Agenda

Probability models of social processes

1. Administrative
2. Probability of unemployment
3. Bayes' rule
4. *Hands on: random samples and grid approximations in R*

Software check in

- Script to test required software: https://soci620.netlify.app/labs/handson_0_testing.R
- Copy and paste, or download to your computer and run

Getting started with R

- Norm Matloff's "fastR" introduction is simple and good: https://github.com/matloff/fasteR
- For this class, completing (and understsanding!) the **first eight lessons** will give you a good foundation

Labs

- Mondays 10-11am look like they'll work, but not in this room
- Next week (Jan 13): Leacock 808
- Subsequent weeks: TBD

PROBABILITY OF UNEMPLOYMENT

Unemployment in Newfoundland and Labrador

- How do we learn something about the risk of unemployment for adult residents of NL?
- Ignoring (for now) contributing factors, we can ask:
  What is the probability that a randomly chosen adult is unemployed?
  i.e. *unemployment rate* (frequentist) – 9.6% according to StatCan

Probability model

- **Strategy**: *model* the process with a parametric probability distibution, and *estimate* the model with a sample
- Assuming we sample n adults, k of whom report being unemployed, we'll model this with a ***binomial distribution***
- *(Details on what this means shortly, but first, a demonstration)*

PROBABILITY OF UNEMPLOYMENT

Data:

Unfortunately, our grant has run out, so we can only afford to sample 10 people:

$$S = (E, E, E, U, U, E, E, E, U, E)$$

$$n = 10$$

$$k = 3$$

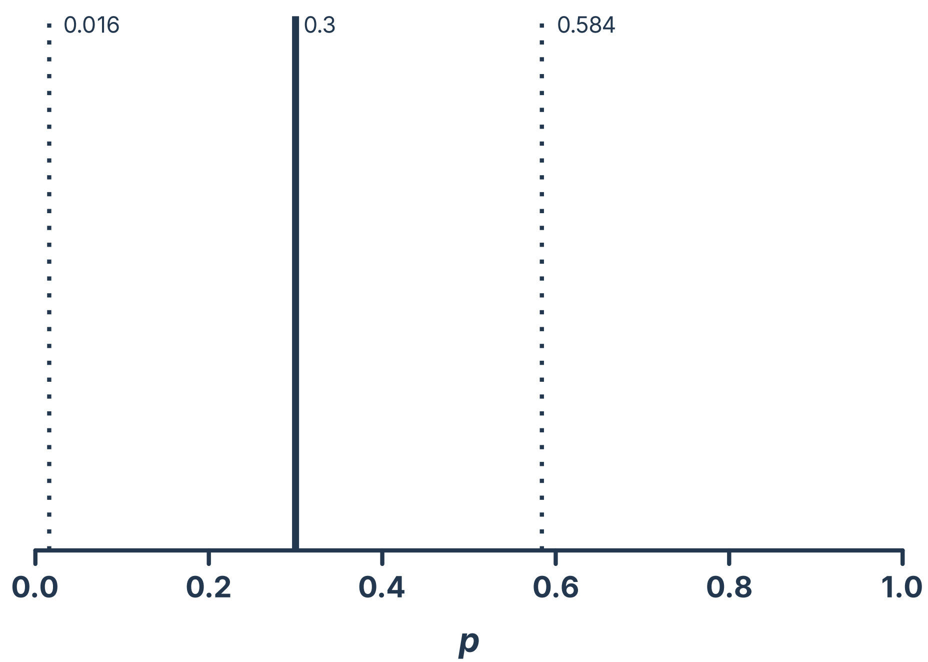
We'll use this data to estimate the probability of unemployment in *two ways*

PROBABILITY OF UNEMPLOYMENT

$$S = (E, E, E, U, U, E, E, E, U, E)$$

Maximimum-likelihood (frequentist) estimation:

1. Pick an *estimator* (such as sample proportion) of the probability p
2. Generate a *point estimate* of p:
$$\hat{p} = \frac{3}{10} = 0.3$$
3. Use an approximation of the sampling distribution to quantify uncertainty:
$$\hat{\sigma} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.145$$
4. Generate a *confidence interval* (e.g. $1.96 \times \hat{\sigma}$ for standard 95% CI)

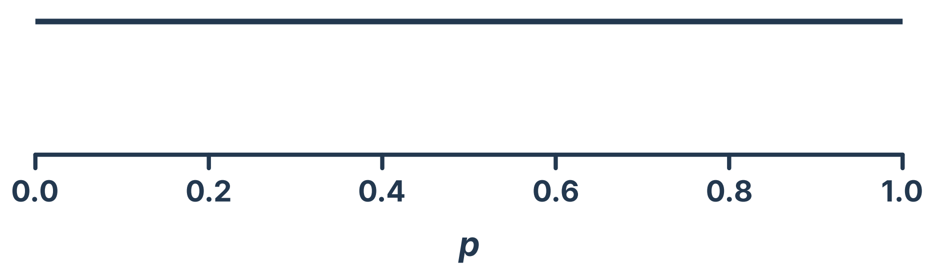


PROBABILITY OF UNEMPLOYMENT

$$S = (E, E, E, U, U, E, E, E, U, E)$$

Posterior (Bayesian) estimation:

