Computational Statistics & Probability

Lab 1 - Bayesian Inference

Fall 2025

Learning Objectives

By the end of this lab, you will be able to: - Construct posterior distributions using grid approximation - Query posterior distributions to answer probability questions - Compare models with different priors - Understand how priors influence inference with limited data

The Globe Tossing Experiment

In lecture, we discussed estimating the proportion of Earth's surface covered by water using a simple experiment: toss a globe and record whether your right index finger lands on water (W) or land (L).

This lab walks you through the computational implementation of Bayesian updating for this problem.

1. Understanding Grid Approximation

Use the following R code to generate a set of `samples` from which to answer questions about its distribution.

```r
set.seed(212)

p_grid <-seq( from=0 , to=1, length.out=1000 )
prior <- rep( 1 , 1000 )
likelihood <- dbinom( 6, size=9, prob=p_grid )
posterior <-likelihood * prior
posterior <- posterior / sum(posterior)

samples <- sample( p_grid , prob=posterior , size=1e4 , replace = TRUE)
```

a) How much posterior probability lies below $p = 0.25$?

```r
# ANSWER 1a
sum(samples < 0.25) / length(samples)
```

```
## [1] 0.0025
```

b) How much posterior probability lies above $p = 0.75$?

```r
sum(samples > 0.75) / length(samples)
```

```
## [1] 0.2209
```

c) How much posterior probability lies between $p = 0.25$ and $p = 0.75$?

```r
sum(samples > 0.25 & samples < 0.75) / length(samples)
```

```
## [1] 0.7766
```

d) 25% of the posterior probability lies below which value of p?

```r
quantile(samples, 0.25)
```

```
##       25%
## 0.5435435
```

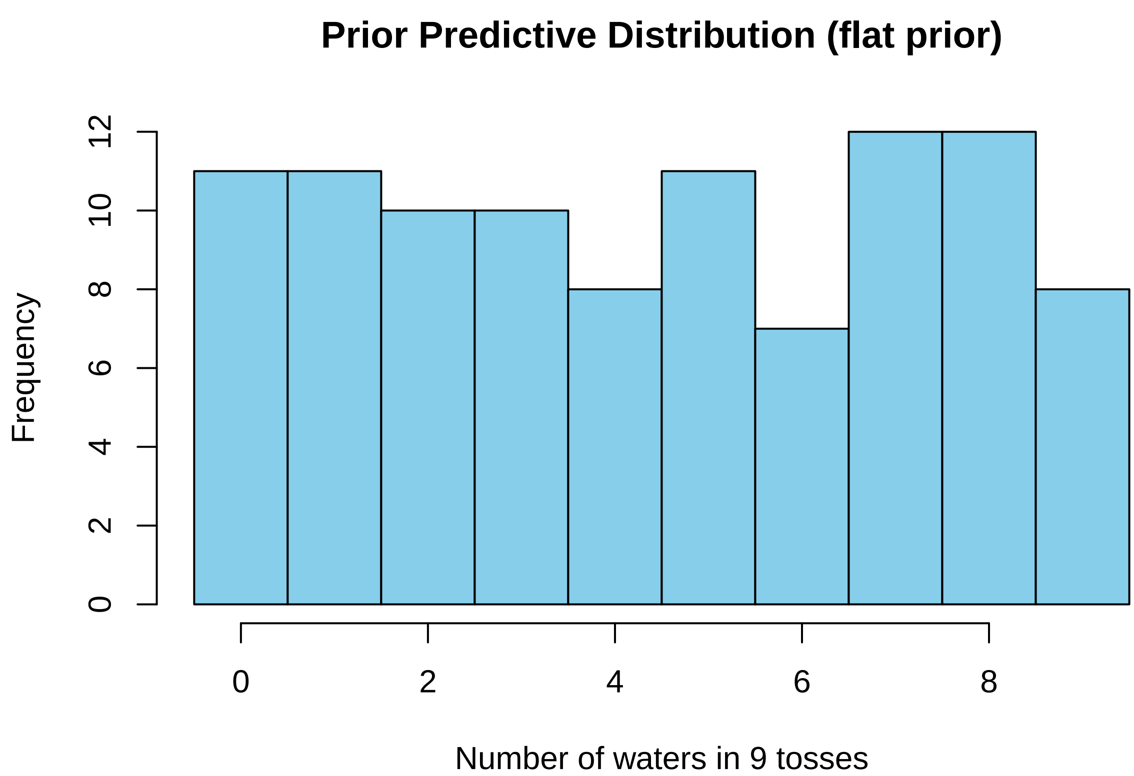
2 Prior Predictive Simulation

Before we fit models with real data, let's check what our priors predict.

