Bayes' Rule

PSYC 573

University of Southern California
January 25, 2022

Inverse Probability

Conditional probability: $P(A \mid B) = \dfrac{P(A, B)}{P(B)}$

which yields $P(A, B) = P(A \mid B)P(B)$ (joint = conditional $\times$ marginal)

On the other side, $P(B \mid A) = \dfrac{P(B, A)}{P(A)}$

Bayes' Theorem

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

Which says how can go from $P(A \mid B)$ to $P(B \mid A)$

Consider B_i $(i = 1, \ldots, n)$ as one of the many possible mutually exclusive events

$$\begin{aligned} P(B_i \mid A) &= \frac{P(A \mid B_i)P(B_i)}{P(A)} \\ &= \frac{P(A \mid B_i)P(B_i)}{\sum_{k=1}^{n} P(A \mid B_k)P(B_k)} \end{aligned}$$

> A police officer stops a driver *at random* and does a breathalyzer test for the driver. The breathalyzer is known to detect true drunkenness 100% of the time, but in **1%** of the cases, it gives a *false positive* when the driver is sober. We also know that in general, for every **1,000** drivers passing through that spot, **one** is driving drunk. Suppose that the breathalyzer shows positive for the driver. What is the probability that the driver is truly drunk?

Gigerenzer (2004)

p value = P(data | hypothesis), not P(hypothesis | data)

- H_0: the person is sober (not drunk)
- data: breathalyzer result

p = P(positive | sober) = 0.01 $\longrightarrow$ reject H_0 at .05 level

However, as we have been, given that $P(H_0)$ is small, $P(H_0 \mid \mathbf{data})$ is still small

Bayesian Data Analysis

Bayes' Theorem in Data Analysis

- Bayesian statistics
  - more than applying Bayes's theorem
  - a way to quantify the plausibility of every possible value of some parameter $\boldsymbol{\theta}$
    - E.g., population mean, regression coefficient, etc
  - Goal: **update one's Belief about θ based on the observed data D**

Going back to the example

Goal: Find the probability that the person is drunk, given the test result

Parameter (θ): drunk (values: drunk, sober)

Data (D): test (possible values: positive, negative)

Bayes' theorem: $\underbrace{P(\theta \mid D)}_{\text{posterior}} = \underbrace{P(D \mid \theta)}_{\text{likelihood}} \underbrace{P(\theta)}_{\text{prior}} / \underbrace{P(D)}_{\text{marginal}}$

Usually, the marginal is not given, so

$$P(\theta \mid D) = \frac{P(D \mid \theta)P(\theta)}{\sum_{\theta^*} P(D \mid \theta^*)P(\theta^*)}$$

- $P(D)$ is also called *evidence*, or the *prior predictive distribution*
  - E.g., probability of a positive test, regardless of the drunk status

Example 2

```
shiny::runGitHub("plane_search", "marklhc")
```

- Try choosing different priors. How does your choice affect the posterior?
- Try adding more data. How does the number of data points affect the posterior?

