Introduction to Bayesian Thinking

Professor Karsten T. Hansen

UC San Diego, Rady School of Management

MGTA 495, Spring 2022

Example: Estimating User Interest in a new App

- Suppose we are interested in estimating the interest among users in using a newly developed subscription smart phone app
- Let the true adoption rate in the market segment of interest be λ.
- Let $Y_i = 1$ if user i is interested in adoption. Then

$$\Pr(Y_i = 1|\lambda) = \lambda.$$

- Suppose we survey $n = 100$ users and ask them about adoption.
- We wish to learn what plausible values of λ might be

Interest in size of λ

- Suppose development cost for the app was C_D
- Furthermore suppose that the monthly fixed cost of maintaining the app is C_M
- Assume the monthly subscription fee is π.
- Finally assume that the target market size is M.

In this case the monthly app profit is

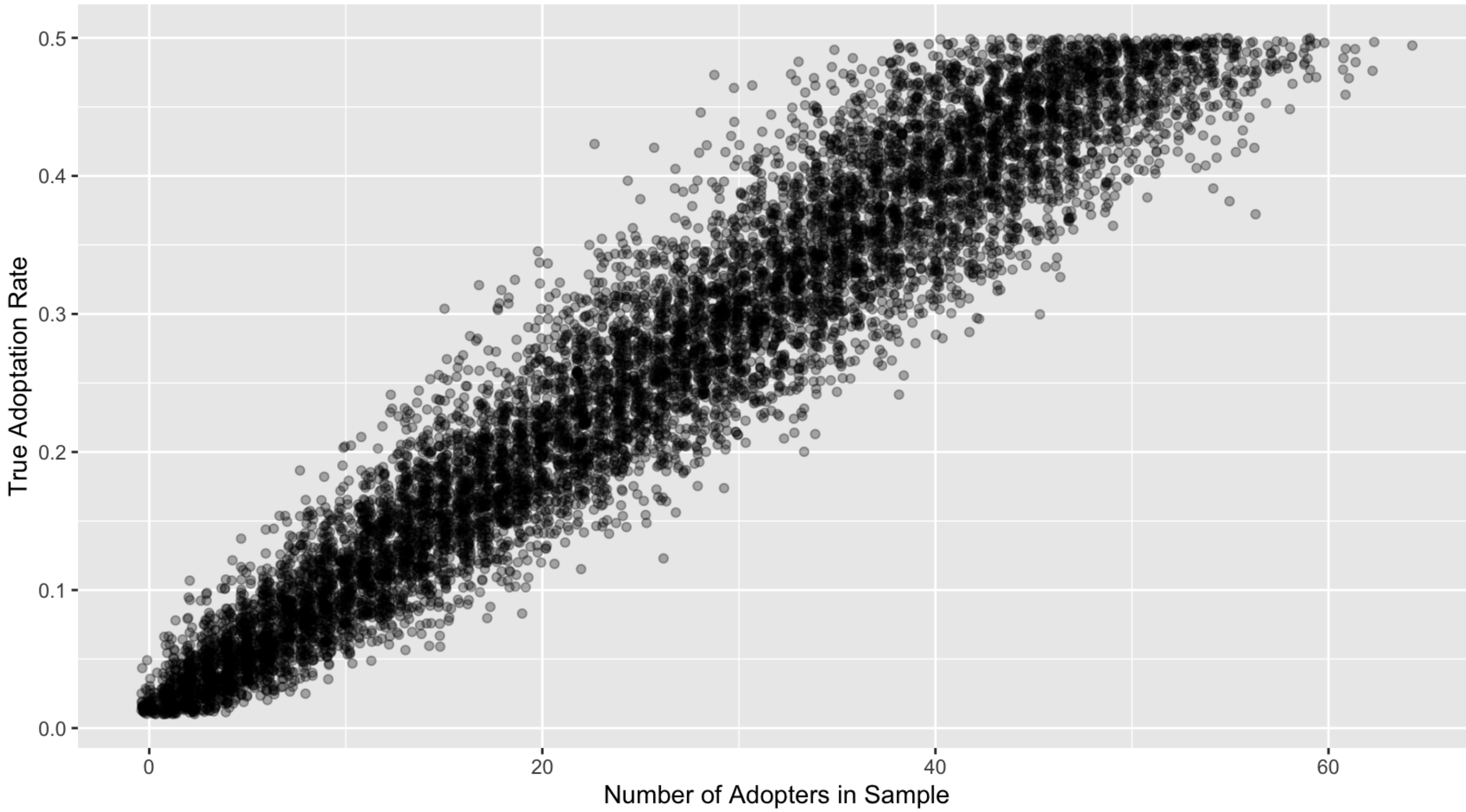
$$\text{profit} = \pi \times \lambda \times M - C_M$$

