Statistical Thinking: Week 8 Lab

Due 12noon Wednesday 14 October 2020

Introduction

The purpose of the Lab this week is again to learn more about the application of Bayes theorem. In this case however, we'll practice using Bayes theorem when the unknown variable takes on an infinite number of values. To keep things manageable, we will focus on conjugate situations.

The Lab is structured in two main sections, Part A and Part B. Part A contains some preliminary activities and questions to help prepare you for Part B, where you will explore a Bayesian A/B testing application. Each part contains several questions for you to answer.

Note that the **Lab08script.R** file contains the **R** code shown in this document. You will need to complete some additional code in order to complete the Lab.

Once you have completed these questions, you can attempt the Lab submission.

Lab Submission

To obtain credit for this lab, you are required to complete a Lab 8 Submission Moodle quiz, which is due in Week 9 on Wednesday 14 October at 12noon. You are not required to submit the .Rmd or .pdf for this Lab.

Good luck and have fun!

Part A: Preparation

The data model

Consider an experiment where a random sample of n observations, each taking one of two possible (binary) outcomes, with

$$u_i = \begin{cases} 1 & \text{if success} \\ 0 & \text{if failure.} \end{cases}$$

Each of the u_i random outcomes (called a "*Bernoulli trial*") has the same probability of a success, θ. The likelihood function corresponding to $u_1, u_2, \ldots, u_n \mid \theta \overset{i.i.d}{\sim} Bernoulli(\theta)$ is

$$L_n(\theta) = \prod_{i=1}^{n} \theta^{u_i}(1-\theta)^{1-u_i} = \theta^{\sum_{i=1}^{n} u_i}(1-\theta)^{n-\sum_{i=1}^{n} u_i}.$$

Rather than thinking of modelling each of the individual $U_i = u_i$ outcomes as a single $Bernoulli(\theta)$ random variable, we can consider the sum of n such $Bernoulli(\theta)$ outcomes, $X = \sum_{i=1}^{n} U_i$. In this case, $X \mid \theta \sim Binomial(n, \theta)$ distribution, taking on a value $x \in \{0, 1, \ldots, n\}$. Once $X = x$ is observed, the likelihood function associated with $X \mid \theta \sim Bernoulli(\theta)$ is given by

$$L_n(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}.$$

Note that the combination term $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ does not depend on θ.

A Beta prior for the Binomial (or Bernoulli) likelihood

The $Beta(\alpha, \beta)$ distribution is the prior conjugate to the Binomial likelihood function. Remember the probability density function (pdf) for $\theta \sim Beta(\alpha, \beta)$ is

$$f(\theta \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \quad \theta \in (0,1) \text{ for } \alpha > 0,\ \beta > 0.$$

Also we know (from the Week 6 lecture slides) the values of the prior mean and variance:

- $E(\theta) = \frac{\alpha}{\alpha+\beta}$
- $Var(\theta) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

Probabilities for θ (e.g. $\Pr(\theta < 0.2)$) and quantiles associated with the $Beta(\alpha, \beta)$ distribution may be obtained using the **R** function *pbeta*().

Bayes Theorem

Bayes theorem, stated for a continuous parameter and a model for continuous data, says that if $X \mid \theta$ has pdf given by $f(x \mid \theta)$, and θ has pdf $f(\theta)$, then the posterior pdf for $\theta \mid X = x$ satisfies

$$\text{Bayes theorem:} \qquad f(\theta \mid x) = \frac{f(x \mid \theta)\ f(\theta)}{\int_\Theta\ f(x \mid \theta)\ f(\theta)\ d\theta}.$$

Beta-Binomial Conjugate Pair

As detailed in the Week 8 lecture slides and discussed in the corresponding video, if $\theta \sim Beta(\alpha,\ \beta)$, for $0 < \theta < 1$, denotes the prior distribution, and if X given n and θ has a $Binomial(n, \theta)$ distribution, then the posterior distribution for θ when $X = x$ will be a $Beta(\tilde{\alpha},\ \tilde{\beta})$, where $\tilde{\alpha} = \alpha + x$ and $\tilde{\beta} = \beta + (n - x)$.