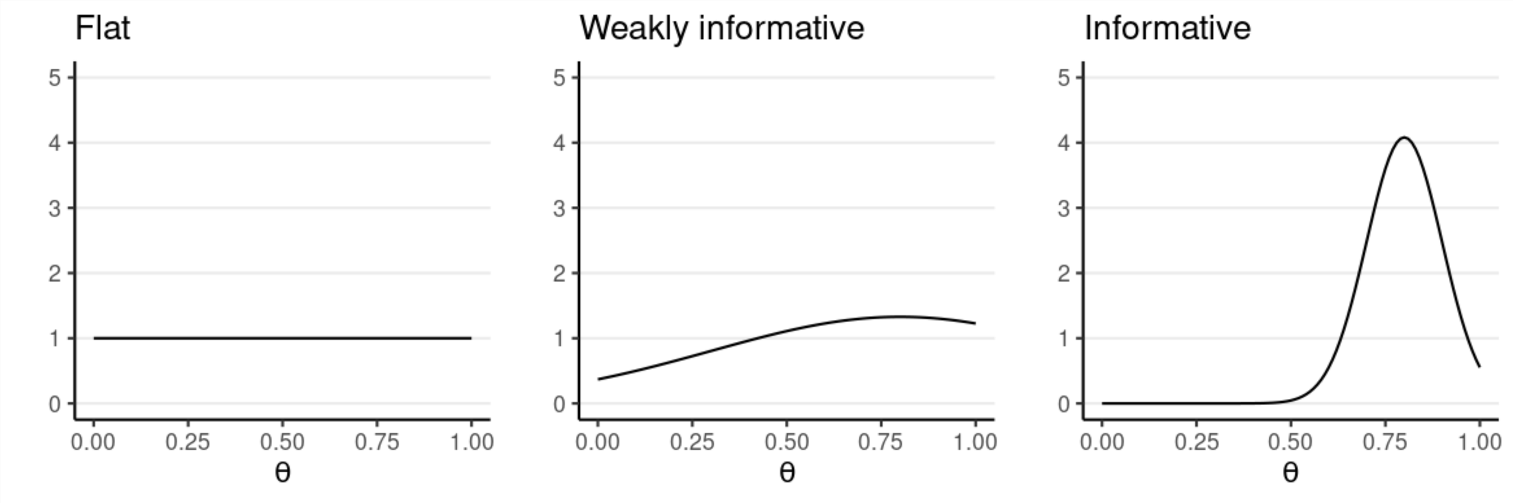
The posterior is a synthesis of two sources of information: prior and data (likelihood)

Generally speaking, a narrower distribution (i.e., smaller variance) means more/stronger information

- Prior: narrower = more informative/strong
- Likelihood: narrower = more data/more informative

Setting Priors

- Flat, noninformative, vague
- Weakly informative: common sense, logic
- Informative: publicly agreed facts or theories

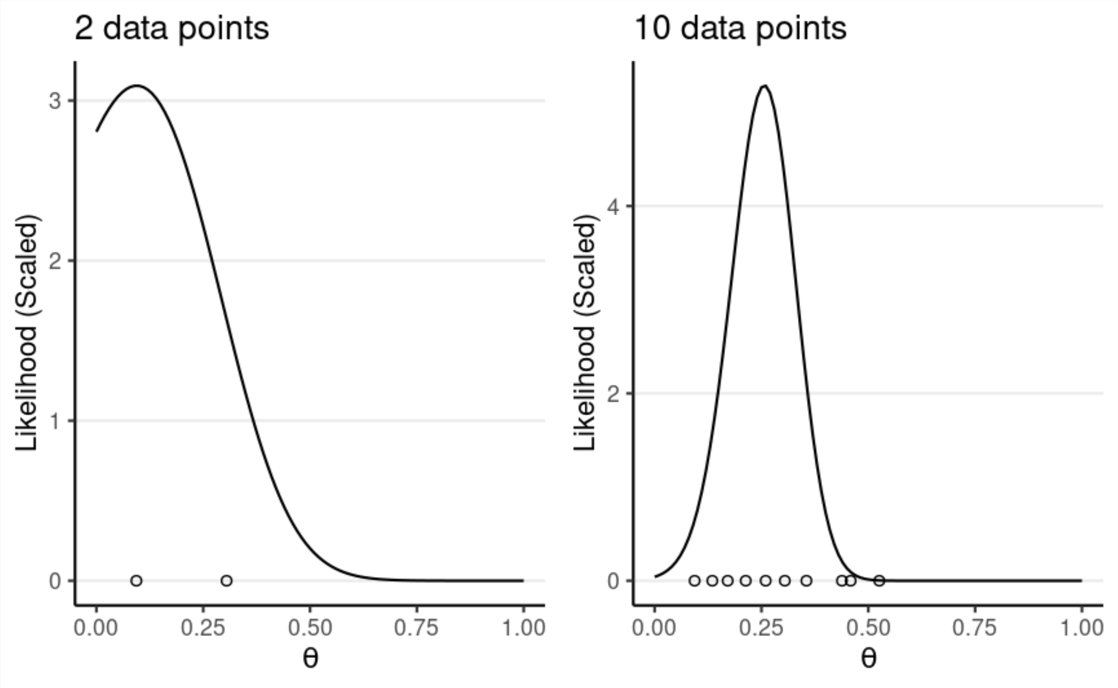


> *Prior beliefs used in data analysis must be admissible by a skeptical scientific audience (Kruschke, 2015, p. 115)*