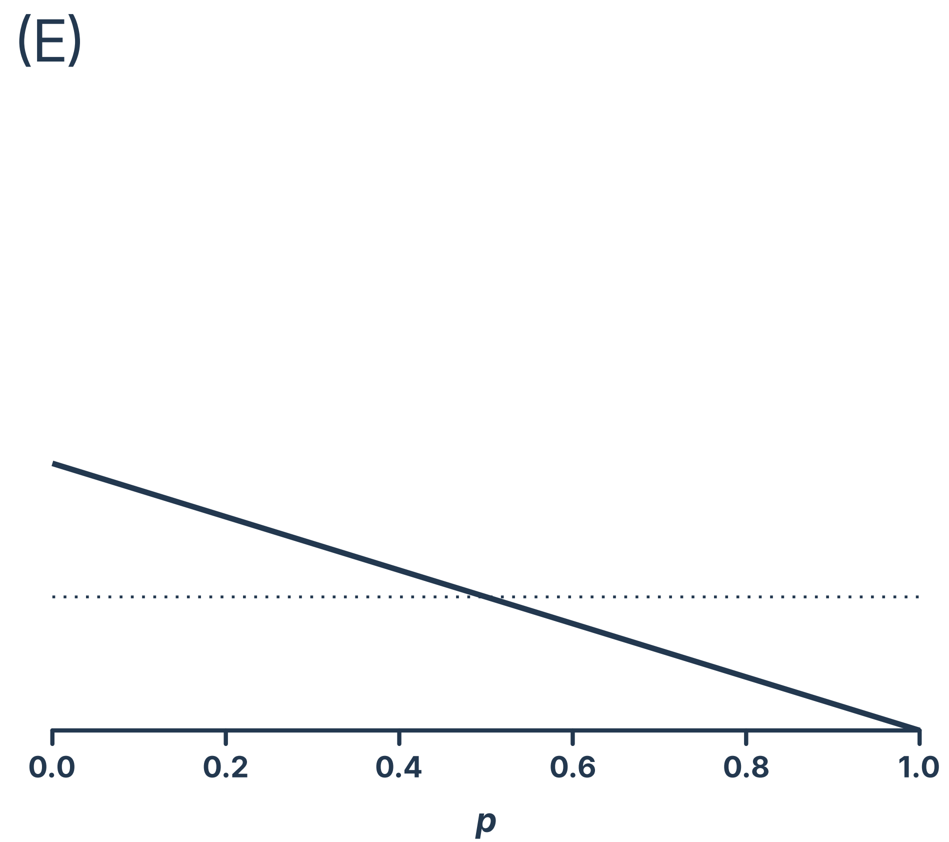
1. Pick a *prior* (such as a uniform distribution) for the probability p



$$S = (E, E, E, U, U, E, E, E, U, E)$$

Posterior (Bayesian) estimation:

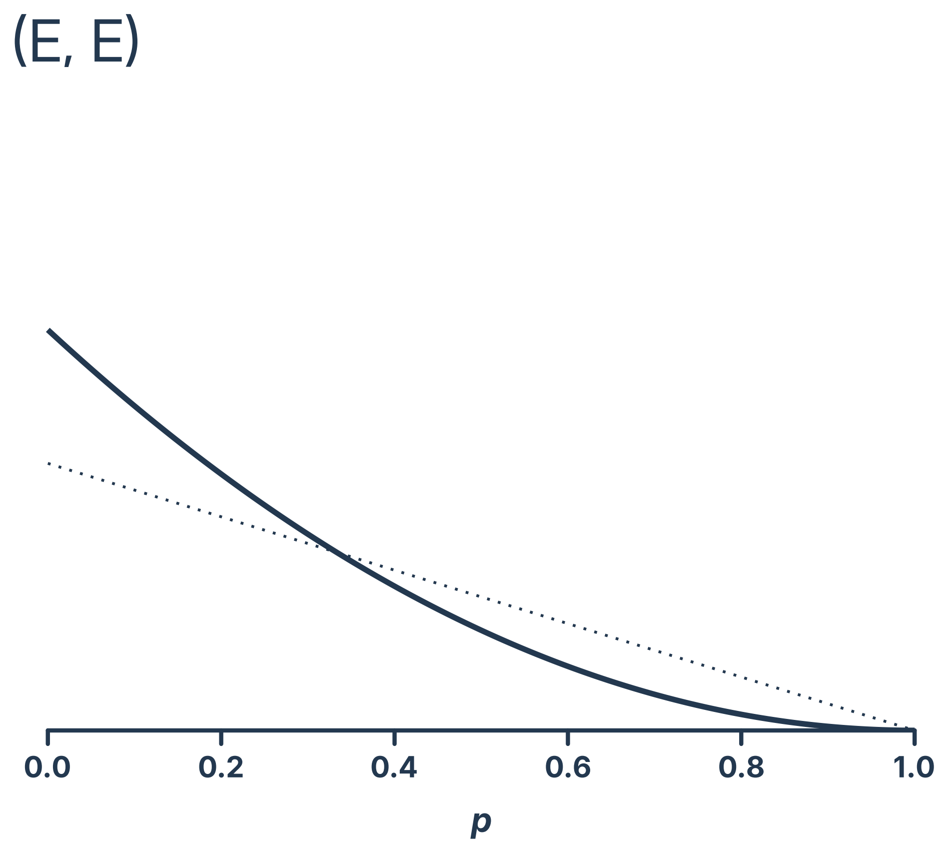
1. Pick a *prior* (such as a uniform distribution) for the probability p
2. Update prior with data (one at a time or all at once)



$$S = (E, E, E, U, U, E, E, E, U, E)$$

Posterior (Bayesian) estimation:

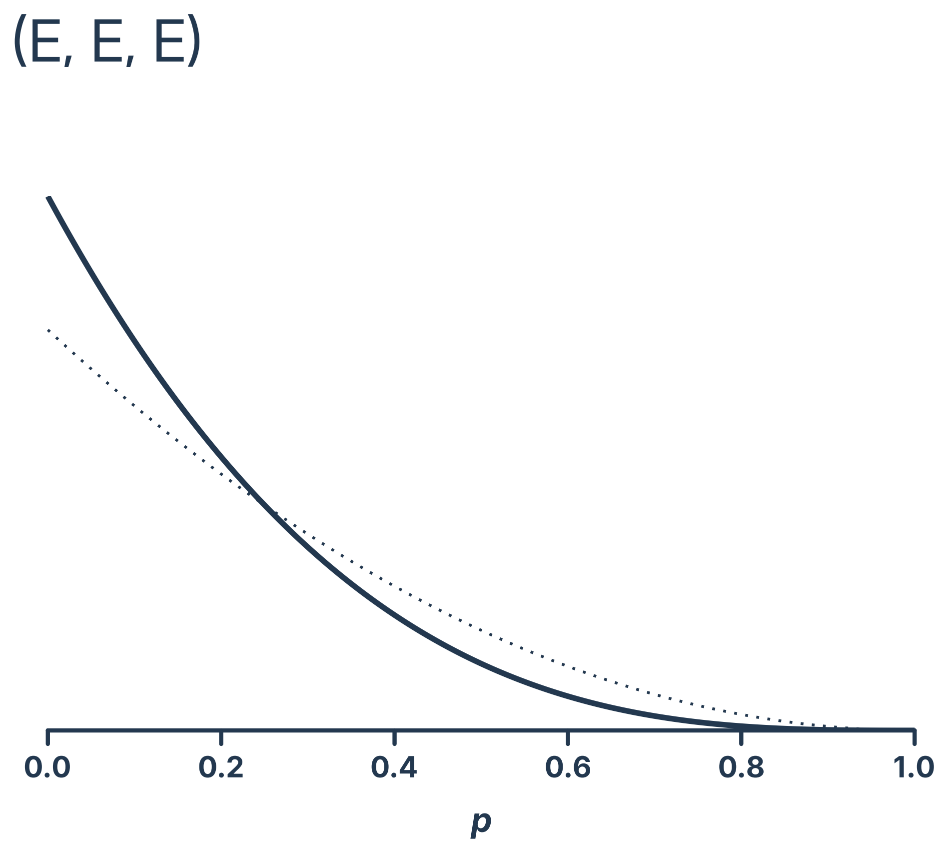
1. Pick a *prior* (such as a uniform distribution) for the probability p
2. Update prior with data (one at a time or all at once)



$$S = (E, E, E, U, U, E, E, E, U, E)$$

Posterior (Bayesian) estimation:

1. Pick a *prior* (such as a uniform distribution) for the probability p
2. Update prior with data (one at a time or all at once)

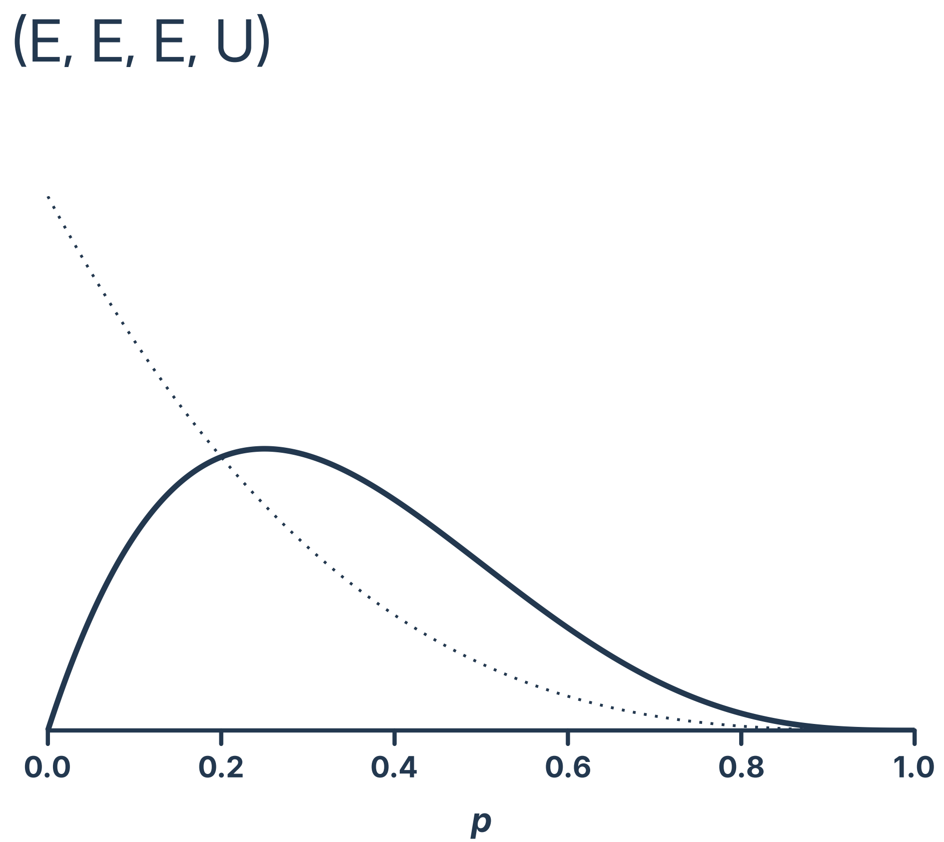


PROBABILITY OF UNEMPLOYMENT

$$S = (E, E, E, U, U, E, E, E, U, E)$$

Posterior (Bayesian) estimation:

1. Pick a *prior* (such as a uniform distribution) for the probability p
2. Update prior with data (one at a time or all at once)



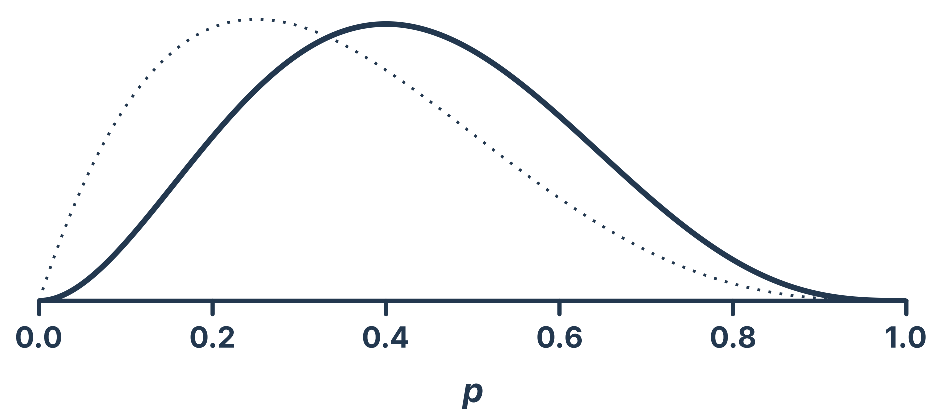
PROBABILITY OF UNEMPLOYMENT

$$S = (E, E, E, U, U, E, E, E, U, E)$$

Posterior (Bayesian) estimation:

1. Pick a *prior* (such as a uniform distribution) for the probability p
2. Update prior with data (one at a time or all at once)