Suppose we decide to launch the app if we can recoup the development cost in 12 months:

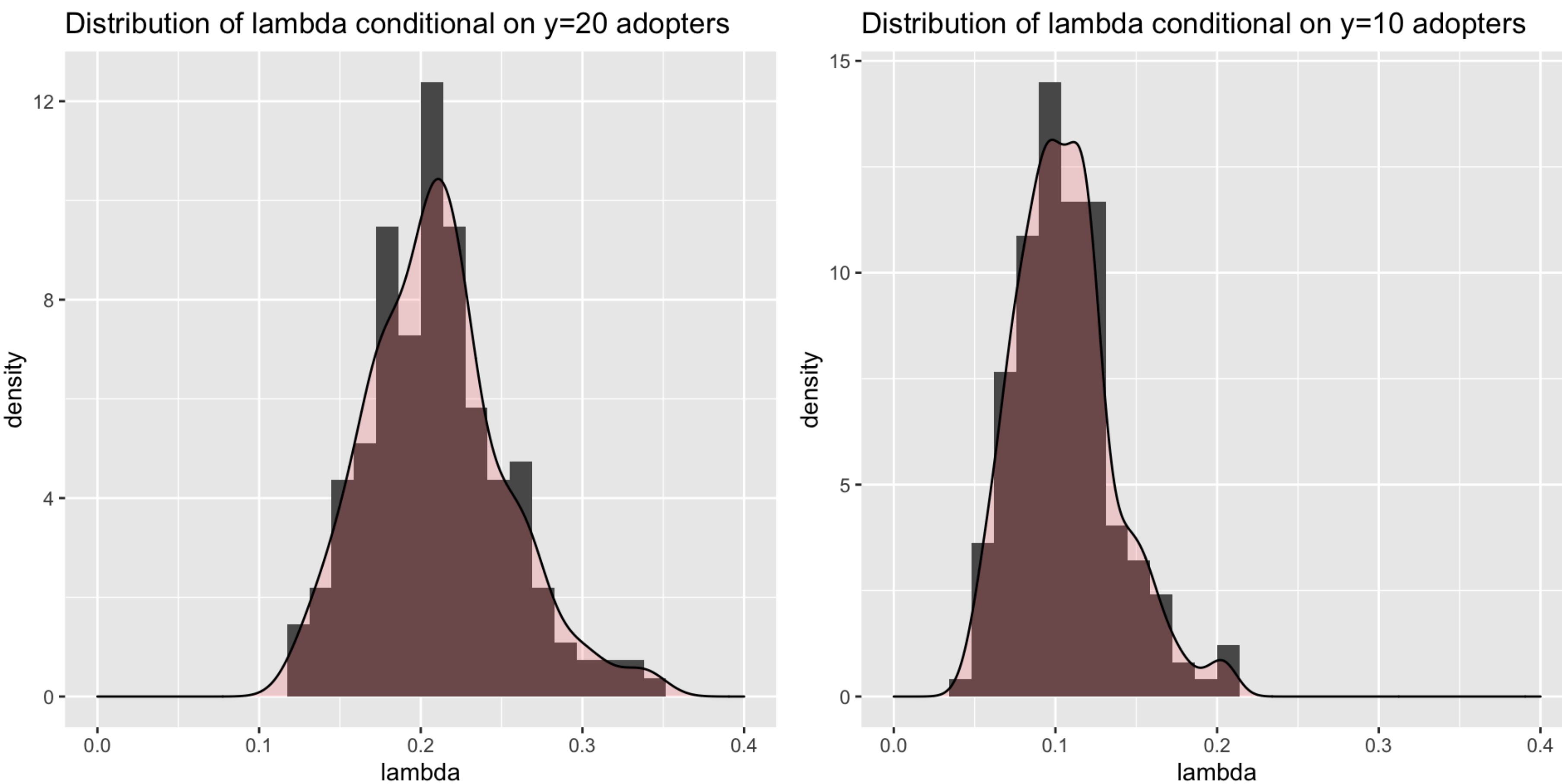
$$12 \times \text{profit} > C_D \iff \lambda > \underline{\lambda} \equiv \frac{\frac{C_D}{12} + C_M}{\pi \times M}$$

$\implies$ We need $\Pr(\lambda > \underline{\lambda})$.