Likelihood/Model/Data $P(D \mid \theta, M)$

> Probability of observing the data **as a function of the parameter(s)**

- Also written as $L(\theta \mid D)$ or $L(\theta; D)$ to emphasize it is a function of θ
- Also depends on a chosen model M: $P(D \mid \theta, M)$



Likelihood of Multiple Data Points

1. Given D_1, obtain *posterior* $P(\theta \mid D_1)$
2. Use $P(\theta \mid D_1)$ as *prior*, given D_2, obtain posterior $P(\theta \mid D_1, D_2)$

The posterior is the same as getting D_2 first then D_1, or D_1 and D_2 together, if

- **data-order invariance** is satisfied, which means
- D_1 and D_2 are **exchangeable**

Joint distribution of the data does not depend on the order of the data

E.g., $P(D_1, D_2, D_3) = P(D_2, D_3, D_1) = P(D_3, D_2, D_1)$

Example of non-exchangeable data:

- First child = male, second = female vs. first = female, second = male
- D_1, D_2 from School 1; D_3, D_4 from School 2 vs. D_1, D_3 from School 1; D_2, D_4 from School 2

An Example With Binary Outcomes

Coin Flipping

Q: Estimate the probability that a coin gives a head

- θ: parameter, probability of a head

Flip a coin, showing head

- $y = 1$ for showing head

> How do you obtain the likelihood?

Bernoulli Likelihood

The likelihood depends on the probability model chosen

- Some models are commonly used, for historical/computational/statistical reasons

One natural way is the **Bernoulli model**

$$P(y = 1 \mid \theta) = \theta$$
$$P(y = 0 \mid \theta) = 1 - \theta$$

The above requires separating $y = 1$ and $y = 0$. A more compact way is

$$P(y \mid \theta) = \theta^y (1 - \theta)^{(1-y)}$$

Multiple Observations

Assume the flips are exchangeable given θ,

$$
\begin{aligned}
P(y_1, \ldots, y_N \mid \theta) &= \prod_{i=1}^{N} P(y_i \mid \theta) \\
&= \theta^{\sum_{i=1}^{N} y_i} (1 - \theta)^{\sum_{i=1}^{N} (1 - y_i)} \\
&= \theta^{z} (1 - \theta)^{N - z}
\end{aligned}
$$

z = # of heads; N = # of flips

> Note: the likelihood only depends on the number of heads, not the particular sequence of observations

Posterior

Same posterior, two ways to think about it

Prior belief, weighted by the likelihood

$$P(\theta \mid y) \propto \underbrace{P(y \mid \theta)}_{\text{weights}} P(\theta)$$

Likelihood, weighted by the strength of prior belief

$$P(\theta \mid y) \propto \underbrace{P(\theta)}_{\text{weights}} P(\theta \mid y)$$