The posterior mean, $E[\theta \mid x]$, can be described as a weighted average of the maximum likelihood estimate, given by $\hat{\theta}_{MLE} = \frac{x}{n}$, and the prior mean, $E[\theta]$, i.e.

$$E[\theta \mid x] = Z\left(\frac{x}{n}\right) - (1 - Z)E[\theta].$$

The so-called *credibility factor*, which determines the weight on the MLE is

$$Z = \frac{n}{\alpha + \beta + n}.$$

Question 1

Create a function like the one below, named *beta_binomial*(), that takes as input values:

- $n =$ the number of independent and identically distributed (i.i.d.) $Bernoulli(\theta)$ draws
- $x =$ the number of successes in the n draws
- $\alpha =$ the first (hyper-)parameter of the $Beta(\alpha, \beta)$ prior distribution (with default value = 1)
- $\beta =$ the second (hyper-)parameter of the $Beta(\alpha, \beta)$ prior distribution (with default value = 1)

and returns as output a list containing the elements:

- $\tilde{\alpha}$ = the first (hyper-)parameter of the $Beta(\tilde{\alpha}, \tilde{\beta})$ posterior distribution
- $\tilde{\beta}$ = the second (hyper-)parameter of the $Beta(\tilde{\alpha}, \tilde{\beta})$ posterior distribution
- Z = the corresponding *credibility factor* for this posterior, representing the proportion of the posterior mean attributable to $\hat{\theta}_{MLE} = x/n$.

```
beta_binomial <- function(n, x, alpha = 1, beta = 1) {

    atil <- alpha + x
    btil <- beta + n - x

    cf <- n/(alpha + beta + n)

    out <- list(alpha_tilde = atil, beta_tilde = btil,
        credibility_factor = cf)
    return(out)
}
```

Then, suppose $\theta \sim Beta(\alpha = 4,\ \beta = 4)$ when X is modelled with a $Binomial(n = 40, \theta)$ distribution, and $x = 12$ is observed. Use the format of the code chunk below together with your *beta_binomial*() function, to find the (hyper-)parameters of the posterior distribution, saving the result of your function evaluation as an object named **Q1out**. Also, to facilitate use of variables in later portions of the Lab, add lines to the code chunk below to define each of the following variables:

- **n** (the number of $Bernoulli(\theta)$ trials)
- **xobs** (the observed value of x)
- **alpha** (the prior value of α)
- **beta** (the prior value of β)
- **Q1out** (the output of beta_binomial(n=n, x=xobs, alpha=alpha, beta=beta))

Detail the form of the posterior distribution when the $Binomial(n = 40, \theta)$ outcome is $x = 12$, and the prior distribution is given by $\theta \sim Beta(\alpha = 4, \beta = 4)$. Explain (in your own words) what this distribution represents.

Once you have obtained your output, you can visualise them using the plot commands below.[1]

Note that the *normalised* likelihood function, given by $\mathcal{L}_n(\theta)/\int_0^1 \mathcal{L}_n(\theta)\ d\theta$, was plotted instead of the usual likelihood function, $\mathcal{L}_n(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$.

```
#### Code to visualise components of Bayes theorem
#### ####

# A colourblind-friendly palette with black:
cbbPal <- c(black = "#000000", orange = "#E69F00",
    ltblue = "#56B4E9", "#009E73", green = "#009E73",
    yellow = "#F0E442", blue = "#0072B2", red = "#D55E00",
```

