Section 1

Introduction to Bayesian Statistics

A motivating example

- Student 1 will write down a number and then flip a coin
- **If the flip is heads**, they will honestly tell student 2 if the number is even or odd
- **If the flip is tails**, they will lie
- Student 2 will then guess if the number is odd or even
- Let θ be probability that student 2 correctly guesses whether the number is even or odd

A motivating example

Prior information

Before we start,

1. What's your best guess about θ?

   0.5 0.4

2. What's the probability that θ is greater than a half?

A motivating example

$$\frac{25}{57} = 0.44$$

The class has 25 successes in 57 trials. In light of these data,

posterior

1. What's your best guess about θ?

   0.5 0.44 0.48

2. What's the probability that θ is greater than a half?

   0.5 0.3

Frequentist approach

- A **frequentist procedure** quantifies uncertainty in terms of *repeating the process that generated the data many times*
- The parameters θ are fixed and unknown
- The sample (data) Y is random
- A frequentest would **never** say Prob$(\theta > 0) = 0.60$ because θ is not a random variable
- All probability statements should be made about randomness in the data

Frequentist approach

- A **frequentist** procedure quantifies uncertainty in terms of *repeating the process that generated the data many times*
- For an illustration see `http://www.rossmanchance.com/applets/ConfSim.html`
- A **statistic** $\hat{\theta}$ is a summary of the sample
- For example, the sample mean $\hat{\theta} = \bar{X}$ is a statistic, and it is an **estimator** of the population mean $\theta = \mu$
- The distribution of $\hat{\theta}$ that arises from repeating the process that generated the data many times is its **sampling distribution**
- A frequentist would **never** say “the distribution of μ is Normal(4.2,1.2)”

Frequentist approach

$\bar{X} \pm 2s$

- A **frequentist** procedure quantifies uncertainty in terms of *repeating the process that generated the data many times*
- A **95% confidence interval** (l, u) is an interval constructed from data in a way that the interval will contain θ 95% of the time when

- A frequentist would **never** say "the probability that the true mean is in the interval (3.4, 4.5) is 0.95"

Frequentist approach

- A **frequentist** procedure quantifies uncertainty in terms of *repeating the process that generated the data many times*
- A common approach for testing a hypothesis is to reject the null if a test statistic exceeds a threshold
- For example, we might reject $\mathcal{H}_0 : \mu \leq 0$ in favor of the alternative $\mathcal{H}_1 : \mu > 0$ if $\bar{X} > T$
- A **p-value** is the prob of observing a test stat more extreme than I obsered if $\mathcal{H}_0$ is trun

  μ is not random
- A frequentist would **never** say "the probability that the null hypothesis is true is 0.03"

Frequentist approach

There is currently an intense discussion of the merits of the p-value in the scientific community:

- http://www.nature.com/news/scientific-method-statistical-errors-1.14700
- http://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/
- http://fivethirtyeight.com/features/science-isnt-broken/
- http://www.tandfonline.com/doi/pdf/10.1080/01973533.2015.1012991

How about a frequentist answer these questions?

Before we start:

1. What's your best guess about θ? I don't know

   ~~0.5~~

2. What's the probability that θ is greater than a half?

   This is a non-sense question

After we have observed some trials:

1. What's your best guess about θ now?

   $\hat{\theta} = 0.44$

2. What's the probability that θ is greater than a half now?

The Bayesian approach

- Bayesians also view θ as fixed and unknown
- However, we express our uncertainty about θ using probability distributions
- The distribution before observing the data is the **prior distribution**
- Example: Prob($\theta >$ 0.5) $=$ 0.6.
- Probability statements like this are intuitive (to me at least)
- This is subjective in that people may have different priors (there is also a field called objective Bayes)

The Bayesian approach

- Our uncertainty about θ is changed (hopefully reduced) after observing the data
- The **Likelihood function** is the distribution of the observed data given the parameters
- This is the same likelihood function used in a maximum likelihood analysis
- Therefore, when the prior information is weak a Bayesian and maximum likelihood analysis are similar

The Bayesian approach

- The uncertainty distribution of θ after observing the data is the **posterior distribution**
- **Bayes theorem** provides the rule for updating the prior

$$f(\theta|Y) = \frac{f(Y|\theta)f(\theta)}{f(Y)}$$

- In words: Posterior $\propto$ Likelihood·prior
- A key difference between Bayesian and frequentist statistics is that all inference is conditional on the single data set we observed Y