(E, E, E, U, U)



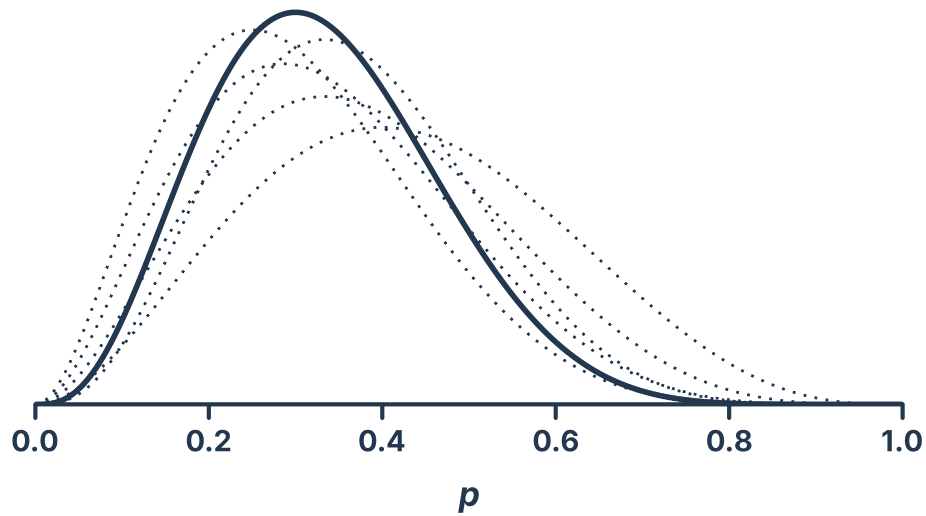
PROBABILITY OF UNEMPLOYMENT

$$S = (E, E, E, U, U, E, E, E, U, E)$$

Posterior (Bayesian) estimation:

1. Pick a *prior* (such as a uniform distribution) for the probability p
2. Update prior with data (one at a time or all at once)

(E, E, E, U, U, E, E, E, U, E)



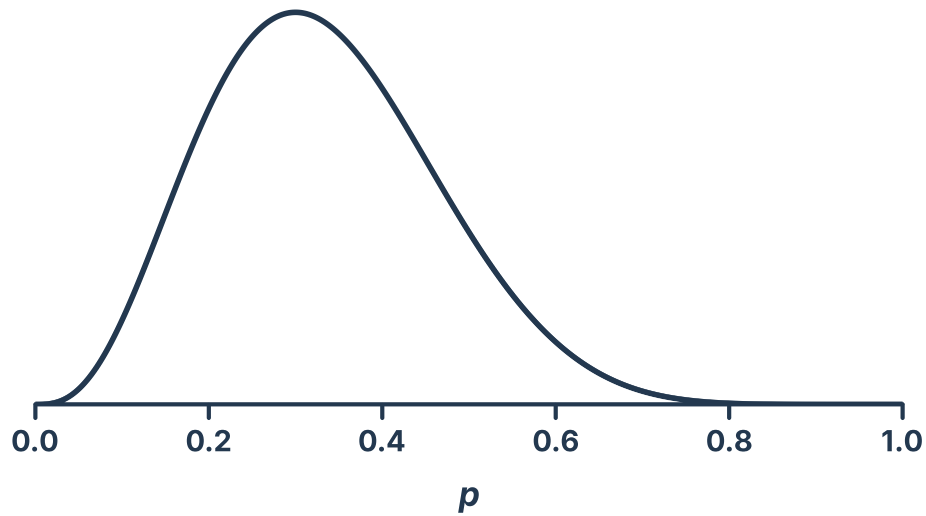
PROBABILITY OF UNEMPLOYMENT

$$S = (E, E, E, U, U, E, E, E, U, E)$$

Posterior (Bayesian) estimation:

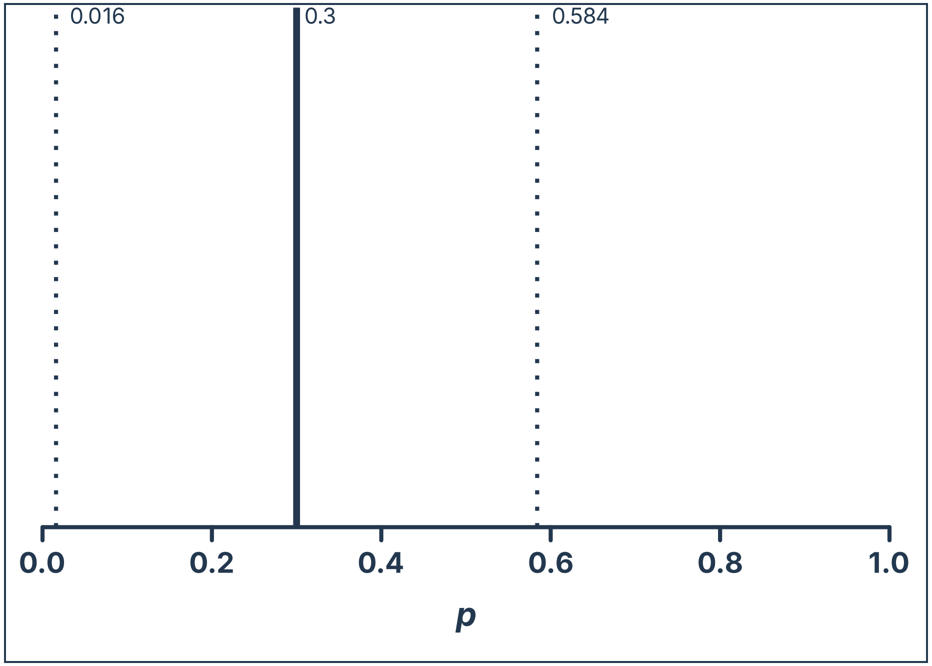
1. Pick a *prior* (such as a uniform distribution) for the probability p
2. Update prior with data (one at a time or all at once)
3. The *posterior distribution* describes the relative *posterior probability* for different values of p

(E, E, E, U, U, E, E, E, U, E)