a) Simulate predictions from the flat prior

The flat prior says all values of p (0 to 1) are equally plausible. What data should we expect if we truly believe this?

```r
# ANSWER FOR 2a
# Set random seed to 212
set.seed(212)

# Set number of simulations to 100
n_sims <- 100

# Draw 100 samples from [0,1]
p_samples <- runif(n_sims, 0, 1)  # Flat prior: any p equally likely

# For each p, simulate 9 globe tosses
sim_data <- rbinom(n_sims, size=9, prob=p_samples)

# Visualize: What do you expect to observe?
hist(sim_data, breaks=seq(-0.5, 9.5, 1),
     col="skyblue",
     main="Prior Predictive Distribution (flat prior)",
     xlab="Number of waters in 9 tosses",
     ylab="Frequency")
```

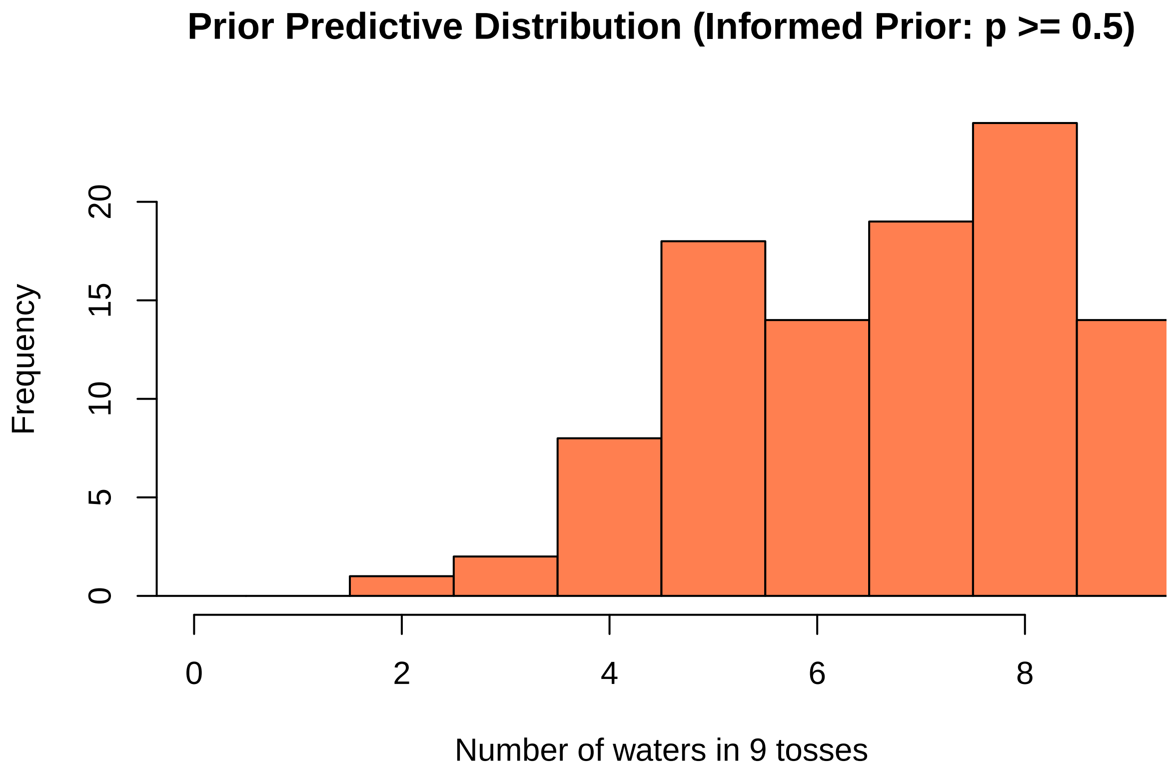



```
# INTERPRETATION: With a flat prior, we expect to see anywhere from 0 to 9
# waters, with all outcomes roughly equally likely. This makes sense in the small
# world where we're saying we have no idea what p might be.
#
# On the other hand, if after seeing this prior simulation the outcome looks
# implausible to you, you might consider a more informed prior.
```