- Consider the following question: For different reasonable values of λ, what is the range of adopters we would expect to see in a sample of size $n = 100$?
- We can easily simulate this.
- Suppose we believe the following: λ is probably bigger than 1 pct. and probably less than 50 pct. In between 0.01 and 0.50 we believe that any value is as likely as any other.
- We can represent this belief as a uniform distribution on $[0.01, 0.50]$
- We can then do the following many times:
  - Draw a random λ from $[0.01, 0.50]$
  - Simulate unemployment status for 100 hypothetical graduates given λ


0.5
0.4
0.3
0.2
0.1
0.0
True Adoptation Rate
0
20
40
60
Number of Adopters in Sample

Distribution of λ conditional on data



Insights

- If we observe 20 adopters in a sample of 100 potential users, then plausible values of λ are between 0.1 and 0.35 with the most likely values around 0.2
- If we observe 10 adopters in a sample of 100 potential users, then plausible values of λ are between 0.02 and 0.2 with the most likely values around 0.1
- Note that you can make probability statements about λ with this approach. For example, we can ask: what is the probability that λ is between 0.15 and 0.25?

Decision

- Suppose $\underline{\lambda} = 0.3$.
- Before observing any data, we have

$$\Pr(\lambda > \underline{\lambda}) = \frac{0.5 - 0.3}{0.5 - 0.01} \approx 41\%$$

- Suppose we observe 20 adopters in the sample - what is $\Pr(\lambda > \underline{\lambda})$ after learning this information?
- We can approximate this probability by looking at the fraction of times $\lambda > 0.3$ in all the simulated samples where $y = 20$. This is

$$\Pr(\lambda > \underline{\lambda}|\text{data}) \approx \frac{\#\{\lambda > 0.3|y = 20\}}{\#\{y = 20\}} = 0.035$$

