STAT 491 - Lecture 5

February 1, 2018

Ch.5 Bayes Rule

Compare the two probability statements:

- P[1 foot (or more) of powder at Bridger]
- P[1 foot (or more) of powder at Bridger | blue light on Baxter Hotel]

The first probability statement consider two possible outcomes: 1 foot (or more) of new snow and less than 1 foot of new snow. The second probability statement incorporates some additional information (data) into the probability statement, namely that 2 or more inches of snow has fallen.

Bayes rule is the mathematical foundation for re-allocating credibility (or probability) when conditioning on data.

Bayes Rule and Conditional Probability

- Recall: the conditional probability $P(A|B) = \frac{P(A \cap B)}{P(B)}$. From here we do some algebra to obtain Bayes rule.

$$
\begin{aligned}
P(A|B) \times P(B) &= \frac{P(A \cap B)}{P(B)} \times P(B) && (1)\\
P(A|B)p(B) &= P(B|A)P(A) && (2)\\
P(A|B) &= \frac{P(B|A)P(A)}{P(B)} && (3)\\
P(A|B) &= \frac{P(B|A)P(A)}{\sum_A^{'} P(B|A')P(A')} && (4)
\end{aligned}
$$

- Either of the last two equations are called **Bayes Rule**, named after Thomas Bayes.

Bayes Rule with two-way discrete table

	COLUMN	
ROW	c	**Marginal**
r	p(r,c) = p(r|c)p(c) = p(c|r)p(r)	p(r)
Marginal	p(c)	

The conditional probability $p(c|r)$ is the joint probability $p(r, c)$ in the cell divided by the marginal probability $p(r)$

Recall the following two-way table:

	Hair Color				
Eye Color	Black	Brunette	Red	Blond	Marginal (eye color)
Brown	0.11	0.20	0.04	0.01	0.37
Blue	0.03	0.14	0.03	0.16	0.36
Hazel	0.03	0.09	0.02	0.02	0.16
Green	0.01	0.05	0.02	0.03	0.11
Marginal (hair color)	0.18	0.48	0.12	0.21	1.0

Previously we calculated:

- What is the probability of a person having red hair given that they have blue eyes

  We now see that this is a simple illustration of Bayes rule.

A classic example of Bayes rule focuses on diagnosing a rare disease. There are a few important values we need to state:

- Let θ be a parameter that determines whether a person has the disease,
- Let T be the result of the diagnostic test,
- Let $Pr(\theta = Yes) = p_\theta$ be the probability a person from the general population has the disease,
- Let $Pr(Test = Yes|\theta = Yes) = p_{T+}$ be the probability of detecting a disease when it is present. This is called the hit rate in the textbook,
- Let $Pr(Test = Yes|\theta = No) = p_{T-}$ be the probability of a false alarm.
- **Question:** do we need to state,
  - $Pr(\theta = No)$
  - $Pr(Test = No|\theta = Yes)$
  - $Pr(Test = No|\theta = No)$

Assume we are testing citizens for Extra Sensory Perception (ESP). The ultimate goal will be to determine the probability that an individual has ESP if they test positive for ESP. Mathematically this is stated as $Pr(\theta = Yes|Test = Yes)$.

First using the generic probability from the previous page, compute $Pr(\theta = Yes|Test = Yes)$.

$$\begin{aligned} Pr(\theta = Yes|Test = Yes) &= \frac{Pr(Test = Yes|\theta = Yes)Pr(\theta = Yes)}{\sum_{\theta}^{'} Pr(Test = Yes|\theta = \theta')Pr(\theta = \theta')} \\ &= \frac{Pr(Test = Yes|\theta = Yes)Pr(\theta = Yes)}{Pr(Test = Yes|\theta = Yes)Pr(\theta = Yes) + Pr(Test = Yes|\theta = No)Pr(\theta = No)} \\ &= \frac{p_{T+} \times p_{\theta}}{p_{T+} \times p_{\theta} + p_{T-} \times (1 - p_{\theta})} \end{aligned}$$

