P(Ace_2) = P(Ace_2|Ace_1) = 3/51 = 0.059$.
- This is an example of *conditional probability* since $P(Ace_2|Ace_1) \neq P(Ace_1)$.

Probability review

Conditional probability and independence

- We can express a conditional probability as:

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

- The probability of A conditional on B is the **joint probability** of A and B, divided by the **marginal probability** of B.
- The denominator the sum of over possible joint probabilities of B and A, $\sum_{A^*} P(B, A^*)$.
  - The $*$ denotes that A^* may take multiple values.

Probability review

Conditional probability and independence

- If two events are independent, then $P(A|B) = P(A)$.
- To reject independence, we need to show that $P(A, B) \neq P(A)P(B)$

Probability review

Bayes' theorem

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

Probability review

Bayes' theorem

- What's the probability it is going to rain given that we can see clouds?

$$P(Rain|Cloud) = \frac{P(Cloud|Rain)P(Rain)}{P(Cloud)}$$

Probability review

Bayes' theorem

- Let's say we live in England...
  - $P(Cloud) = 0.7$
  - $P(Rain) = 0.3$
  - $P(Cloud|Rain) = 1$

$$P(Rain|Cloud) = \frac{P(Cloud|Rain)P(Rain)}{P(Cloud)} = \frac{1 * 0.3}{0.7} = \frac{0.3}{0.7} \approx 0.429$$

Probability review

Deriving Bayes' theorem

- Start with the definition of conditional probability:

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

- Multiply each side by $P(B)$:

$$P(A|B)P(B) = P(B,A)$$

- Analogously, if we start with $P(B|A)$ we can get:

$$P(B|A)P(A) = P(B,A)$$

Probability review

Deriving Bayes' theorem

- The previous example shows that the following quantities are equal:

$$P(A|B)P(B) = P(B|A)P(A)$$

- Divide both sides by $P(B)$ to get Bayes' theorem:

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

Bayes' theorem

COVID-19 tests

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

$$P(C19|+) = \frac{P(+|C19)P(C19)}{P(+)}$$

Bayes' theorem

COVID-19 tests

$$P(C19|+) = \frac{P(+|C19)P(C19)}{P(+)}$$

- $P(C19|+)$: Probability you have COVID-19 given that you test positive.
- $P(+|C19)$: Probability you test positive given that you have COVID-19.
- $P(C19)$: Probability you have COVID-19 given population infection rates.
- $P(+)$: Probability a test returns a positive result.

Bayes' theorem

COVID-19 tests

- Assume there is a 1% chance you have COVID-19.
- Assume a test has a false negative rate of 2%.
  - 98% of the time it correctly diagnoses COVID-19, 2% of the time it fails to detect it.
- Assume the same test has a false positive rate of 5%
  - 95% of the time it correctly rejects COVID-19 when a person is negative, 5% of the time it falsely diagnoses COVID-19.
- What is the probability you really have COVID-19 following a positive test?

Bayes' theorem

COVID-19 tests: P(+|C19)

$$P(C19|+) = \frac{P(+|C19)P(C19)}{P(+)}$$

- If we assume a false negative rate of 2%. Then the probability of a positive test given COVID-19 is $P(+|C19) = 1 - 0.02 = 0.98$.

Bayes' theorem

COVID-19 tests: P(C19)

$$P(C19|+) = \frac{P(+|C19)P(C19)}{P(+)}$$

- Assume 1% of the population has COVID-19, then $P(C19) = 0.01$.