Classical Approach: Distribution of Estimator Conditional on Fixed Parameter

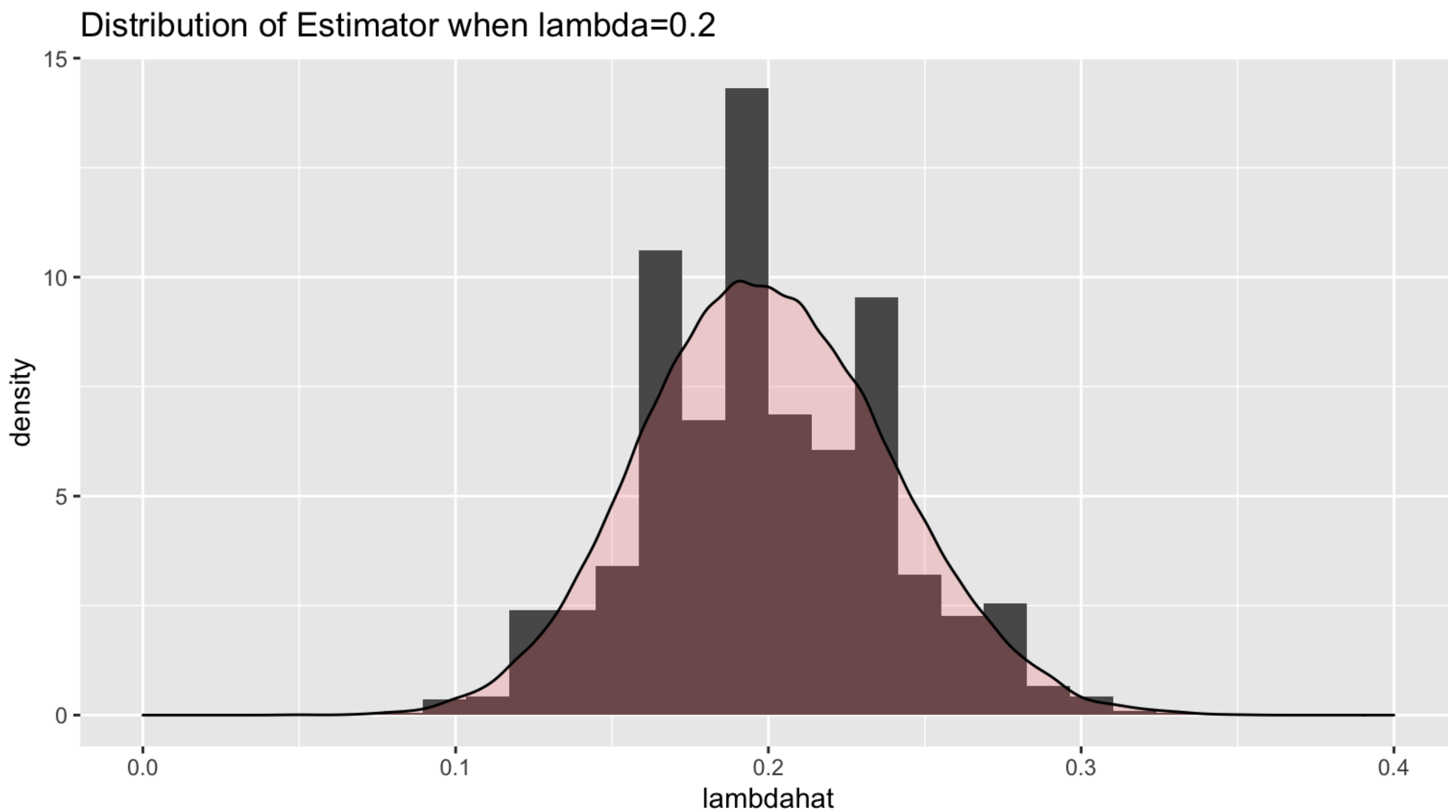
- In the classical approach we start by proposing an **estimator** $\hat{\lambda}$ of λ. This is just some function of the data.
- We then study the properties of this estimator in **repeated samples**, that is, we consider the variation of $\hat{\lambda}$ across repeated hypothetical samples. This is just a thought experiment - we always only have one sample
- The standard estimator for this problem is

$$\hat{\lambda} = \frac{\sum_i y_i}{100}$$

- So if we observe 20 adopters in a sample of 100, then $\hat{\lambda} = 0.2$.
- The estimator is out best guess of the true λ. But how do we get plausible values of λ?

Repeated Sample Distribution

- Suppose we get repeated samples of $N = 100$ and we keep applying the estimator $\hat{\lambda}$. What is the distribution of the realized estimates $\hat{\lambda}_1, \hat{\lambda}_2, \ldots$?



Summary

- First Approach = Bayesian
  - Statements about λ are made conditional on the observed data
  - No requirements of population/repeated sample set-up
  - You can make probability statements about parameters (λ) and hypotheses (e.g., $0.1 < \lambda < 0.2$)
  - You can add prior information about λ
- Second Approach = Classical
  - Parameters are fixed constants
  - Estimators are evaluated in repeated samples from population
  - You cannot make probability statements about parameters or hypotheses (e.g., you cannot evaluate the probability that $0.1 < \lambda < 0.2$).
  - Hard to add prior information

Classical Approach: Probability = ?

- Long run frequency of outcome of a “repeated random experiment”
- But..
  - Hard to define precisely what a random experiment is!
  - What about situations where repeated random experiments doesn’t make sense?
  - Can we ever get repeated random samples - where nothing else changes - except a random draw?

Bayesian Approach: Probability = ?

- Everything not observed has a probability distribution attached to it
- This probability distribution encodes the uncertainty associated with the corresponding quantity
- For example, in the example above we had $\lambda \in \mathrm{Uniform}[0.01, 0.50]$ before we observed any data. This reflected our current beliefs about the unknown quantity λ.
- In this interpretation probabilities are detached from the idea of describing something “random”. Instead probabilities encode how uncertain something unknown is.

Bayesian Foundations

Two Required Ingredients to a Bayesian Model

- Generative Model of Data:

$$p(Y|\theta)$$

  where $Y = \text{observed data}$. This is also called the the likelihood function. It specifies the joint distribution of the observed data, conditional on the unknown parameters/weights.

- Prior knowledge:

$$p(\theta)$$

  This is called the prior distribution. It characterizes the state of our knowledge about the parameters θ before we observe any data.