Now to make this concrete assume:

- The rate of ESP in the population is 1 in 100,000, so $Pr(\theta = Yes) = p_{\theta} = 0.00001$
- The hit rate of the test is 9999 in 10000 %, so $Pr(Test = Yes|\theta = Yes) = p_{T+} = .9999$
- The false detection rate is 1 in 10000 %, so $Pr(Test = Yes|\theta = No) = p_{T-} = 0.0001$
- **Question:** Before doing any math, what is your guess for the probability that a person receiving a positive test actually has ESP?

```
p.theta <- 1 / 100000
p.t.plus <- 9999 / 10000
p.t.minus <- 1 / 10000
p.theta.true <- p.t.plus * p.theta / (p.t.plus * p.theta + p.t.minus * (1 - p.theta))
```

It turns out that the probability that a person actually has ESP given they had a positive test is $Pr(\theta = Yes|Test = Yes) = 0.0909$.

This example allows us to understand the mechanisms behing Bayes rule.

- ESP is rare (if you believe in that kind of thing), so there is low prior probability of a person having this ability
- The test provides some additional information to support the person having ESP - that is there is a shift in the plausibility of this outcome, but the combination of the rareness of ESP and the false detection rate still make it more likely that the person does not have ESP.

Bayes rule with parameters and data

The previous example was essentially a probability exercise and we were not doing Bayesian statistical analysis per se, but rather just using Bayes rule. Bayesian statistical analysis refers to a fairly specific application of this theorem where:

- there is a statistical model of observed data, conditional on parameter values, $p(\mathcal{D}|\theta)$. We will use $\mathcal{D}$ to represent data and θ to represent the parameters. (Note the statistical model of observed data exists in a frequentist paradigm as well, as this function is sometimes called a *statistical likelihood* which is used for Maximum Likelihood Estimation (MLE).)

- there exists a **prior** belief on the parameter values, $p(\theta)$, and

- Bayes rule is used to convert the prior belief on the parameters *and* the statistical model into a **posterior** belief $p(\theta|\mathcal{D})$.

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})}$$

The denominator $p(\mathcal{D})$ is referred to as the marginal likelihood of the data and is computed as:

- $p(\mathcal{D}) = \sum_{\theta'} p(\mathcal{D}|\theta^{'})p(\theta^{'})$ in the case where the parameters are discrete and $p(\mathcal{D}) = \int p(\mathcal{D}|\theta^{'})p(\theta^{'})d\theta^{'}$ in the case where the parameters are continuous. The $\theta^{'}$ is used as we enumerate or integrate across all possible values of the parameter values.

Example of Bayesian Analysis on a binary outcome

Consider estimating the probability that a die will roll a six and recall the 5 steps in a Bayesian analysis

1. Identify the data relevant to the research question.
2. Define a descriptive model for the relevant data.
3. Specify a prior distribution on the parameters.
4. Use Bayesian inference to re-allocate credibility across parameter values.
5. Check that the posterior predictions mimic the data with reasonable accuracy.

1. Identify the data relevant to the research question.

- What data do we need to determine of the probability that a die lands on a six?

  This amounts to rolling the die a large number of times and recording whether each roll ended up as a six or not.

2. Define a descriptive model for the relevant data.

A descriptive model denoted as $p(\mathcal{D}|\theta)$ is needed for the die rolling experiment.