Bayes' theorem

COVID-19 tests: P(+)

- To calculate the proportion of positive tests we need to count all the positive tests.
- We can thus reformulate Bayes rule as

$$\frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A*)P(A*)}$$

$$\frac{P(+|C19)P(C19)}{P(+|C19)P(C19) + P(+|C19-)P(C19-)}$$

Bayes' theorem

COVID-19 tests: P(+)

$$P(C19|+) = \frac{P(+|C19)P(C19)}{P(+)}$$

- We already know the first part of the denominator, $P(+|C19)P(C19) = 0.98 * 0.01$.
- If the test has a false positive rate of 5%, $P(+|C19-) = 0.05 * (1 - 0.01)$
- Thus, we take the sum of these probabilities to get the marginal probability of a positive test: $P(+) = (0.98 * 0.01) + (0.05 * (1 - 0.01))$

Bayes' theorem

COVID-19 tests: Calculating P(C19|+)

- If we plug the numbers into Bayes' theorem we get

$$P(C19|+) = \frac{0.98 * 0.01}{0.98 * 0.01 + 0.05 * 0.99}$$

- We can use R to do the calculation for us

```
(0.98*0.01) / ((0.98*0.01) + (0.05*(1-0.01)))
```

```
## [1] 0.1652614
```

Bayes' theorem

Terminology

Posterior $\propto$ **Likelihood** x **Prior**

- In the previous example,
  - $P(C19|+)$ is the **posterior**.
  - $P(+|C19)$ is the **likelihood of the data**.
  - $P(C19)$ is the **prior**.
- The denominator $P(+)$ is ensures the result is a probability. It is often described as the **evidence** or the **marginal likelihood**.

Bayes' theorem

COVID-19 tests: Tabular explanation

- The four cells in the middle of the table represent the *joint probabilities* of two events.
- The row and column totals represent the *marginal probabilities* of each event.
  - θ is used to denote the parameters we are estimating.

\begin{table}

Test result	$\theta = C19+$	$\theta = C19-$	Marginal (Test
$+$	P(+|C19)P(C19)	P(+|C19-)P(C19-)	$\sum_\theta P(+\|\theta)P(\theta$
-	P(-|C19)P(C19)	P(-|C19-)P(C19-)	$\sum_\theta P(-\|\theta)P(\theta$
Marginal C19	P(C19+)	P(C19-)	1.0

\end{table}

Bayes' theorem

COVID-19 tests: Tabular explanation

- To calculate $P(C19|+)$ we can take the *joint probability* of C19 and a positive test and divide it by the *marginal probability* of a positive test.
  - We can get the relevant values directly from the table: $0.98 * 0.01/0.06$. \begin{table}

Test result	$\theta = C19+$	$\theta = C19-$	Marginal (Test)
$+$	0.98*0.01	0.05*(1-0.01)	0.06
-	(1-0.98)*0.01	(1-0.05)*(1-0.01)	0.94
Marginal C19	0.01	(1-0.01)	1.0

\end{table}

Bayes' theorem

Changing our priors

- Let's change our prior to assume 10% COVID-19 prevalence in the population (perhaps this is a more reasonable assumption at the moment...)

```
(0.98*0.1) / ( (0.98*0.1) + (0.05*0.9) )
```

```
## [1] 0.6853147
```

- Now we get a much higher posterior probability.
- We could easily alter the calculation by incorporating other prior information, e.g. symptoms, exposure

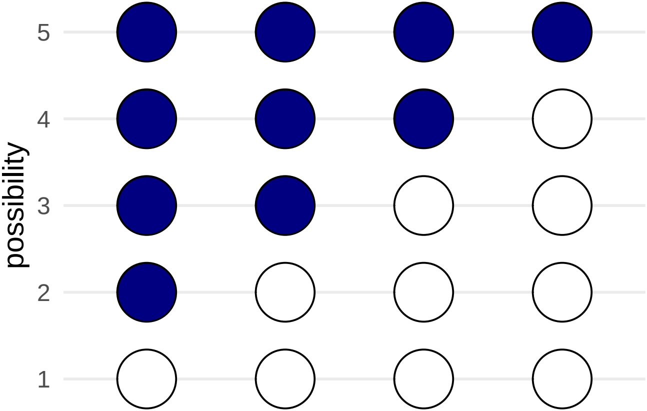
Bayesian inference as counting

McElreath's marble counting example

- Consider a bag containing four marbles
- The marbles can be white or blue
- We draw a sample of marbles from the bag (with replacement)