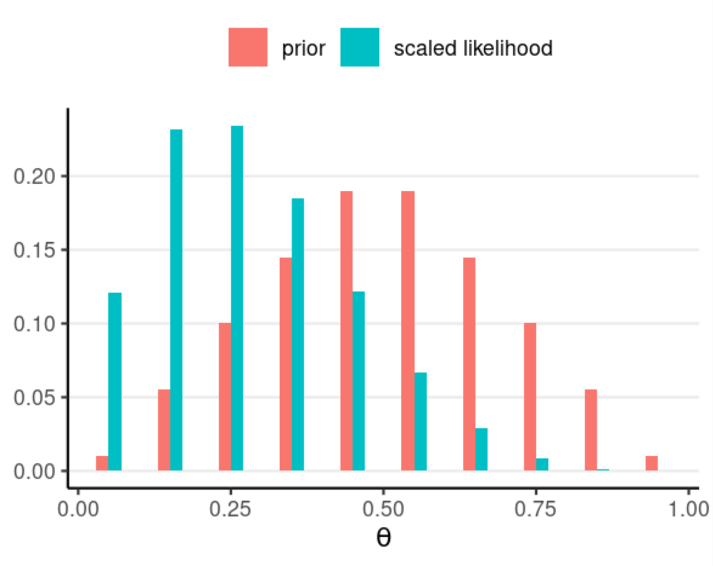
Posterior

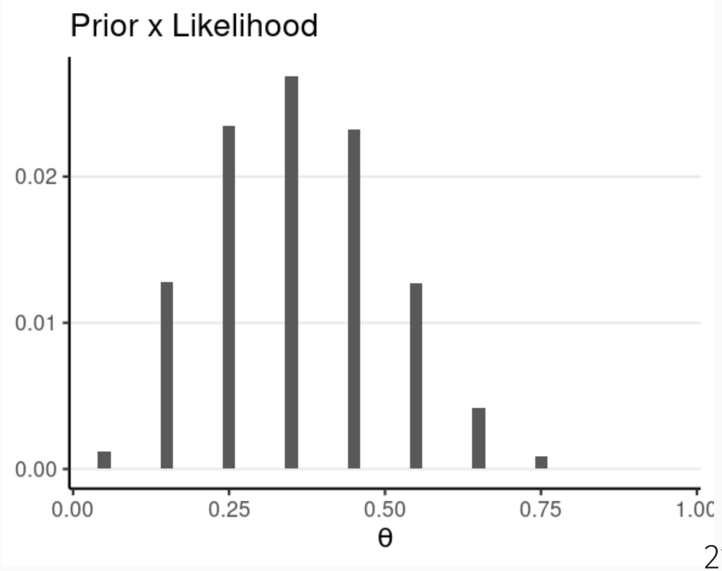
Say N = 4 and z = 1

E.g., $P(\theta \mid y_1 = 1) \propto P(y_1 = 1 \mid \theta) P(\theta)$

For pedagogical purpose, we'll discretize the θ into [.05, .15, .25, ..., .95]

- Also called **grid approximation**





How About the Denominator?

Numerator: relative posterior plausibility of the θ values

We can avoid computing the denominator because

- The sum of the probabilities need to be 1

So, for **discrete** parameters:

- Posterior probabilitiy = relative plausibility / sum of relative plausibilities

However, the denominator is useful for computing the *Bayes factor*

Summarizing a Posterior Distribution

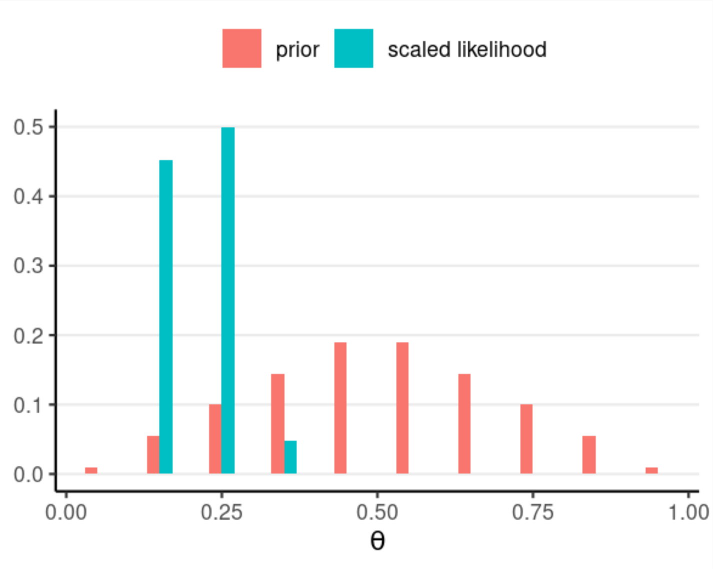
Simulate (i.e., draw samples) from the posterior distribution