Bayesian Updating

- After having observed the data Y we update our knowledge about the parameters θ using Bayes Rule:

$$p(\theta|Y) = \frac{p(Y|\theta)p(\theta)}{p(Y)} = \frac{p(Y|\theta)p(\theta)}{\int p(Y|\theta)p(\theta)d\theta}$$

后验分布

This is called the posterior distribution. It characterizes the state of our knowledge about the parameters θ after having observed the data. simulate the distrution

- Note that θ is typically a large dimensional array of parameters. The posterior is a multidimensional distribution.

- A Bayesian analysis involves a full characterization of the posterior distribution

- Only in the simplest model can the posterior distribution be derived analytically. In complex models this distribution is characterized using numerical techniques.

Example I

- Let's derive the posterior distribution of λ in the example above. To make the math a little easier assume that the prior is a standard uniform distribution: $U(0, 1)$.
- The full model is

$$\Pr(Y_i = y_i|\lambda) = \lambda^{y_i}(1-\lambda)^{1-y_i}, \qquad i = 1, \ldots, N;$$
$$p(\lambda) = \mathrm{U}(0, 1). \longrightarrow \text{assumption}$$

- The likelihood function is

$$\Pr(Y_1 = y_1, \ldots, Y_N = y_N|\lambda) = \prod_{i=1}^{N} \Pr(Y_i = y_i|\lambda) = \prod_{i=1}^{N} \lambda^{y_i}(1-\lambda)^{1-y_i}.$$
$$= \lambda^{N_1}(1-\lambda)^{N-N_1},$$

where $N_1 = \#\{i : y_i = 1\}$.

- The posterior distribution is then

$$
\begin{aligned}
p(\lambda|Y) &= \frac{\lambda^{N_1}(1-\lambda)^{N-N_1}\ \mathrm{U}(\lambda|0,1)}{\int \lambda^{N_1}(1-\lambda)^{N-N_1}\ \mathrm{U}(\lambda|0,1)d\lambda}, \\
&= \frac{\lambda^{N_1}(1-\lambda)^{N-N_1}\ \mathbb{I}(\lambda \in (0,1))}{\int \lambda^{N_1}(1-\lambda)^{N-N_1}\ \mathbb{I}(\lambda \in (0,1))d\lambda}, \\
&= \frac{1}{\mathrm{B}(N_1+1, N-N_1+1)}\lambda^{N_1}(1-\lambda)^{N-N_1},
\end{aligned}
$$

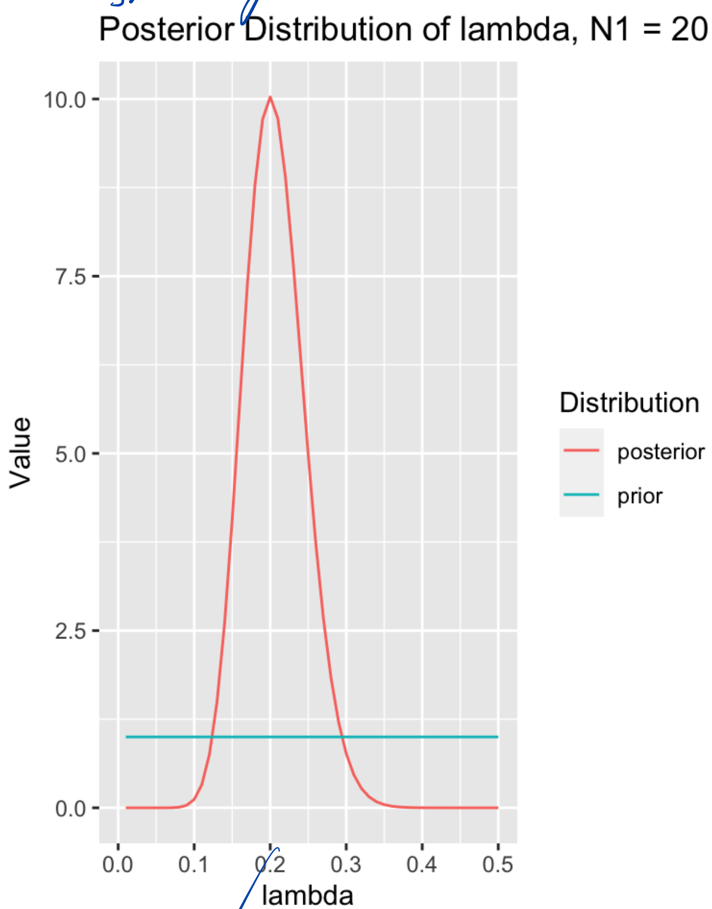
where $B(a, b)$ is the beta function defined as