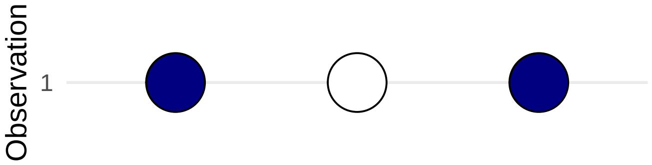
Bayesian inference as counting

Conjecture: Five possibilities



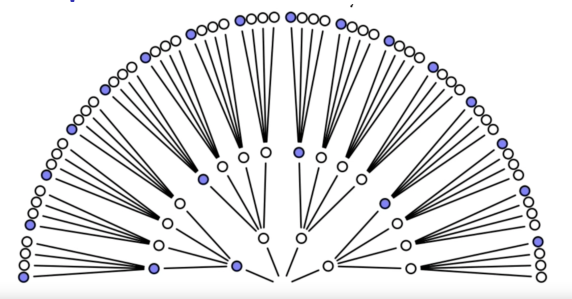
Bayesian inference as counting

A sample from the bag produces



Bayesian inference as counting

Sampling and possibilities



McElreath 2020, Fig. 2.2 (p. 22)

Bayesian inference as counting

Counting the possibilities

Conjecture	Ways to produce [B,W,B]
[W,W,W,W]	0 x 4 x 0 = 0
[B,W,W,W]	1 x 3 x 1 = 3
[B,B,W,W]	2 x 2 x 2 = 8
[B,B,B,W]	3 x 1 x 3 = 9
[B,B,B,B]	4 x 0 x 4 = 0

Bayesian inference as counting

From counts to probability

Conjecture	Propoportion B	Ways [B,W,B]	Plausibility
[W,W,W,W]	0.00	0	0.00
[B,W,W,W]	0.25	3	0.15
[B,B,W,W]	0.50	8	0.40
[B,B,B,W]	0.75	9	0.45
[B,B,B,B]	1.00	0	0.00

Bayesian inference as counting

Summary

- We enumerated the set of plausible data generating processes p
- We counted the ways we could produce the data given each value of p. This is known as the *likelihood*.
- We normalized these counts to get *posterior* probabilities, which indicate the relative plausibility of each option p.
- The most plausible value is the one that has the most ways of generating the data.

Bayesian inference as counting

Incorporating prior information

- Now let's say we pick another marble and it's blue. We can use the prior information to update our counts.

Conjecture	Ways to produce [B]	Prior counts	New counts
[W,W,W,W]	0	0	$0 \times 0 = 0$
[B,W,W,W]	1	3	$3 \times 1 = 3$
[B,B,W,W]	2	8	$8 \times 2 = 16$
[B,B,B,W]	3	9	$9 \times 3 = 27$
[B,B,B,B]	4	0	$0 \times 4 = 0$

Bayesian inference as counting

Bayes' theorem and data analysis

- In a general sense, we can think about Bayesian inference as calculating the posterior distribution in the following way:

$$Posterior = \frac{Probability\ of\ the\ data\ * Prior}{Average\ probability\ of\ the\ data}$$

Bayesian inference

"Bayesian inference is reallocation of credibility across possibilities" - John Kruscke[2]

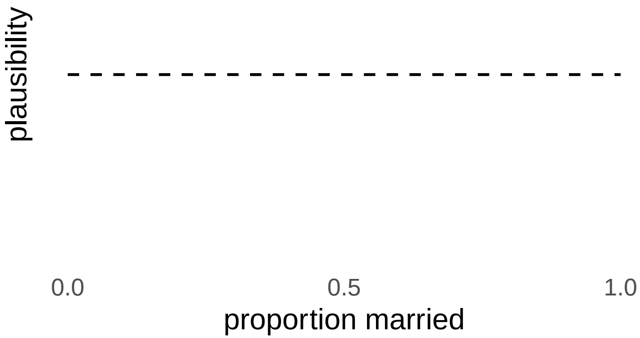
[2] Chapter 2 of Kruschke's 2015 book *Doing Bayesian Data Analysis* provides an outline of his argument and is available online.