b) Simulate predictions from the informative prior

```
# ANSWER FOR 2b
# Set random seed to 212
set.seed(212)

# Set number of simulations to 100
n_sims <- 100

# Draw 100 samples from [0.5,1]
p_samples_informed <- runif(n_sims, 0.5, 1)  # informed prior: any p greater
                                              # than 1/2

# For each p, simulate 9 globe tosses
sim_data_informed <- rbinom(n_sims, size = 9, prob = p_samples_informed)

# Visualize: What do you expect to observe?
hist(sim_data_informed,
     breaks = seq(-0.5, 9.5, by = 1),
     col = "coral",
     main = "Prior Predictive Distribution (Informed Prior: p >= 0.5)",
     xlab = "Number of waters in 9 tosses",
```

```r
    ylab = "Frequency",
    xlim = c(0, 9))
```




```r
# INTERPRETATION:
# With the informed prior (p >= 0.5), we expect to see MORE waters than with
# the flat prior. The distribution is shifted toward higher values:
# - Most predictions fall between 5-9 waters
# - We rarely expect 0-3 waters because we've ruled out low values of p
# - This makes sense: if we believe p >= 0.5, then in 9 tosses we should
#   see at least 4-5 waters most of the time
#
# Compare this to the flat prior results: the informed prior produces
# more concentrated predictions because we've incorporated prior knowledge
# that constrains which outcomes are plausible.

# Summary statistics
cat("Informed prior predictions:\n")
```

```
## Informed prior predictions:
```

```r
cat("  Mean:", mean(sim_data_informed), "waters\n")
```

```
##   Mean: 6.65 waters
```

```r
cat("  Range:", min(sim_data_informed), "to", max(sim_data_informed), "waters\n")
```

```
##   Range: 2 to 9 waters
```

3 Build Your Own

Suppose the globe tossing experiment yielded the following sequence of 15 observations,

$$[W, L, W, W, L, L, W, L, W, L, L, W, L, W, W]$$

where W denotes 'water' and L denotes 'land'.

Using grid approximation, construct the posterior using: - Grid approximation with 1000 points - A flat prior - The binomial likelihood

a) Write the code (i.e., modify the example from Part 1)

```
# ANSWER
p_grid <-seq( from=0 , to=1, length.out=1000 )
prior <- rep( 1 , 1000 ) # flat prior
likelihood <- dbinom( 8, size=15, prob=p_grid )
posterior <-likelihood * prior
posterior <- posterior / sum(posterior)
set.seed(212)
samples_a <- sample( p_grid , prob=posterior , size=1e4 , replace = TRUE)
```

b) Using grid approximation, construct the posterior distribution with a prior that is 0 below $p = 0.5$ and otherwise constant.

```
# ANSWER
p_grid <-seq( from=0 , to=1, length.out=1000 )
prior <- ifelse(p_grid < 0.5, 0, 1)  # informed prior
likelihood <- dbinom( 8, size=15, prob=p_grid )
posterior <-likelihood * prior
posterior <- posterior / sum(posterior)
set.seed(212)
samples_b <- sample( p_grid , prob=posterior , size=1e4 , replace = TRUE)
```

c) Explain the difference between model (a) and (b).

```
# Model (b) encodes your prior knowledge that at least one-half of the Earth's
# surface is covered with water, whereas Model (a) encodes that you believe any
# proportion, from all water to all land, is equally plausible
```

d) Which prior, (a) or (b), is better? Explain why.

```
# There's no universally "better" prior - it depends on what you know BEFORE
# seeing the data.

# Prior (a) - Flat prior: Use this when you genuinely have no information
# about p. It lets the data dominate the inference. After 15 observations,
# the posterior mean is 0.53.

mean(samples_a)
```

```
## [1] 0.5289651
```

```
# Prior (b) - Informative prior: Use this when you have legitimate prior
# knowledge. If you KNOW that Earth has at least 50% water, this prior
# incorporates that knowledge. The posterior mean is 0.61.

mean(samples_b)
```