Back to the example

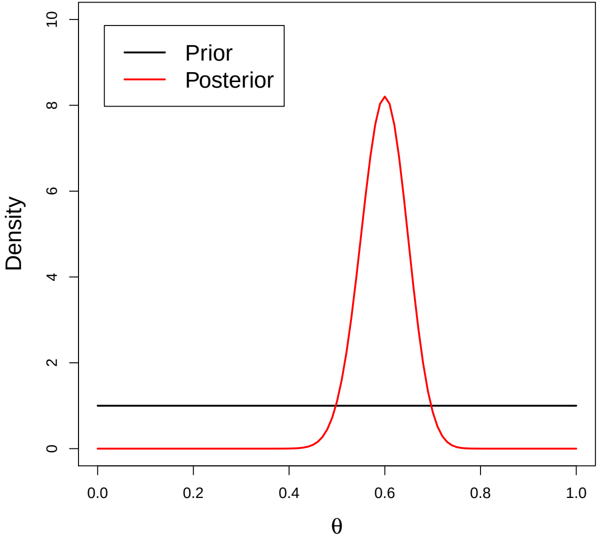
- Say we observed $Y = 60$ successes in $n = 100$ trials
- The parameter $\theta \in [0, 1]$ is the true probability of success
- In most cases we would select a prior that puts probability on all values between 0 and 1
- If we have no relevant prior information we might use the prior
$$\theta \sim \text{Uniform}(0, 1)$$
so that all values between 0 and 1 are equally likely
- This is an example of an **uninformative prior**

Posterior distribution

- The likelihood is $Y|\theta \sim \text{Binomial}(n, \theta)$
- The uniform prior is $\theta \sim \text{Uniform}(0, 1)$
- Then it turns out the posterior is

$$\theta|Y \sim \text{Beta}(Y + 1, n - Y + 1)$$

Bayesian learning: $Y = 60$ and $n = 100$

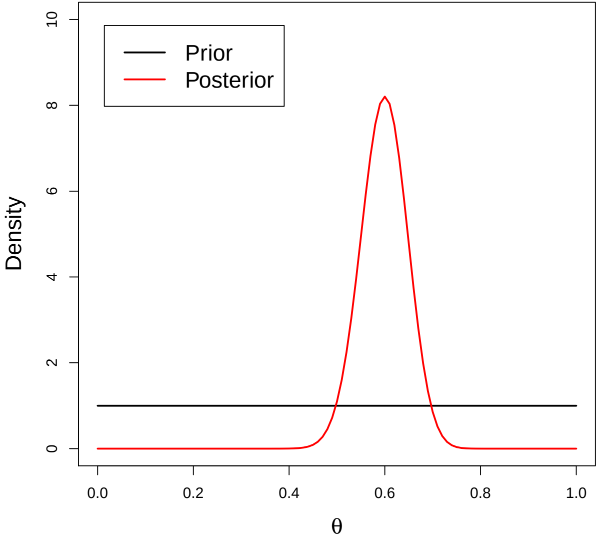


Beta prior

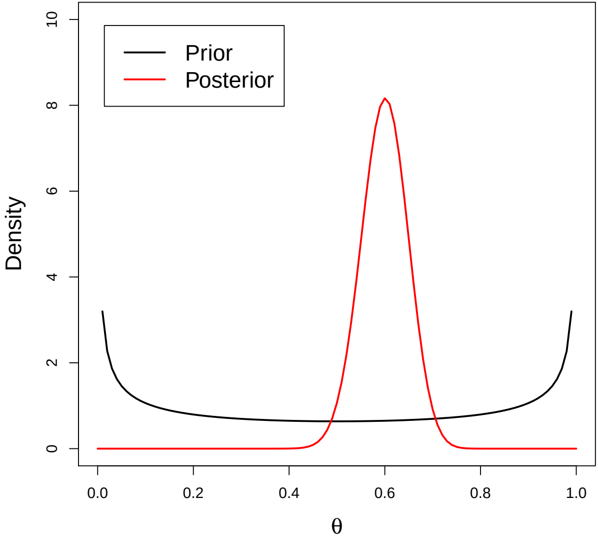
- The uniform prior represents prior ignorance
- To encode prior information we need a more general prior
- The beta distribution is a common prior for a parameter that is bounded between 0 and 1
- If $\theta \sim \text{Beta}(a, b)$ then the posterior is

$$\theta | Y \sim \text{Beta}(Y + a, n - Y + b)$$

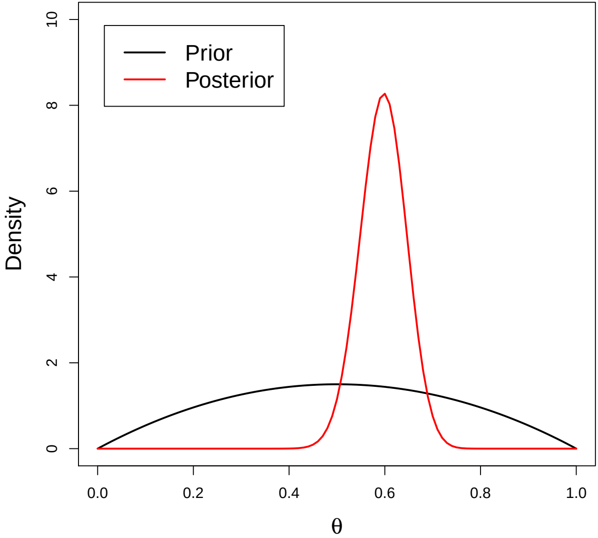
Prior 1: $\theta \sim \text{Beta}(1, 1)$