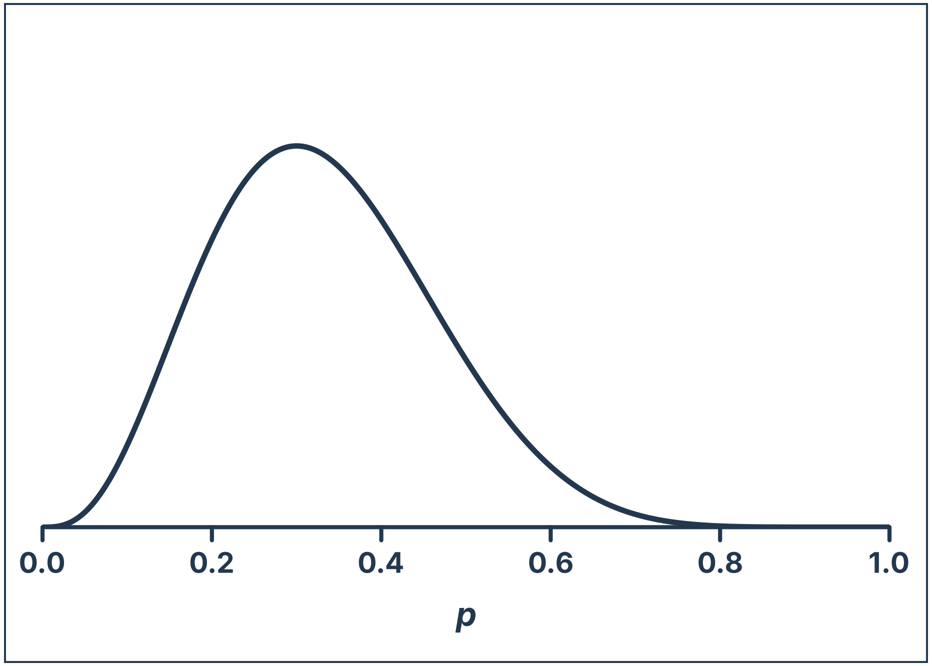


COMPARING ESTIMATES

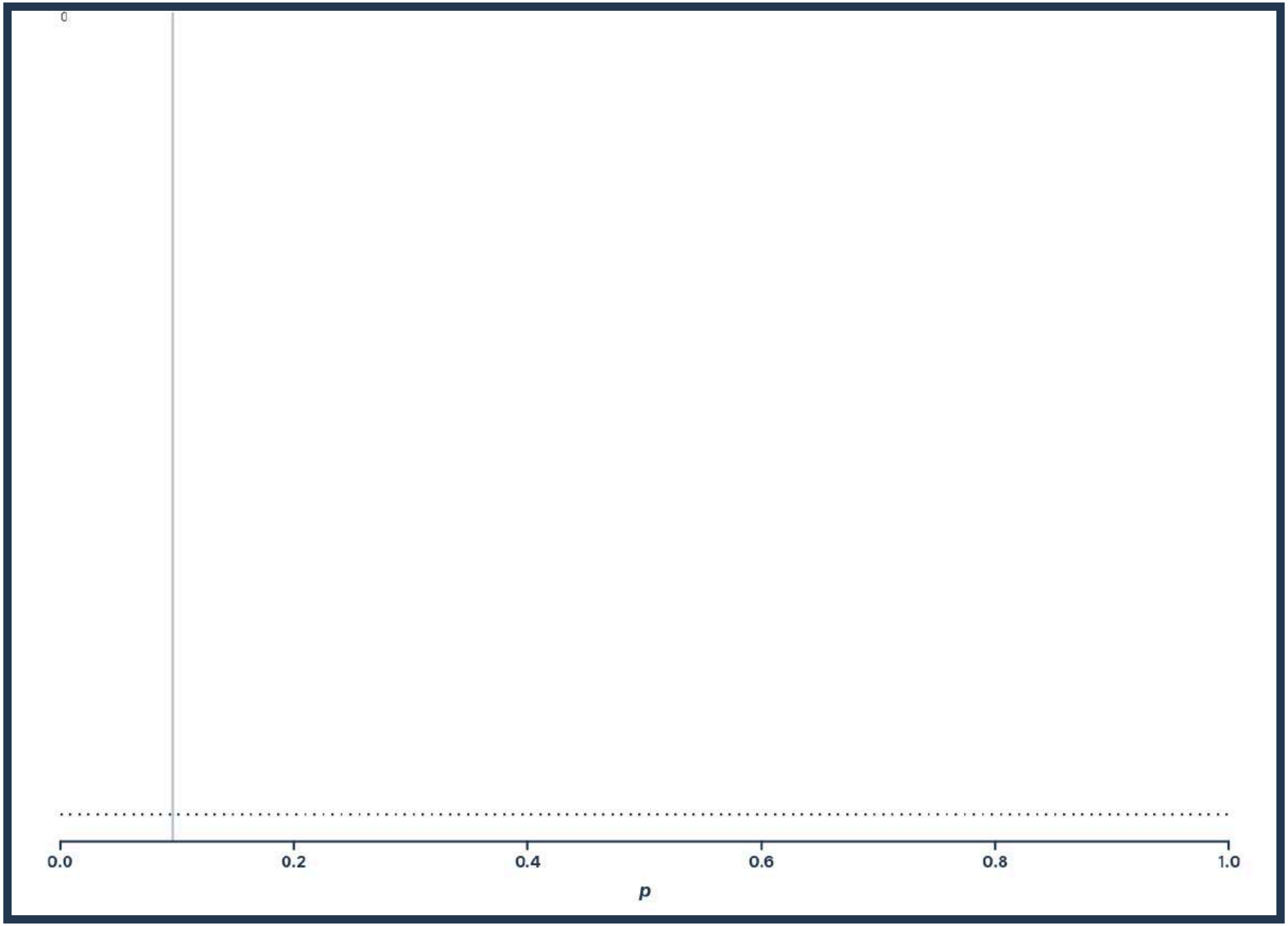
Maximum likelihood:



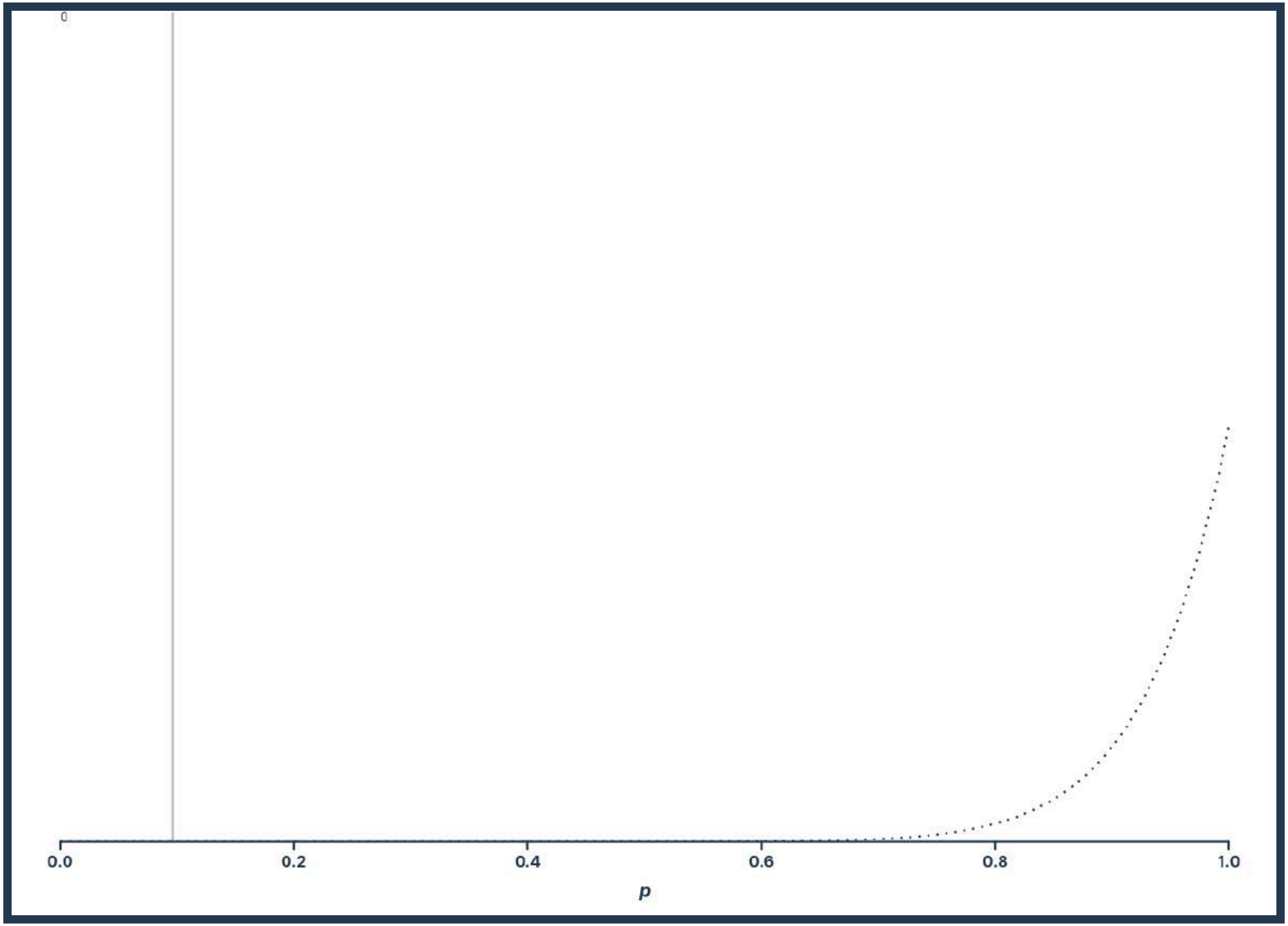
Posterior:



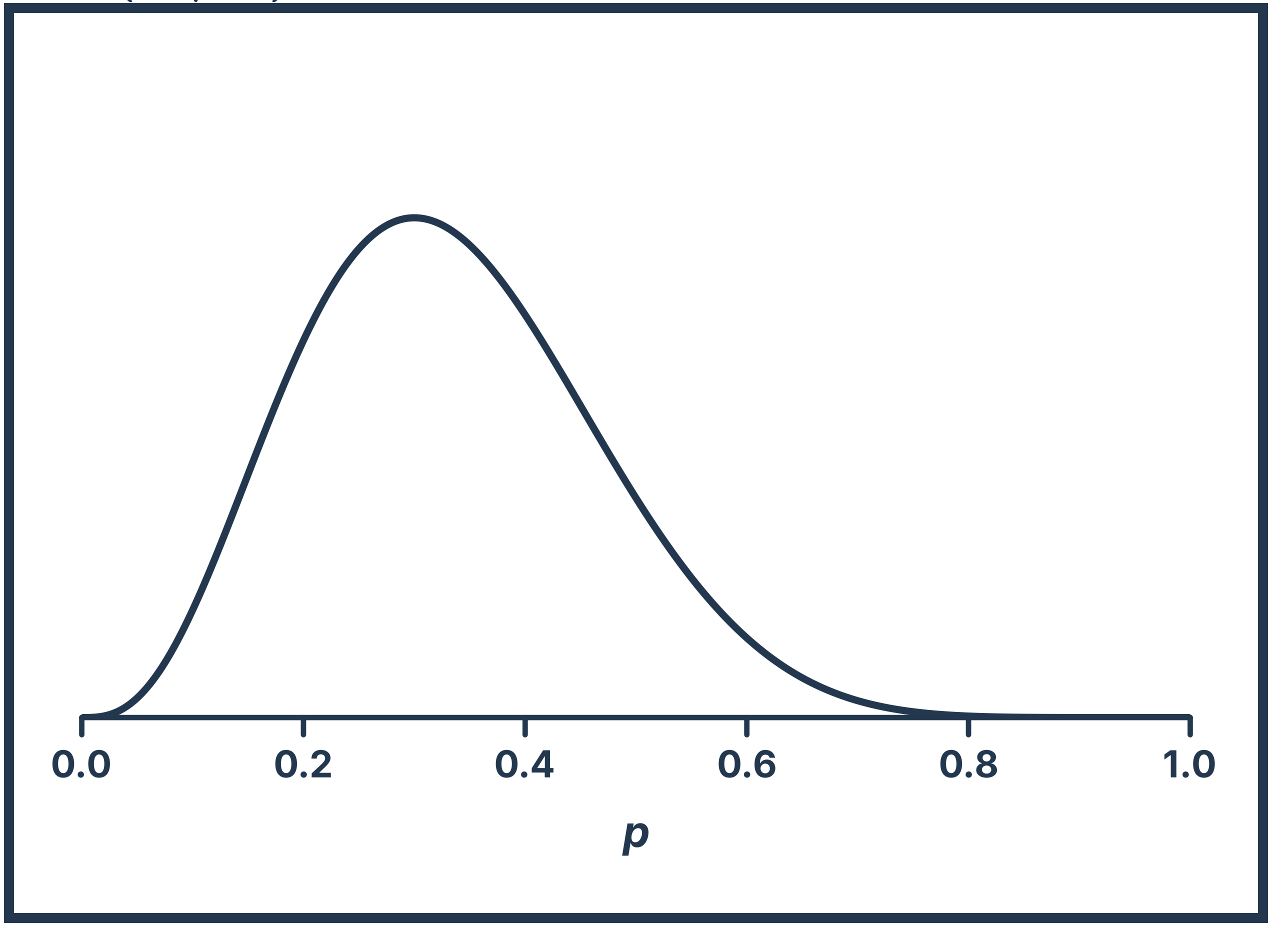
500 samples; uniform prior (click to animate)