- what is $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}$ the data for binary outcomes of rolling the die

- what is θ: the probability of rolling a six

- what is a descriptive model for $p(\mathcal{D}|\theta)$, For a single roll of the die, $d_i = 1$ if a 6 is rolled and $d_i = 0$ otherwise, use the Bernoulli distribution:

$$p(d_i = 1|\theta) = \theta^{d_i}(1-\theta)^{1-d_i}$$

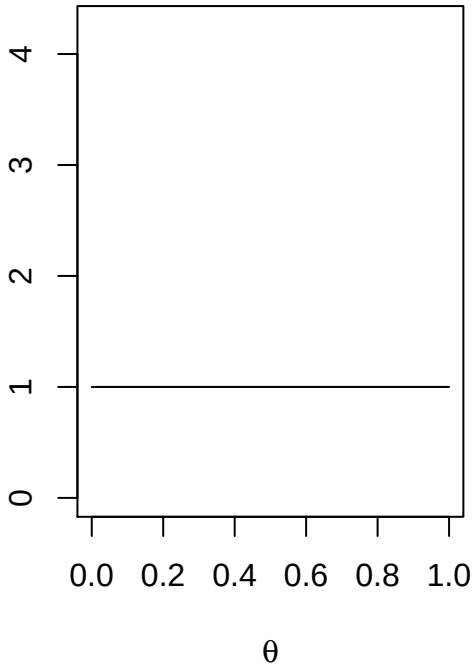
  By assuming that each die roll is independent, they can be combined as:

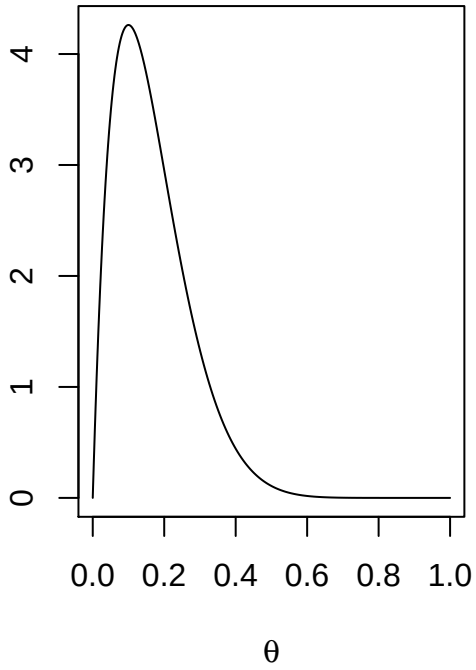
$$\begin{aligned} p(\mathcal{D}|\theta) &= \prod_i p(d_i|\theta) \\ &= \prod_i \theta^{d_i}(1-\theta)^{1-d_i} \\ &= \theta^{\sum_i d_i}(1-\theta)^{\sum_i (1-d_i)} \\ &= \theta^{\#\text{ of heads}}(1-\theta)^{\#\text{ of tails}} \end{aligned}$$

This model is related to a binomial distribution and will be the mathematical machinery that allows updated prior beliefs in a formulaic manner.

3. Specify a prior distribution on the parameters

Here are a couple of reasonable prior distributions on θ, the probability of rolling a 6.





Discuss the implications behind each figure.

It turns out that distribution for each figure can be formulated in terms of a Beta distribution:

$$p(\theta) = \theta^{\alpha-1}(1-\theta)^{\beta-1}\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)},$$

for $\theta \in [0, 1]$ and 0 otherwise.

- The first figure has: $\alpha = 1$ and $\beta = 1$, which results in a uniform distribution.

- The second figure has: $\alpha = 2$ and $\beta = 10$.