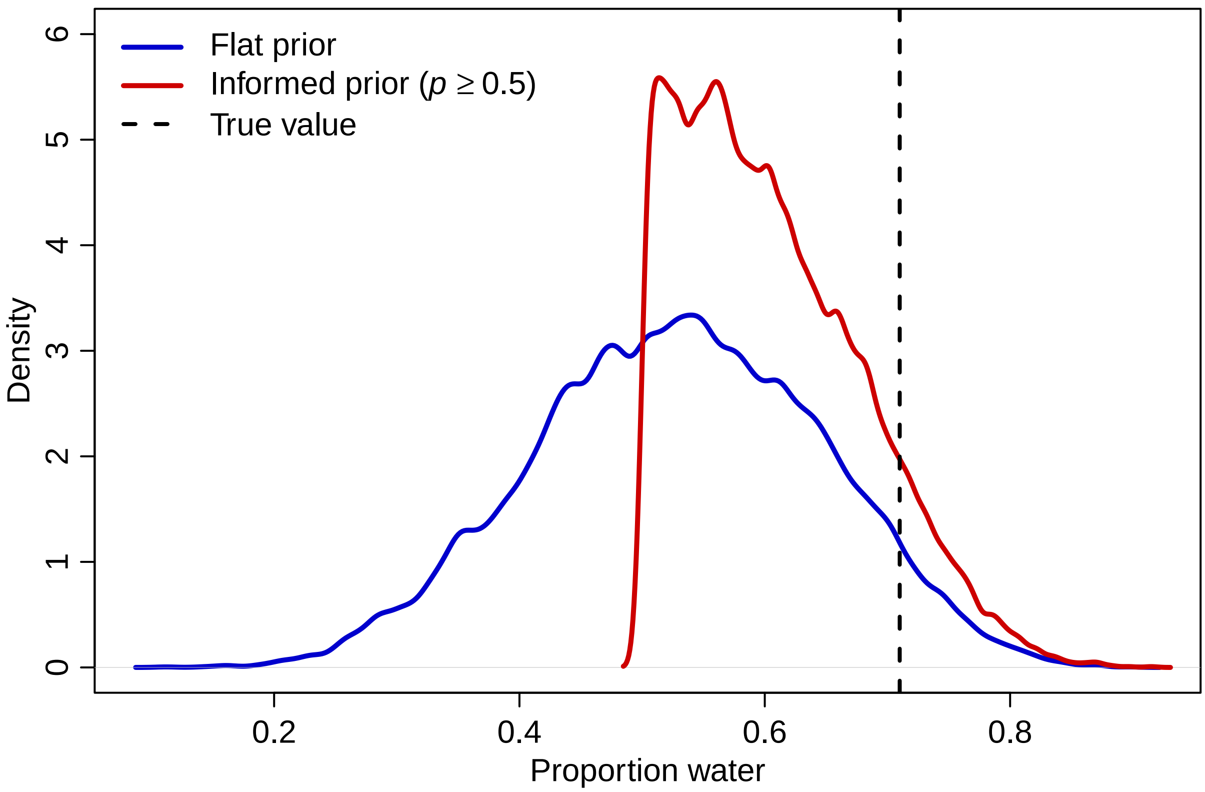
```
## [1] 0.6060088
```

```
# Key insight: With only 15 observations, the prior still matters! Prior (b)
# pulls the estimate toward higher values. With MORE data (say, 150 tosses),
```

```
# both priors would converge to approximately 0.71 (the true value).

# Which is "better"?
# - If you have legitimate prior knowledge → use an informative prior (b)
# - If you want to let data speak → use a flat/weak prior (a)
# - NEVER choose a prior because it gives you the answer you want!

# The flat prior (a) is more conservative and honest when you're truly uncertain.
# The informative prior (b) is appropriate when you have genuine prior knowledge
# (e.g., from previous studies or physical constraints).
```