500 samples; “informative” prior (click to animate)



$\Pr(p|S)$



The "posterior" is represented as a *conditional probability distribution* (the probability of varying values of p *conditional on* the value of S).

Bayes' rule is a simple formula that allows us to 'flip' a conditional probability

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

And for our unemployment model this becomes

$$\Pr(p|S) = \frac{\Pr(S|p)\Pr(p)}{\Pr(S)}$$

Posterior probability:

The posterior probability is our ***answer***. It tells us everything we know about the probability of unemployment rate (p) given what we've learned from our sample (S).

Posterior probability

$$\boxed{\Pr(p|S)} = \frac{\Pr(S|p)\Pr(p)}{\Pr(S)}$$

Prior probability:
The prior probability is everything we claim to know about the probability of unemployment (p) *before* we ask anybody about their employment. It is the *unconditional* distribution of p.

Prior

$$\Pr(p|S) = \frac{\Pr(S|p)\boxed{\Pr(p)}}{\Pr(S)}$$

Evidence:

The evidence is just the average probability of seeing our sample across all possible values of p (normalizing the posterior). It is often the hardest part of a posterior to calculate.
Fortunately we can almost always ignore it.

$$\Pr(p|S) = \frac{\Pr(S|p)\Pr(p)}{\boxed{\Pr(S)}}$$

↑

Evidence

Likelihood:
The likelihood is where our ***model*** lives.

Likelihood ↓

$$\Pr(p|S) = \frac{\boxed{\Pr(S|p)}\Pr(p)}{\Pr(S)}$$

How to build a parametric model

- *Pretend* that we already know the probability of being unemployed (p)
- Tell a story about what our sample S might look like, assuming we already know p

Reverse the logic of your question

In ***reality*** we know S and want to learn about p:

$$\Pr(p|S)$$

In ***our model*** we know p and want to describe S:

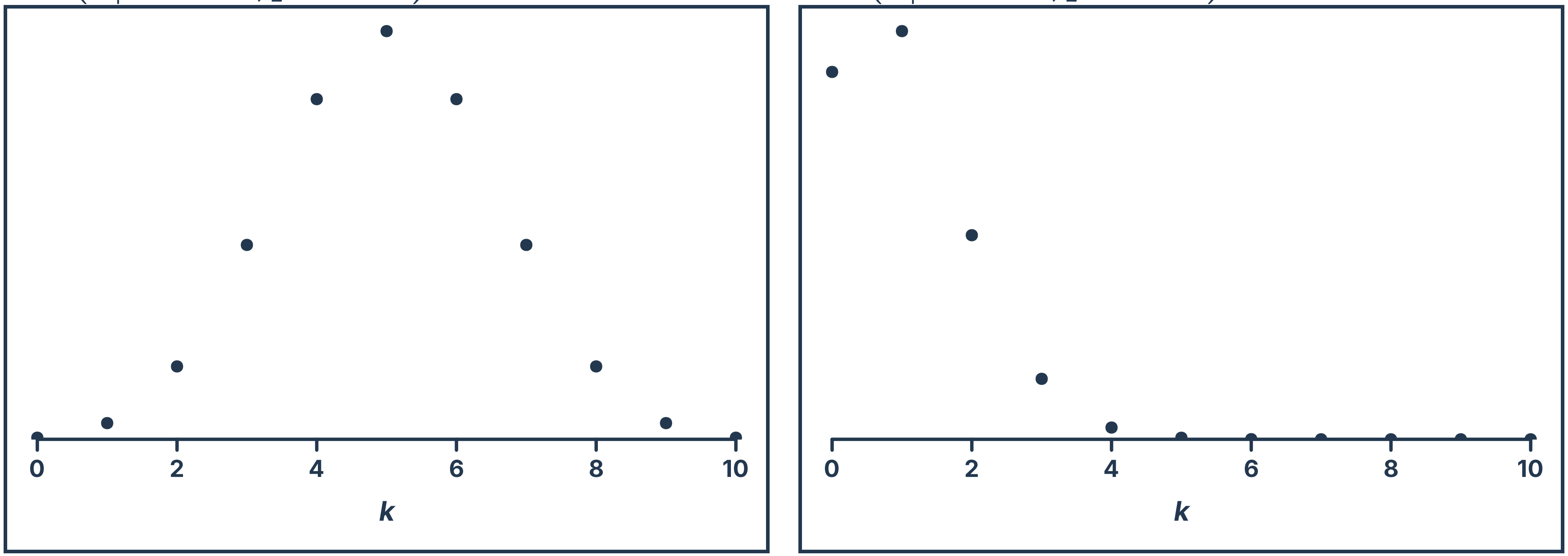
$$\Pr(S|p)$$

Binomial distribution

The probability of getting k 'successes' in n trials if the probability of success is p:

$$\mathrm{Bin}(k|n,p) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

$\mathrm{Bin}(k|n=10, p=0.5)$ $\mathrm{Bin}(k|n=10, p=0.1)$



Likelihood:

The likelihood is where our model lives.
In this case, a ***binomial distribution*** is a good choice. Given a particular probabily of unemployment p (and a sample size n), $Bin(k|n,p)$ tells us how likely our sample is.