[1]For information about the “colour blind palette with black” (named here “cbbPal”), see the R Graphics Cookbook

```
    pink = "#CC79A7")

cbbP <- cbbPal[c("orange", "blue", "pink")]  #choose colours for p1

thetaxx <- seq(0.001, 0.999, length.out = 100)
priorxx <- dbeta(thetaxx, alpha, beta)
postxx <- dbeta(thetaxx, Q1out$alpha_tilde, Q1out$beta_tilde)
likexx <- dbinom(xobs, size = n, prob = thetaxx)
nlikexx <- 100 * likexx/sum(likexx)

df <- tibble(theta = thetaxx, `prior pdf` = priorxx,
    `normalised likelihood` = nlikexx, `posterior pdf` = postxx)

df_longer <- df %>% pivot_longer(-theta, names_to = "distribution",
    values_to = "density")

p1 <- df_longer %>% ggplot(aes(x = theta, y = density,
    colour = distribution, fill = distribution)) +
    geom_line() + scale_fill_manual(values = cbbP) +
    theme_bw()
```

Question 2

Given your prior and posterior distributions from **Question 1**, find both the *prior probability* $\Pr(\ 0.2 \leq \theta \leq 0.4)$, and the *posterior probability* $\Pr(\ 0.2 \leq \theta \leq 0.4 \mid x = 12)$.

Question 3

Write a function, named *beta_meanvar*(), that takes as input the parameter values *alpha* and *beta*, corresponding to α and β, respectively, and returns a list containing the mean (named *mean*) and variance (named *var*) of the $Beta(\alpha, \beta)$ distributions.

Then, use your function to calculate the mean and variance for both the $\theta \sim Beta(\alpha = 4, \alpha = 4)$ *prior* distribution, and for the corresponding *posterior* distribution for θ when $x = 12$, obtained in **Question 1**.

Question 4

Given the prior distribution, values of x and n, the output of your *beta_binomial*() and *beta_meanvar*() functions from **Question 1** and **Question 3**, respectively, confirm that the posterior mean satisfies the relation

$$E[\theta \mid x] = Z\left(\frac{x}{n}\right) + (1 - Z)E[\theta].$$

Question 5

Explain why the denominator in Bayes theorem above, once evaluated, is **not** a function of θ. Then, using the data $x = 12$ from n Bernoulli trials, and the $Beta(\alpha = 4, \beta = 4)$ prior, calculate the value of this denominator as follows:

$$f(x) = \int_{\Theta} f(x \mid \theta)\ f(\theta)\ d\theta = \frac{f(x \mid \theta^*)\ f(\theta^*)}{f(\theta^* \mid x)},$$

for any $\theta = \theta^* \in (0, 1)$.
[Hint: remember you can evaluate the likelihood using the *dbinom*() function, and you can evaluate both the prior and the posterior pdfs using the *dbeta*() function. As for θ^*, any value in $(0, 1)$ should give you the same answer, so just pick a few to check that you always get the same answer!]

Part B: A/B testing

What is A/B testing? See the Wikipedia entry. Focus on the first paragraph, and the first two sentences of the second paragraph, replicated here for your convenience.

A/B testing (bucket tests or split-run testing) is a randomized experiment with two variants, A and B. It includes application of statistical hypothesis testing or "two-sample hypothesis testing" as used in the field of statistics. A/B testing is a way to compare two versions of a single variable, typically by testing a subject's response to variant A against variant B, and determining which of the two variants is more effective.

As the name implies, two versions (A and B) are compared, which are identical except for one variation that might affect a user's behavior. Version A might be the currently used version (control), while version B is modified in some respect (treatment).

A promotional email campaign

Further down in the Wikipedia entry, you will find a section entitled: **Example**. This is the setting for the activities in this computer lab assignment.[2] Read the example description on Wikipedia.

We consider here a similar **Call to Action** experiment involving emails sent to 2000 different people independently sampled from a very large customer database. All email messages contain a promotional offer to buy a certain product with a discount, and are **identical** except for one sentence:

Call to Action A: **Offer ends this Saturday! Use code A1**

Call to Action B: **Offer ends soon! Use code B1**

Half of the emails sent ($n_A = 1000$) include the Call to Action A message and the other half of the emails ($n_B = 1000$) include the Call to Action B message. Those sent the Call to Action A message belong to **group A** and the remaining who received the Call to Action B message email belong to **group B**. There is no overlap between the two groups.