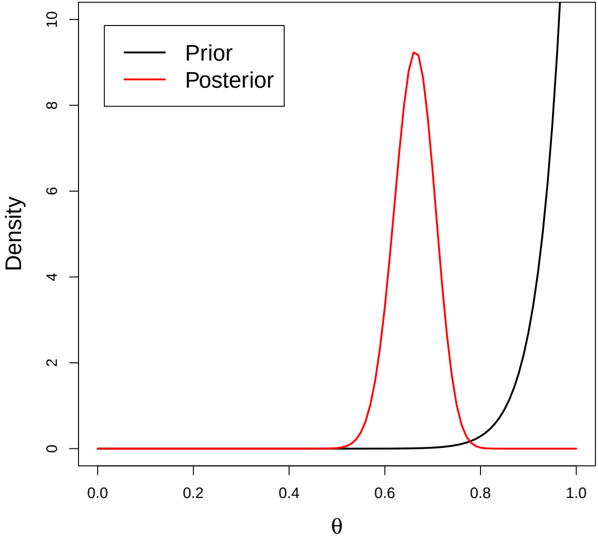
Prior 2: $\theta \sim \text{Beta}(0.5, 0.5)$



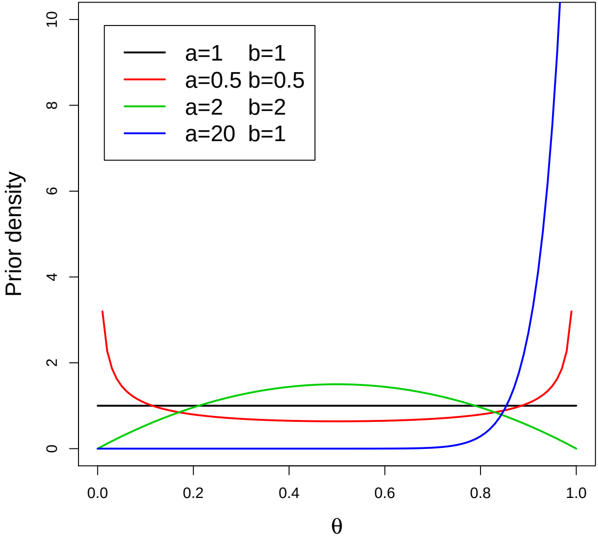
Prior 3: $\theta \sim \text{Beta}(2, 2)$



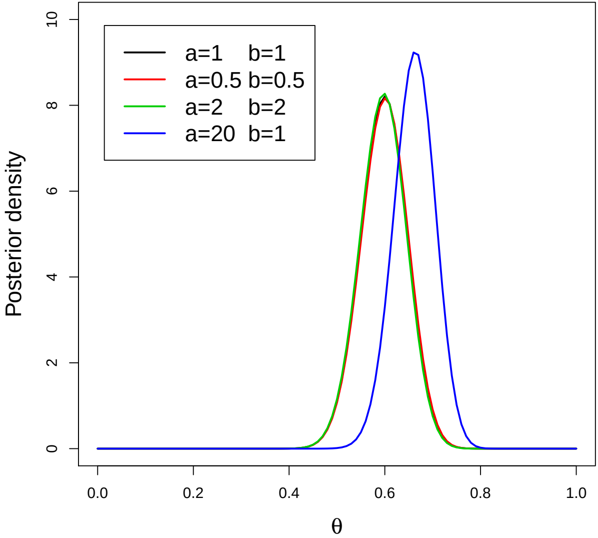
Prior 4: $\theta \sim \text{Beta}(20, 1)$



Plot of different beta priors



Plots of the corresponding posteriors



Senstivity to the prior

		Prior			Posterior		
a	b	Mean	SD	P>0.5	Mean	SD	P>0.5
1	1	0.50	0.29	0.50	0.60	0.05	0.98
0.5	0.5	0.50	0.50	0.50	0.60	0.05	0.98
2	2	0.50	0.22	0.50	0.60	0.05	0.98
20	1	0.95	0.05	1.00	0.66	0.04	1.00

Summary

- The first three priors give essentially the same results
- Say the objective is to test $\mathcal{H}_o : \theta \leq 0.5$ versus $\mathcal{H}_A : \theta > 0.5$
- In these three cases we can say that after observing the data the probability of the null is only 0.02 and the alternative is 50 times more likely than the null
- The final prior strongly favored large θ and gave different results
- How would we argue this analysis is useful?

Advantages of the Bayesian approach

- Bayesian concepts (posterior prob of the null) are arguably easier to interpret than frequentist ideas (p-value)
- We can incorporate scientific knowledge via the prior
- Excellent at quantifying uncertainty in complex problems (e.g., missing data, correlation, etc.)
- In some cases the computing is easier
- Provides a framework to incorporate data/information from multiple sources

Disadvantages of Bayesian methods

Go Bayes!

- Picking a prior is subjective (we'll study objective priors)
- Procedures with frequentist properties are desirable (we'll study the frequentist properties of Bayesian methods)
- Computing can be slow or unstable for hard problems
- Less common/familiar
- Nonparametric methods are challenging