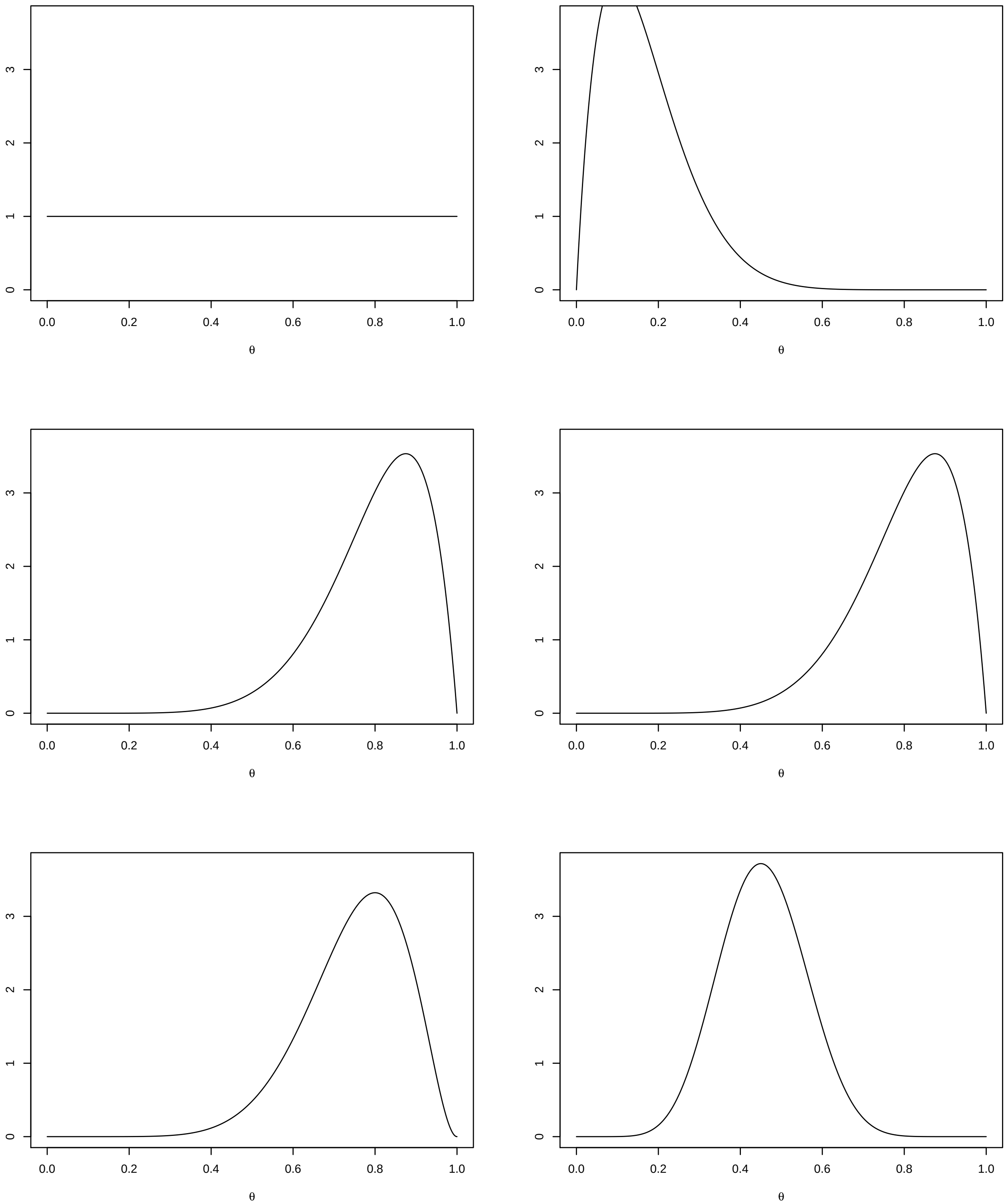
4. Use Bayesian inference to re-allocate credibility across parameter values.

Recall the goal of this analysis was to learn about θ the probability of rolling a six. Specifically, we are interested in the posterior distribution $p(\theta|\mathcal{D})$.