$$B(a,b) \equiv \int_0^1 t^{a-1}(1-t)^{b-1}dt$$

- This is the density of the beta distribution:

$$p(\lambda|Y) = \mathrm{Beta}(\lambda|N_1+1, N-N_1+1).$$

showing how we learned



0.2 is the peak

- The beta distribution $B(a, b)$ has mean $a/(a+b)$
- Therefore the posterior mean of λ is

$$E[\lambda|Y] = \frac{N_1 + 1}{N + 2} = \frac{21}{102} \approx 0.206$$

- Under the uniform prior we have for $\underline{\lambda} = 0.3$:

$$\Pr(\lambda > \underline{\lambda}) = 70\%,$$
$$\Pr(\lambda > \underline{\lambda}|Y) = 1.4\%,$$

  where the second probability is the right tail probability at 0.3 for a Beta(21,81) distribution:

```
pbeta(0.3,21,81,lower.tail = F)
```

Different Prior

- Suppose the uniform prior doesn't capture our prior state of knowledge
- A more general prior for a fraction is

$$\lambda \sim \mathrm{Beta}(a_0, b_0)$$

- This has the uniform distribution as a special case ($a_0 = b_0 = 1$)
- This prior can characterize asymmetric distributions of λ, e.g., $a_0 =$ and $b_0 = 10$.
- The posterior can easily be derived to be

$$p(\lambda|Y) = \mathrm{Beta}(\lambda|N_1 + a_0, N - N_1 + b_0).$$

Posterior Predictive Distribution

- What is the distribution of a new data point y_{N+1} conditional on observing $Y = \{y_i\}_{i=1}^N$?
- If we knew θ this would simply be

$$p(y_{N+1}|\theta)$$

- In general we don't know θ, but our current state of knowledge is summarized by the posterior $p(\theta|y)$
- We define the posterior predictive distribution as

$$p(y_{N+1}|Y) = \int p(y_{N+1}|\theta)p(\theta|Y)d\theta$$

- Note that we can think of this as an ensemble method:

$$p(y_{N+1}|Y) \approx \frac{1}{S}\sum_{s=1}^{S} p(y_{N+1}|\tilde{\theta}_s),$$

  where $\{\tilde{\theta}_s\}_{s=1}^S$ is a large set of random draws from $p(\theta|Y)$.

Example I revisited

- The posterior predictive distribution is very simple in this case:

$$\begin{aligned}
\Pr(Y_{N+1} = 1|Y) &= \int \Pr(Y_{N+1} = 1|\lambda) p(\lambda|Y) d\lambda \\
&= \int \lambda \; p(\lambda|Y) d\lambda \\
&= \mathrm{E}[\lambda|Y] \\
&= \frac{N_1 + 1}{N + 2}
\end{aligned}$$

Example II: Bayesian A/B Testing

- A company is testing two different online ads - A and B
- Suppose ad A had 10,000 impressions with 317 click-throughs and B had 5,000 impressions with 152 click-throughs
- Which ad should we pick?
- The raw estimate of the CTR for A is $317/10000 \approx 0.032$ and $152/5000 \approx 0.03$ for B
- So we should pick A?

Posterior Calculation

- Let λ_A and λ_B be the true click-through rates under ad A and B
- Suppose we assume a prior as

$$\lambda_A, \lambda_B \sim \text{Beta}(1, 20)$$

- Using the results from above we then have posteriors

$$\begin{aligned} \lambda_A|Y &\sim \text{Beta}(317 + 1, 10000 - 317 + 20), \\ \lambda_B|Y &\sim \text{Beta}(152 + 1, 5000 - 152 + 20). \end{aligned}$$

- What is evidence for $\delta \equiv \lambda_A - \lambda_B > 0$?