e) Compare the two posteriors visually

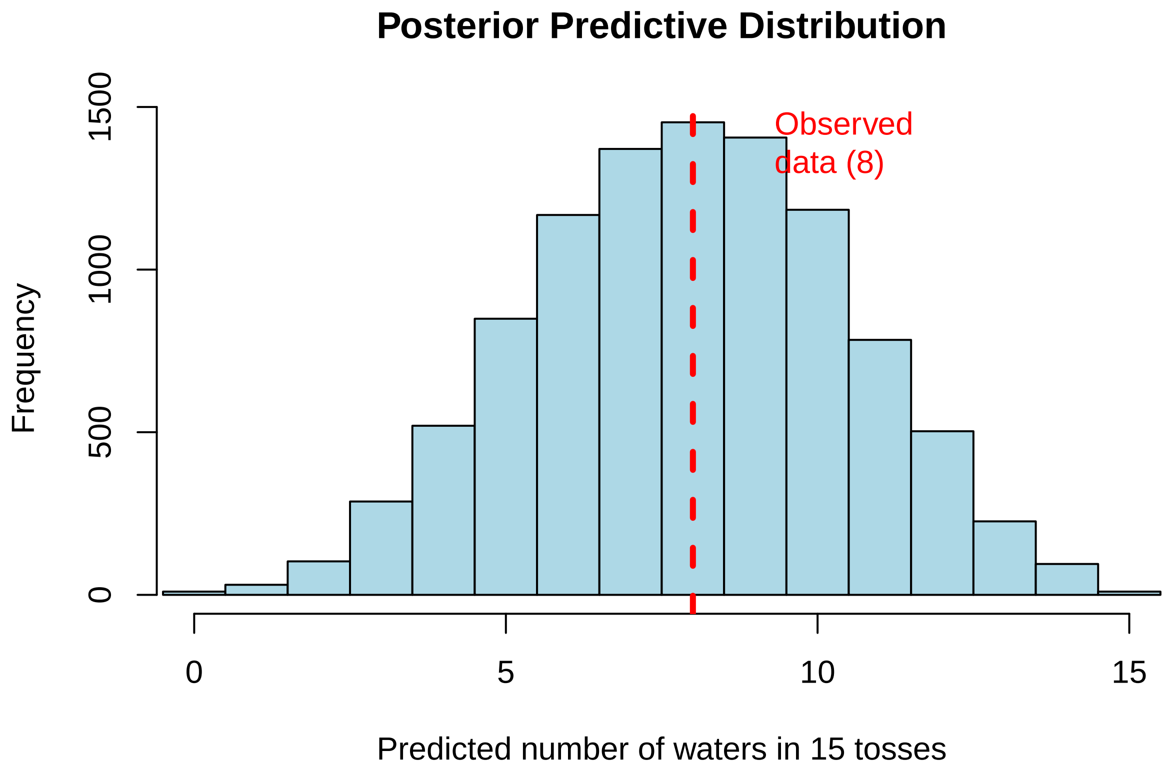
```
dens(samples_a,
     xlab = "Proportion water",
     ylab = "Density",
     ylim = c(0, 6),
     col = "blue3",
     lwd = 2.5)

dens(samples_b, add = TRUE, col = "red3", lwd = 2.5)

abline(v = 0.71, lty = 2, lwd = 2)

legend("topleft",
       legend = c("Flat prior",
                  expression(paste("Informed prior (", italic(p) >= 0.5, ")")),
                  "True value"),
       col = c("blue3", "red3", "black"),
       lty = c(1, 1, 2),
       lwd = c(2.5, 2.5, 2),
       bty = "n")
```



4 Posterior Predictive Check

Now that we've seen the data (8 waters in 15 tosses), does our model make sensible predictions?

```
set.seed(212)

posterior_predictions <- rbinom(n = 1e4, size = 15, prob = samples_a)

# Visualize the posterior predictive distribution
hist(posterior_predictions,
     breaks = seq(-0.5, 15.5, by = 1),
     col = "lightblue",
     main = "Posterior Predictive Distribution",
     xlab = "Predicted number of waters in 15 tosses",
     ylab = "Frequency",
     xlim = c(0, 15))

# Mark the observed data
abline(v = 8, col = "red", lwd = 3, lty = 2)
text(x = 9, y = par("usr")[4] * 0.9,
     labels = "Observed\ndata (8)",
     pos = 4,
     col = "red")
```




```r
# Check: How often does the model predict exactly 8 waters?
prob_8 <- mean(posterior_predictions == 8)
cat("Probability of observing exactly 8 waters:", round(prob_8, 3), "\n")
```

```
## Probability of observing exactly 8 waters: 0.145
```

```r
# Check: Is 8 in a reasonable range?
cat("89% prediction interval:", PI(posterior_predictions, prob = 0.89), "\n")
```

```
## 89% prediction interval: 4 12
```

```r
# What probability does the model assign to the observed outcome?
# mean(posterior_predictions == 8)
## Should be around 0.15-0.20

# INTERPRETATION:
# If our observed value (8) falls within the bulk of the posterior predictive
# distribution, our model is consistent with the data. If it's in the tails,
# we might need to reconsider our model.
```