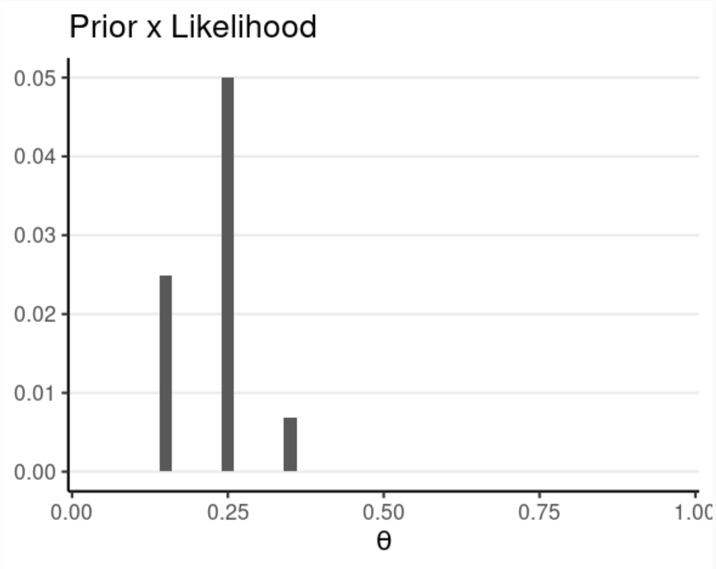
R code Summary

```
th <- seq(.05, .95, by = .10)
pth <- c(.01, .055, .10, .145, .19, .19, .145, .10, .055, .01
py_th <- th^1 * (1 - th)^4
pth_y_unscaled <- pth * py_th
pth_y <- pth_y_unscaled / sum(pth_y_unscaled)
post_samples <- sample(th,
  size = 1000, replace = TRUE,
  prob = pth_y
)
```