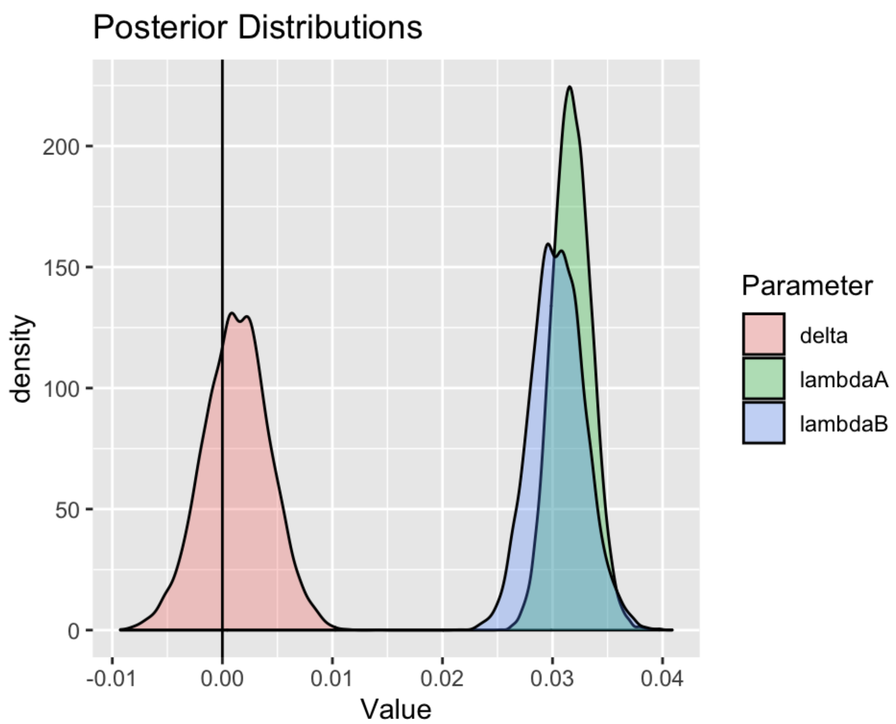
Posterior Simulation

- How to simulate posterior of λ_A, λ_B and δ?
- We can use the following procedure:
  - First sample $nSim$ draws of λ_A and λ_B from their respective Beta distributions
  - Let these be $\{\tilde{\lambda}_{A,s}, \tilde{\lambda}_{B,s}\}_{s=1}^{nSim}$
  - Next define $\tilde{\delta}_s = \tilde{\lambda}_{A,s} - \tilde{\lambda}_{B,s}$ for each $s = 1, \ldots, nSim$
  - Then $\{\tilde{\delta}_s\}_{s=1}^{nSim}$ will be draws from the implied posterior of δ
  - We can approximate $\Pr(\delta > 0|Y)$ simply as the fraction of positive $\tilde{\delta}_s$

```
nSim <- 10000
lambdaAPost <- rbeta(nSim,yA + a0,nA - yA + b0)
lambdaBPost <- rbeta(nSim,yB + a0,nB - yB + b0)
deltaPost <- lambdaAPost - lambdaBPost

ProbDeltaPos <- sum(deltaPost > 0)/nsim
```

Result



- $\Pr(\delta > 0) \approx 0.67$
- Should we go with option A?

Accounting for Risk

- What is the associated risk of a decision?
- Example of loss function:

$$L(\lambda_A, \lambda_B, D) = \begin{cases} \lambda_B - \lambda_A, & \text{if } D = A \text{ and } \lambda_B > \lambda_A, \\ 0, & \text{if } D = A \text{ and } \lambda_A > \lambda_B, \\ \lambda_A - \lambda_B, & \text{if } D = B \text{ and } \lambda_A > \lambda_B, \\ 0, & \text{if } D = B \text{ and } \lambda_B > \lambda_A \end{cases}$$

- We evaluate the loss of a decision D as

go with A or go with B

$$\hat{L}(D) \equiv \int L(\lambda_A, \lambda_B, D)\, p(\lambda_A, \lambda_B | Y)\, d\lambda_A d\lambda_B,$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} L(\tilde{\lambda}_{A,s}, \tilde{\lambda}_{B,s}, D).$$ monte Carlo simulation

S large error ↓

Posterior Risk

$$\hat{L}(A) = 0.00067$$

$$\hat{L}(B) = 0.0019$$

- The risk of choosing A is about three times lower than choosing B

- We can also include other information in the decision analysis, e.g., costs of different decisions

Example III: Gaussian Model with known variance

avg of output

$$\begin{aligned} Y_i|\mu &\sim \mathrm{N}(\mu, \sigma^2), \qquad i = 1, \dots, N, \\ \mu &\sim \mathrm{N}(\mu_0, \sigma_0^2), \end{aligned}$$

my prior belief of data

where we assume that σ is known (as well as the prior parameters μ_0, σ_0).