Bayesian inference for a continuous parameter

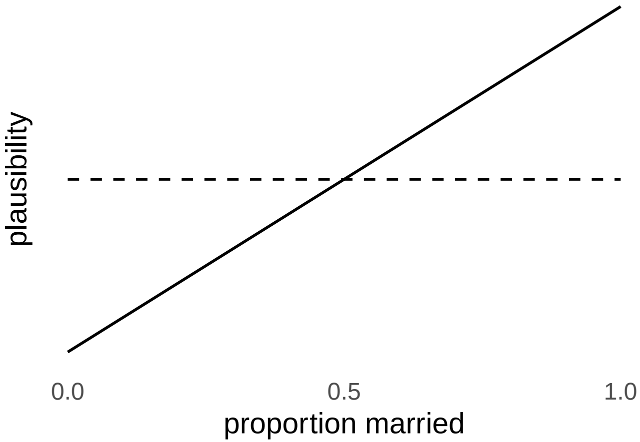
Estimating the marriage rate

- Assume a demographer is interested in estimating the marriage rate in the population.
- The demographer starts out with a "flat" prior
  - The marriage rate could be anywhere from 0 (nobody is married) to 1 (everybody is married).
- The demographer samples people at random and asks them their marital status.

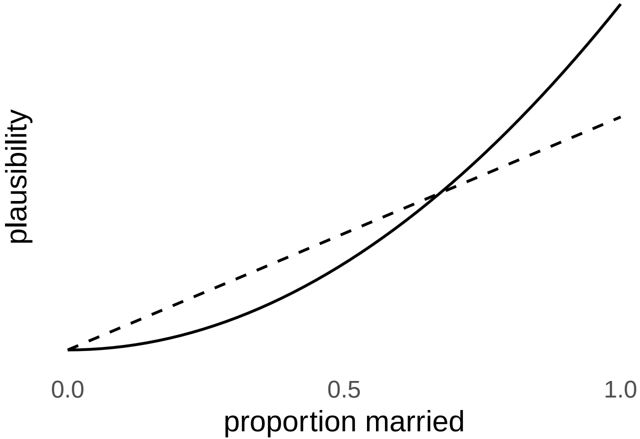
Assume zero knowledge with a flat (uniform) prior