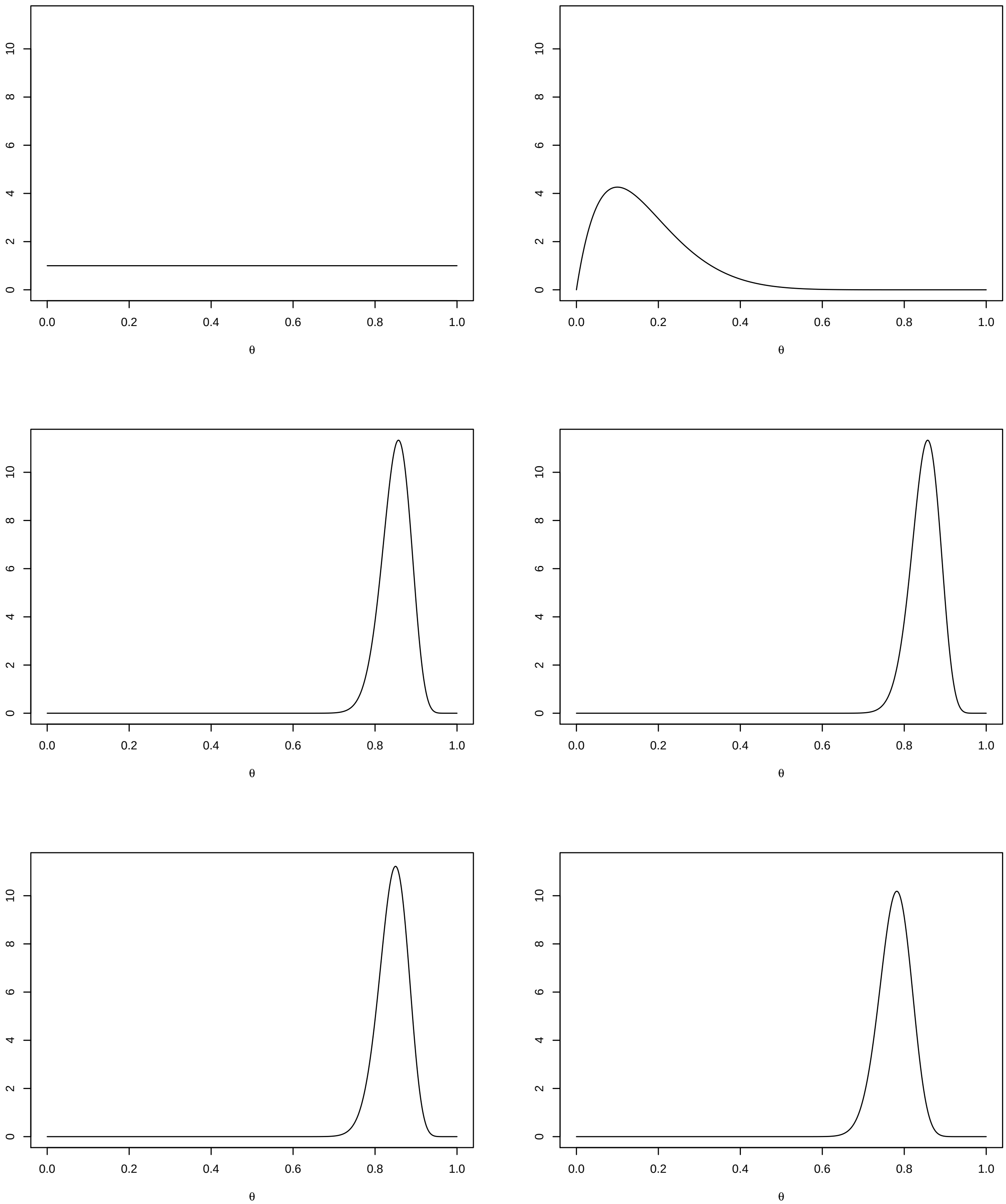
Lets assume a few data collection procedures:

1. 10 rolls of the die, with 8 6's
2. 25 rolls of the die, with 20 6's
3. 100 rolls of the die, with 85 6's

With 10 rolls



Then with 100 rolls



influence of the sample size and prior on posterior

With Bayesian statistics there is an interplay between the strength of our prior beliefs and the amount of data collected. - The posterior distribution can be considered as a weighted average between the prior distribution and the data.

- If the prior is strong relative to the amount of data collected, the posterior will be largely influenced by the prior distribution.

- If the sample size is large relative to the strength of the prior distribution, the posterior will be largely influenced by the data.