- This model is simple enough that we can solve for the posterior distribution analytically
- The likelihood $p(Y_1, \dots, Y_N|\mu)$ is

because we assume they're independent

$$\begin{aligned} p(Y_1, \dots, Y_N|\mu) &= \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2}(Y_i - \mu)^2\}, \\ &= \frac{1}{(\sigma\sqrt{2\pi})^N} \exp\{-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - \mu)^2\} \end{aligned}$$

“Completing the Square”

- A useful result:

$$\sum_{j=1}^{J} c_j(x - m_j)^2 = c(x - m)^2 + C, \quad \text{doesn't depend on } x$$

occurs J times occurs 1 time

where

$$c = \sum_{j=1}^{J} c_j, \quad \text{sum}$$

$$m = \frac{\sum_{j=1}^{J} c_j m_j}{\sum_{j=1}^{J} c_j}, \quad \text{weighted avg of } m$$

and C is some constant that doesn't involve x

Deriving posterior

$p(Y_1, \dots | \mu) = k_1 \cdot e^{-\frac{1}{2\sigma}\sum(Y_i - \mu)^2}$

$p(\mu) = k_2 e^{-\frac{1}{2\sigma_0}(\mu - \mu_0)^2}$

$$p(\mu|Y) = \frac{\cancel{k_1 k_2}\, e^{-\frac{1}{2}\left[\frac{1}{\sigma_0^2}(\mu - \mu_0)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{N}(Y_i - \mu)^2\right]}}{\int \cancel{k_1 k_2}\, e^{-\frac{1}{2}[\ ]}\, du}$$

- The posterior for μ is then

$$p(\mu|Y) = \frac{\exp\left\{-\frac{1}{2}h(\mu)\right\}}{\int \exp\left\{-\frac{1}{2}h(\mu)\right\} d\mu},$$

where

$$h(\mu) \equiv \left[\frac{1}{\sigma_0^2}(\mu - \mu_0)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{N}(\mu - Y_i)^2\right],$$

and we have canceled all constants not depending on μ.

Deriving posterior

- Using the result from above we then get

$$h(\mu) = c(\mu - m)^2 + C,$$

where

$$c \equiv \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2},$$

$$m \equiv \frac{\frac{1}{\sigma^2}\sum_{i=1}^{N} Y_i + \frac{1}{\sigma_0^2}\mu_0}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}} = \frac{\frac{N}{\sigma^2}\bar{Y} + \frac{1}{\sigma_0^2}\mu_0}{\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}}$$

avg of Y

$\sigma_0 \downarrow$

prior weight in the posterior is big

Deriving posterior

- Since C is just constant that doesn't depend on μ we then have

$$p(\mu|Y) = \frac{\exp\left\{-\frac{c}{2}(\mu-m)^2\right\}}{\int \exp\left\{-\frac{c}{2}(\mu-m)^2\right\}d\mu} = K \times \underbrace{\exp\left\{-\frac{c}{2}(\mu-m)^2\right\}}_{\text{another normal dist.}},$$

  where K is another constant that doesn't depend on μ.

- We recognize this as the density of a normal distribution with mean m and variance $1/c$:

$$p(\mu|Y) = \mathrm{N}(\mu|m, c^{-1})$$

Bayesian update will be normal

if $\begin{cases} \text{prior is normal} \\ \text{model is normal} \end{cases}$

Posterior Analysis

- Note that as the prior gets “flat”, i.e., σ_0 gets large, the posterior concentrates around the sample average:

$$\mathrm{E}[\mu|Y] = m \rightarrow \bar{Y} \text{ as } \sigma_0 \rightarrow \infty$$

- On the other hand, when the prior has a large weight, i.e., $1/\sigma_0^2$ is large, then the posterior mean is pulled towards the prior mean μ_0.
- This is an illustration of the “regularizing” effect of a prior. This is beneficial when the prior encodes information about μ that we already have prior to observing the data

Posterior with Different Prior Strength