Influence of more samples

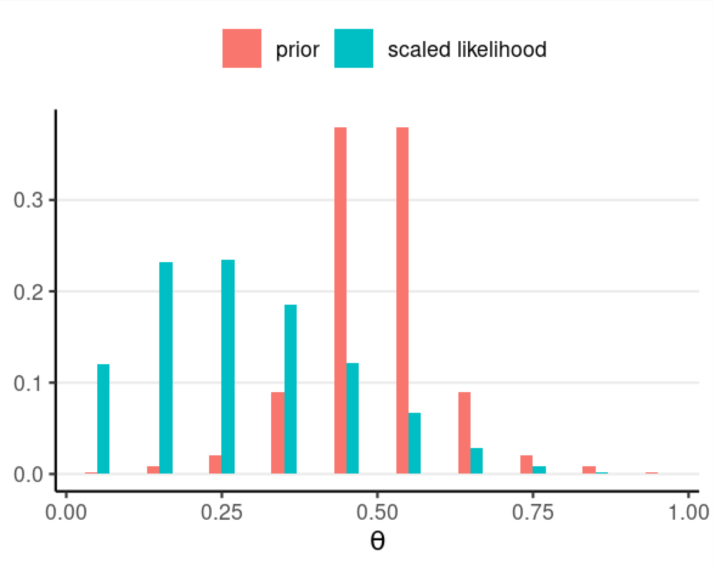
$N = 40$, $z = 10$

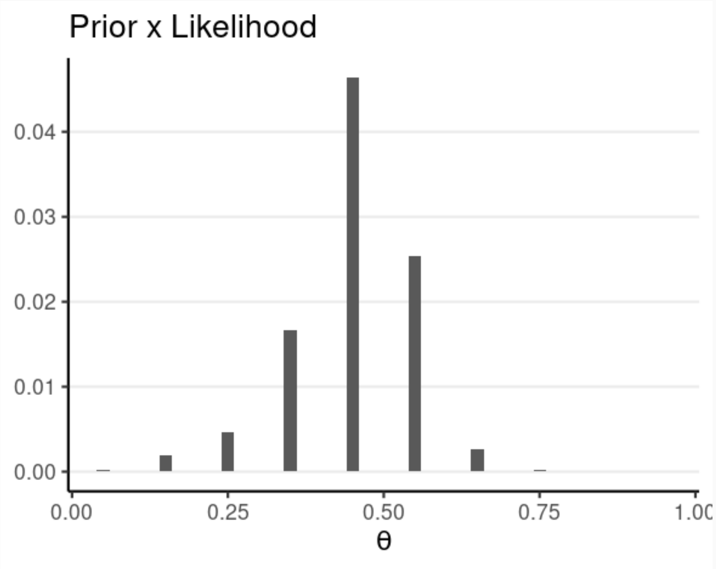




Influence of more informative priors

$N = 4$, $z = 1$





The prior needs to be well justified

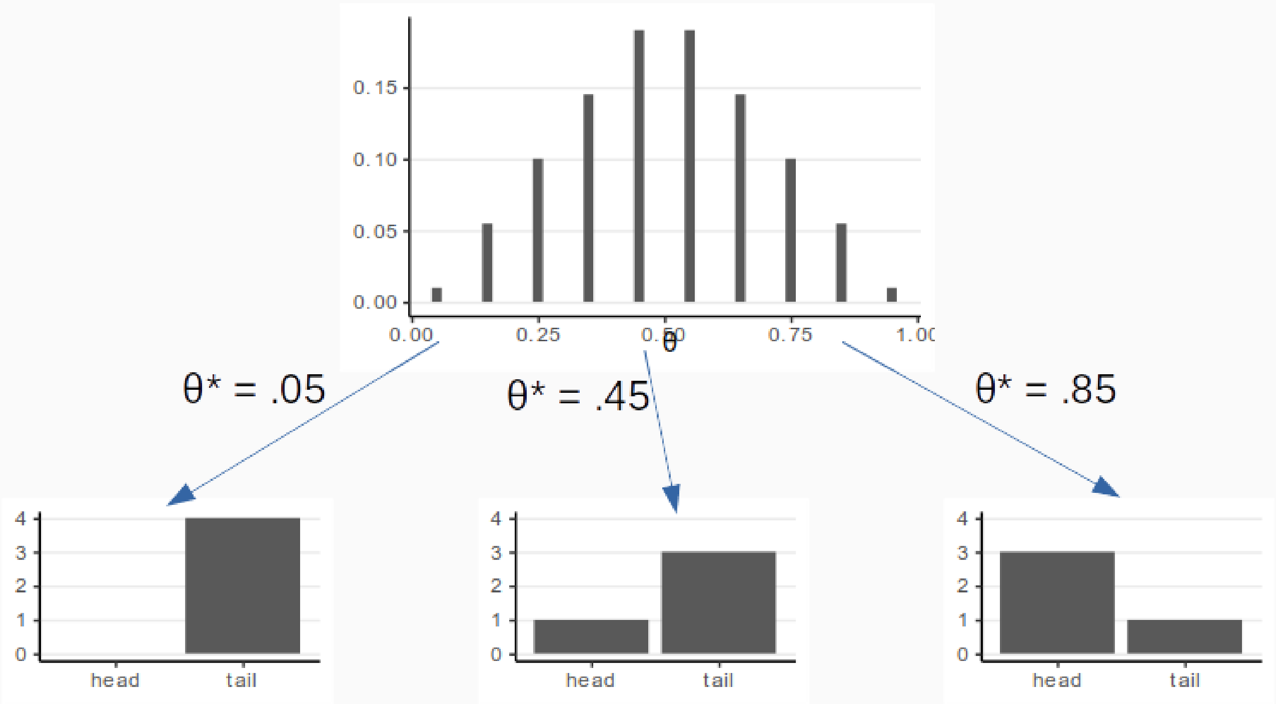
Prior Predictive Distribution

> Bayesian models are **generative**

Simulate data from the prior distribution to check whether the data fit our intuition

- Clearly impossible values/patterns?
- Overly restrictive?

$P(y) = \int P(y|\theta^*)P(\theta^*)d\theta^*$: Simulate a θ^* from the prior, then simulate data based on θ^*



Criticism of Bayesian Methods

Criticism of "Subjectivity"

Main controversy: subjectivity in choosing a prior

- Two people with the same data can get different results because of different chosen priors

Counters to the Subjectivity Criticism

- With enough data, different priors hardly make a difference
- Prior: just a way to express the degree of ignorance
    - One can choose a weakly informative prior so that the Influence of subjective Belief is small

Counters to the Subjectivity Criticism 2

Subjectivity in choosing a prior is

- Same as in choosing a model, which is also done in frequentist statistics
- Relatively strong prior needs to be justified,
    - Open to critique from other researchers
- Inter-subjectivity $\longrightarrow$ Objectivity

Counters to the Subjectivity Criticism 3

The prior is a way to incorporate previous research efforts to accumulate scientific evidence

> Why should we ignore all previous literature every time we conduct a new study?