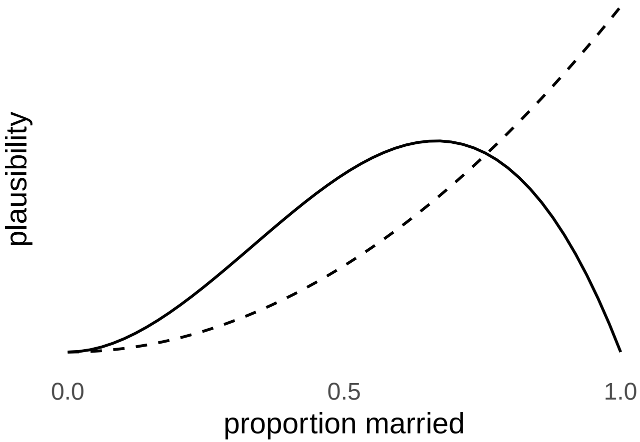
First observation: Married



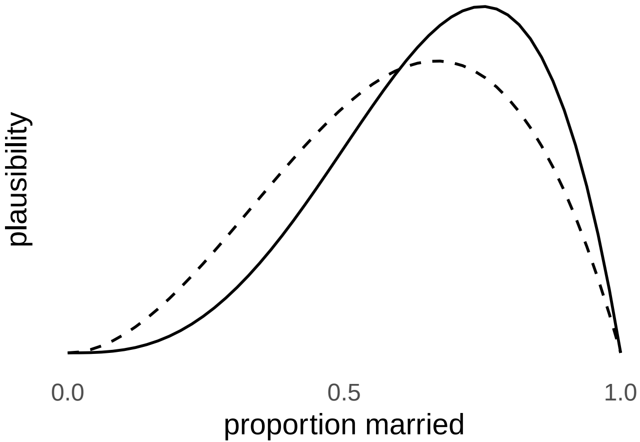
Second observation: Married



Third observation: Single



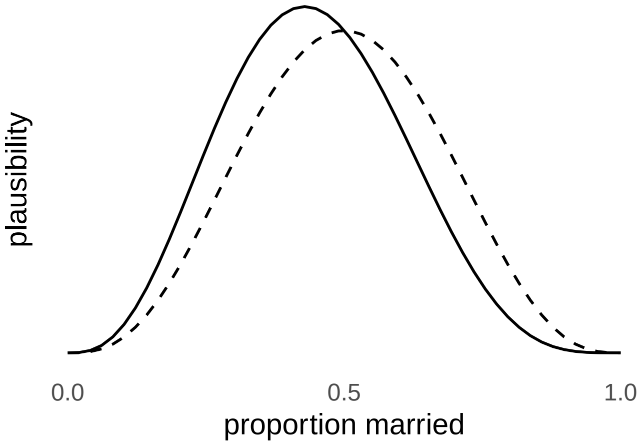
Fourth observation: Married



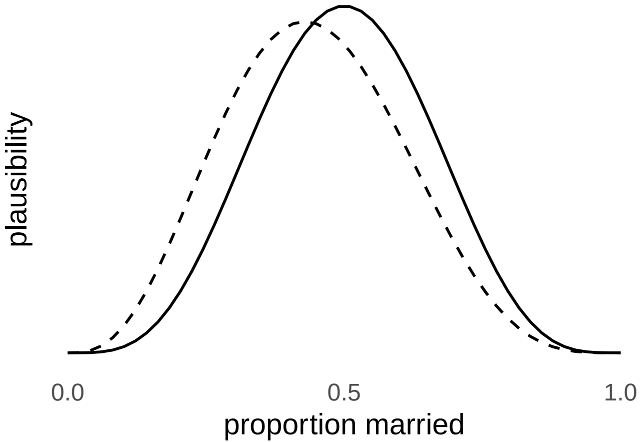
Fifth observation: Single

Sixth observation: Single

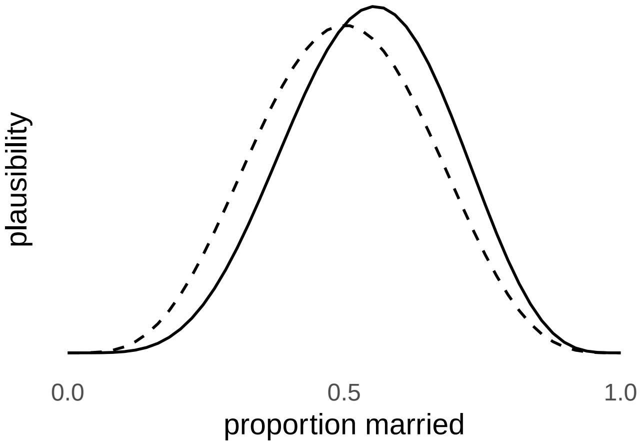
Seventh observation: Single



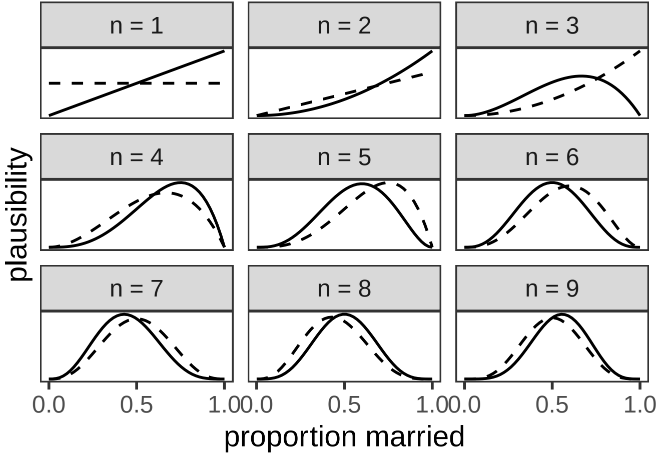
Eighth observation: Married



Nineth observation: Single



Overview



n = 1
n = 2
n = 3
n = 4
n = 5
n = 6
n = 7
n = 8
n = 9
plausibility
0.0
0.5
1.0
proportion married

Bayesian Updating

- This example demonstrates the concept of **Bayesian updating**
  - We use new information to update our beliefs
- Each time we update we use the previous **posterior** as the new **prior**!