Likelihood

$$\Pr(p|S) = \frac{\boxed{\Pr(S|p)}\Pr(p)}{\Pr(S)}$$

BAYES' RULE

Posterior probability

Likelihood

Prior

$$\Pr(p|S) = \frac{\Pr(S|p)\Pr(p)}{\Pr(S)}$$

Evidence

PROPORTIONAL POSTERIOR

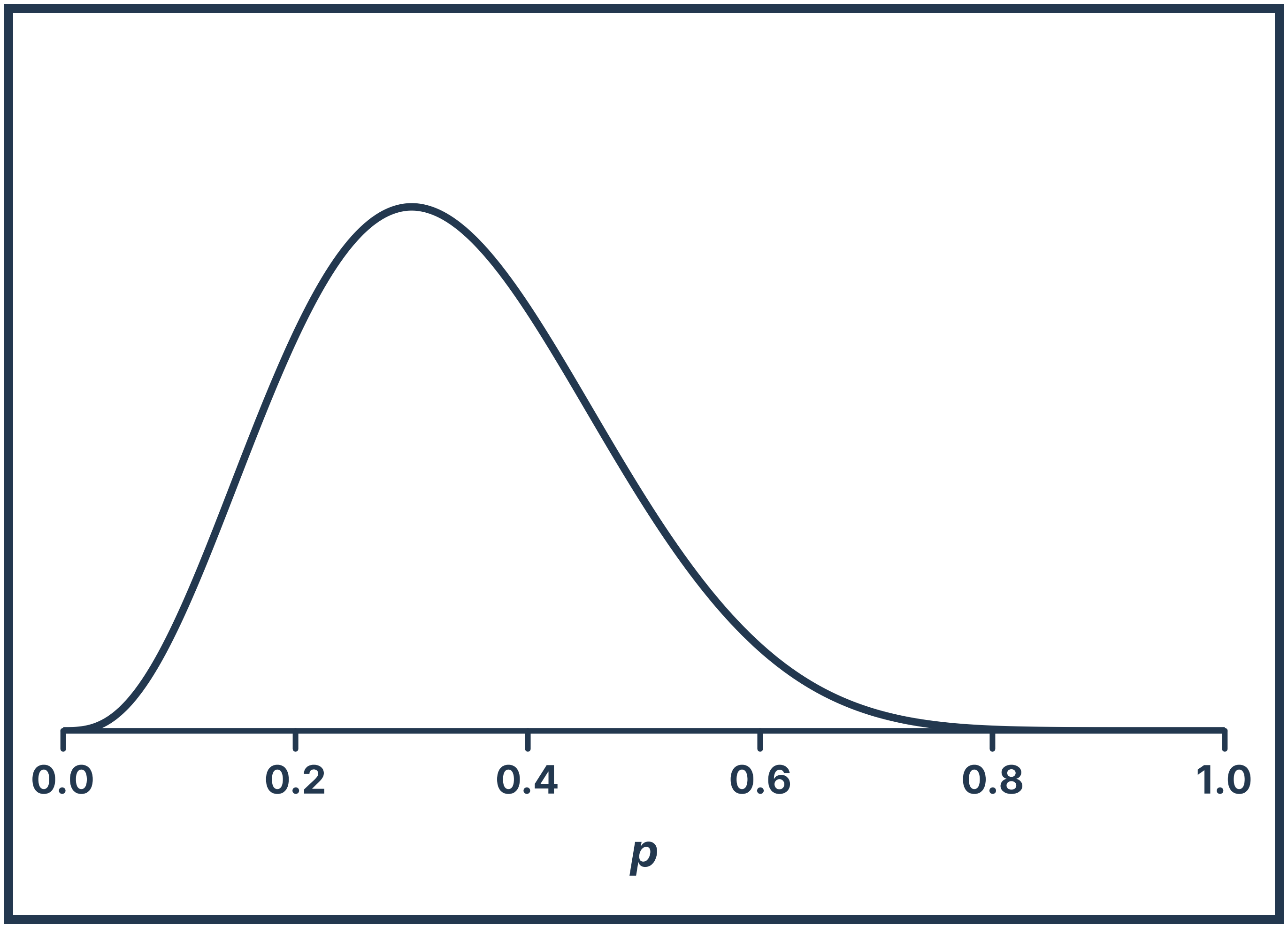
In practice, we rarely need to calculate the “evidence” (the denominator) in Bayes’ formula:

$$\Pr(p|S) \propto \Pr(S|p)\Pr(p)$$

The *posterior probability* is proportional to ($\propto$) the *likelihood* times the *prior*

PROPORTIONAL POSTERIOR

$$\Pr(p|S) \propto \Pr(S|p)\Pr(p)$$



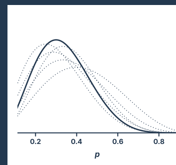
Sample R script

- R scripts are plain-text files containing commands to be interpreted by R
- Example: https://soci620.netlify.app/labs/handson_01_RvsRmarkdown.R

Sample RMarkdown document

- RMarkdown documents are plain-text files that mix regular text with R code.
- RMarkdown files can be 'rendered' to PDF, HTML, or MS Word files that are suitable for distribution.
- Example .Rmd file: https://soci620.netlify.app/labs/handson_01_RvsRmarkdown.Rmd
- Resulting HTML: https://soci620.netlify.app/labs/handson_01_RvsRmarkdown.html

Image credit



Figures by Peter McMahan (source code)

Derrick Mercer, CC BY-SA 2.0, via Wikimedia Commons