In addition, we refer to the **effectiveness rate** (or *effectiveness probability*) for group A as θ_A and the effectiveness rate for group B as θ_B. These rates reflect the chance that a person from the relevant group will purchase the product using the corresponding promotional code.

We want to answer questions such as:

1. **Which of the two Calls to Action, A or B, is more effective?** i.e., which Call to Action strategy has the higher effectiveness rate?

[2] While the setting of this activity has been taken from the Wikipedia entry, many of the specific details of the assignment differ from those described on the Wikipedia page.

2. **What is the distribution of possible values for the difference between θ_A and θ_B?**

In this particular situation, the firm undertaking the A/B testing has a lot of experience using Call to Action strategies, and have characterised the effectiveness rate for each scenario A or B as having a $Beta(\alpha = 2, \beta = 26)$ distribution. They also assume, prior to undertaking the experiment, that θ_A is independent of θ_B.

Note that given n, if $X \mid \theta \sim Binomial(n, \theta)$ and θ is considered random, then $E_{(X,\theta)}[X] = E_\theta[n\theta] = nE_\theta[\theta]$. (The subscript on the E refers to the random variables inside the square brackets over which the expectation is to be calculated.)

Question 6

According to the prior distribution, what is the expected number of people in group A who will take up their promotional offer? Also report a 95% "credible" interval for, satisfying

$$\Pr(L \leq n\theta_A \leq U) = 0.95,$$

for any two values $0 \leq L < U \leq n$.
[Hint: Remember that $n = 1000$ is fixed.]

Question 7

According to the prior distributions for effectiveness rates θ_A and θ_B, and given $n_A = n_B = 1000$, what is the difference between the expected number of people from group A who will take up their promotional offer and the expected number of people from group B who will take up their promotional offer? Explain why.

The experimental data

After running the A/B testing protocol, $x_A = 63$ customers purchased the product using discount code A and $x_B = 45$ customers purchased the product using discount code B.

Question 8

Given the results of the experiment, what are the posterior distributions for the effectiveness rates for group A and for group B, respectively?

Let $\Delta = \theta_B - \theta_A$ denote the **difference** between the effectiveness rates associated with Call to Action B and Call to Action A.

Question 9

Given the results of the experiment, what is the (posterior) mean and variance of Δ?
[Hint: Are the two posterior distributions independent? Why or why not?]

Using simulation to approximate the posterior of Δ

Although we can work out the posterior mean and variance of Δ fairly easily, we do not have access to the pdf of Δ. However, given what we know about the posterior distributions of θ_A and θ_B, it is easy to obtain a *sample* from this distribution using simulation.

To use simulation, draw a large number, say $R = 5000$ values of $\theta_A^{(r)} \overset{i.i.d.}{\sim} p(\theta_A \mid x_A = 63)$ and another $R = 5000$ values of $\theta_B^{(r)} \overset{i.i.d.}{\sim} p(\theta_B \mid x_B = 45)$, and compute the $R = 5000$ replications $\{\Delta^{(r)} = \theta_B^{(r)} - \theta_A^{(r)}$. Use the empirical sample of these $\Delta^{(r)}$ draws to approximate the posterior distribution, $p(\Delta \mid x_A = 63, x_B = 45)$.

Question 10

Simulate a sample of $R = 10000$ replicated values of Δ from $p(\Delta \mid x_A = 63, x_B = 45)$. Use this sample to approximate each of the following items:

a) the posterior mean of Δ,

b) the standard deviation of Δ, and

c) the probability that the two effectiveness rates will differ by more than 0.01, i.e. $\Pr(|\theta_B - \theta_A| > 0.01 \mid x_A = 63, x_B = 45)$.

Question 11: What would a frequentist do?

Of the frequentist techniques we have considered thus far in the *Statistical Thinking* unit, which do you think would be appropriate to apply in this setting? How would you implement them, and what type of 'inference' would they provide?