Bayesian Updating

- This example demonstrates the concept of **Bayesian updating**
  - We use new information to update our beliefs
- Each time we update we use the previous **posterior** as the new **prior**!
- Most of the time we use all our data at once to get the final posterior rather than iteratively updating.

Bayesian Updating

- This example demonstrates the concept of **Bayesian updating**
  - We use new information to update our beliefs
- Each time we update we use the previous **posterior** as the new **prior**!
- Most of the time we use all our data at once to get the final posterior rather than iteratively updating.
- Bayesian updating is order invariant: we will get the same result regardless of the way observations are ordered.

Formalizing a model

- The previous calculations are an example of the *binomial distribution*
  - Recall the distribution has two parameters N and p
- The goal of this analysis is to produce an estimate of the parameter p.
- We can thus write down a model to describe our analysis of marriage:

$$Marriage \sim Binomial(N, p)$$

$$p \sim Uniform(0, 1)$$

- The goal of this analysis is to produce an estimate of the parameter p. In this case, we started with a flat, uniform prior.

Comparing Bayesian and Frequentist approaches

Thomas Bayes (1701-1761)

Source: Wikipedia.

Comparing Bayesian and Frequentist approaches

Pierre-Simon Laplace (1749-1827)

Source: Wikipedia.

Comparing Bayesian and Frequentist approaches

Ronald Fisher (1890-1962)

Source: Wikipedia.

Comparing Bayesian and Frequentist approaches

Historical developments

- Frequentist (or "Fisherian") statistics dominated for most of the 20th century.
- Bayesian inference critiqued as too subjective and difficult to implement for complex problems.
- Reversal over the past couple of decades as critiques of Bayesian approach debunked, cheap compute power makes it tractable, and key tenets of Frequentist statistics are questioned (e.g. controversy over p-hacking[3]).
- The Bayesian approach is now mainstream in statistics and much of the natural sciences, but the social sciences have been slower to adopt.[4]

[3] See Imbens 2021 reading from Week 1.

[4] See Scott and Bartlett 2019.

Comparing Bayesian and Frequentist approaches

Theoretical foundations

- Frequentist
  - Long-run probabilities
  - Sampling distributions
- Bayesian
  - Probability theory

Comparing Bayesian and Frequentist approaches

Sample size

- Frequentist
  - Properties of estimators depend on minimal sample size
- Bayesian
  - No minimum sample size
  - But larger samples allow for more precise estimates

Comparing Bayesian and Frequentist approaches

Point estimates

- Frequentist
  - Models produce point estimates
- Bayesian
  - No singular point estimates
    - Many different summaries of the posterior distribution are possible (e.g. mean, median, mode)

Comparing Bayesian and Frequentist approaches

P-values

- Frequentist
  - p-values used to communicate statistical significance
- Bayesian
  - Critique: p-values are based on arbitrary distributional assumptions
  - Uncertainty is captured by entire posterior distribution
  - *Bayes' Factor* is a Bayesian version of a p-value[5]

[5] See Kruschke and Liddell 2018.

Comparing Bayesian and Frequentist approaches

Confidence intervals

- Frequentist
  - Confidence intervals defined using test statistics and conventions
  - Assumption that a parameter is fixed and that interval is derived from a sample
- Bayesian
  - Critique: Frequentist conventions are arbitrary
  - Assumption that a parameter has a distribution
  - *Credible intervals* or *compatibility intervals* can be used to summarize the posterior distribution

Comparing Bayesian and Frequentist approaches

Confidence intervals: Interpretation of a 95% interval

- Frequentist
  - Over many repeat samples, 95% of calculated confidence intervals would contain the true value of the parameter
- Bayesian (assume an interval over 95% of the posterior distribution)
  - There is a 95% probability that the estimated parameter lies within the defined range, given the model and the data.
  - "What the interval indicates is a range of parameter values compatible with the model and the data." McElreath, p. 54.

Computation and Bayesian Estimation

Bayesian Estimation

- Three methods for estimating the posterior distribution
  - Analytical calculations
  - Grid and quadratic approximation
  - Markov Chain Monte Carlo

Computation and Bayesian Estimation

Analytical calculations

- For simple problems we can use calculus to provide an analytical solution for the posterior distribution
- But this approach does not scale well beyond simple problems like the marriage example

Computation and Bayesian Estimation

Grid and quadratic approximation

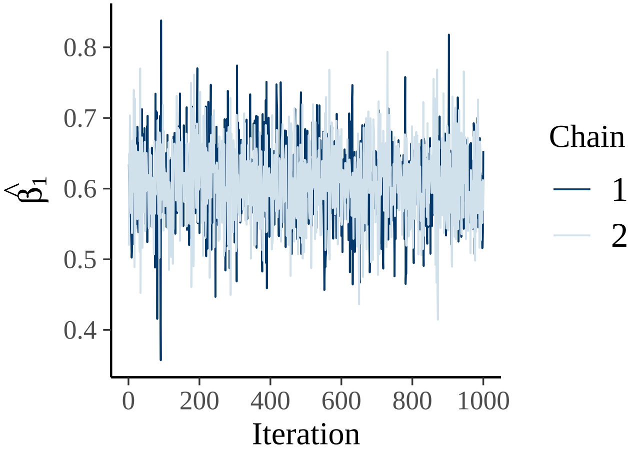
- Grid approximation (see McElreath 2.4.3)
  - We can approximate continuous spaces by using grids
    - But the method also scales very poorly to complex examples
- Quadratic approximation (see McElreath 2.4.4)
  - A more robust approach that involves using distributions to approximate the posterior
  - Flexible for many regression problems but also has trouble scaling

Computation and Bayesian Estimation

Markov Chain Monte Carlo (MCMC)

- Use simulation to draw samples from the posterior distribution
  - A computationally intensive approach
  - Samples provide an approximation for complex spaces
  - More efficient for complex models than quadratic approximation
- MCMC has led to major advances in Bayesian methods since the 1990s (see McElreath 2.4.5).

Samples from a Markov Chain



Computation and Bayesian Estimation

Stan and Hamiltonian Monte Carlo

- `Stan` is a programming language developed for statistical computing
- It implements **Hamiltonian Monte Carlo (HMC)** sampling
  - A variant of MCMC methods based on Hamiltonian physics
  - Approximates the posterior by "flicking" a particle and observing its movement
- HMC is highly effective at solving even complex problems[6]
  - It provides lots of useful diagnostics making it easier to debug than early MCMC approaches
  - Greater flexibility as it not require *conjugacy*

[6] See McElreath Chapter 9 and Betancourt 2018 for a more advanced conceptual overview.

Bayesian Regression

- Regression coefficients are the *unknown* parameters that we want to estimate given a model and the observed data.
- We can formalize these assumptions by writing down a model that looks something like this:

$$y_i \sim Normal(\mu_i, \sigma)$$

$$\mu_i = \beta_0 + \beta_1 x_i$$

$$\beta_0 \sim Normal(0, 10)$$

$$\beta_1 \sim Normal(0, 1)$$

$$\sigma \sim Uniform(0, 1)$$

Bayesian Regression

- In this case, we make the *assumption* that y_i is normally distributed and that we can express its mean in terms of x (recall that $E[y|x] = \beta_0 + \beta_1 x_i$)
- Each *parameter* in the model has a *prior* distribution. We specify these before we have seen any data
- After estimating a model using the data we get the *posterior* distribution for each parameter

Bayesian Regression

- We will be using `stan_glm` to estimate these kinds of models using HMC
- The *posterior distributions* of the parameters are then analyzed to make inferences about the relationship between x and y
- We can also use the posterior to generate new data consistent with the model

Final remarks

"All models are wrong, but some are useful" - George Box[7]

[7] This aphorism is attributed to statistician George Box. See Wikipedia for further discussion.