Edps 590BAY
Fall 2019
C.J.Anderson

Homework 2
Answer Key

1. *For each of the scenarios, in your opinion would an informative or non informative prior be most appropriate? Briefly explain why you selected the prior.*

   (a) *A research group at small college has done a study which yielded a surprising result. You decide to replicate the study are UIUC and use Bayesian method to analyze your data.*

   I would use a non informative prior because the students as UIUC may be different. Also, when doing a replication study, I don't want previous results to influence my results.

   (b) *You are conducting an experiment using fMRI, which is very expensive. Since there is only one machine available, you collect data from subjects one at a time. To try to minimize your cost and time to complete the project, you decide to only collect data until you obtain a reasonable result within a pre-determined level of precision.*

   I would use an informative prior which was based on my results up to the current point. I would not be changing procedures, population, etc so using previous results is reasonable.

   (c) *You are trying to predict the probability of getting lung cancer given that a person is a smoker. Data from retrospective studies exist that give the incidence of lung cancer in the population of interest, the incidence of smoking in the population, and the incidence of a person being a smoker among lung cancer patients.*

   I would use an informative prior because previous information exists on

2. *Compute the probability among 40 year old or older women of having breast cancer given that they have a positive mammogram result using the following information:*

   - *According to major journals,* 4/10 *get a mammogram*
   - *The probability of breast cancer patients having a positive results is* 32/40.
   - *The probability of a positive results is* 1/10.
   - *The probability of cancer is* 1/69

So I have the following information:

$$\begin{aligned} Pr(\text{has mamogram}) &= .4000 \\ Pr(\text{positive result}|\text{cancer}) &= 32/40 = .8000 \\ Pr(\text{postiive}) &= .1000 \\ Pr(\text{cancer}) &= .01449 \end{aligned}$$

$$\begin{aligned} Pr(\text{cancer}|\text{positive}) &= \frac{Pr(\text{positive}|\text{cancer})Pr(\text{cancer})}{Pr(\text{positive})} \\ &= \frac{.80(.01449)}{.10} \\ &= .1192 \end{aligned}$$

Small probability of having cancer given positive results on mammogram (i.e., High false positive) .

3. *Below is data from the 2006, 2016 and 2018 General Social Survey where respondents indicated whether they favor or oppose a law which would require a person to obtain a police permit before he or she could buy a gun. Note that these are cross-sectional data (i.e., different respondents each year).*

	Response	
Year	favor	oppose
2006	1568	395
2016	1330	528
2018	1102	439

(a) *For the year* 2006*,*

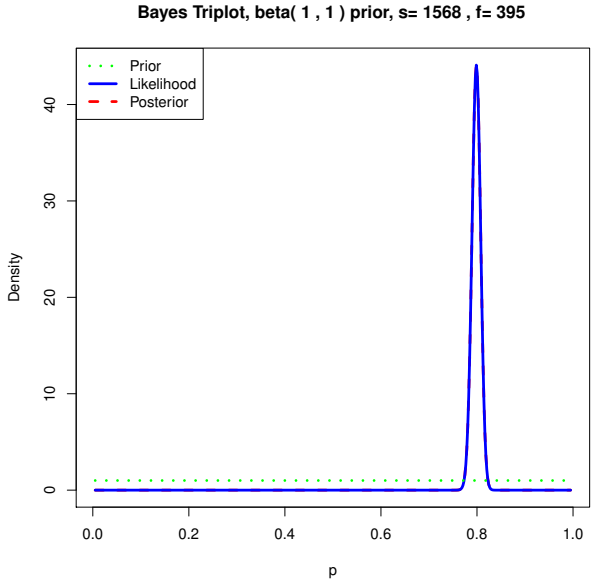
i. *What is the sample proportion to favor a law?*

$$p_{2006} = 1568/(1568 + 395) = .799$$

ii. *What is the posterior distribution for the proportion of people who favor gun control (use Uniform prior)? Also, plot the prior, likelihood and posterior and comment on the plot.*

```
library(LearnBayes)
triplot(c(1,1),c(1568,395),where="topleft")
```

The posterior distribution is a beta(1569,396).



iii. *What is the mean of this distribution? Comment.*
Comment: The likelihood and posterior are on top of each other because the prior has no impact (it's flat).

$$\text{mean} = a/(a+b) = 1569/(1569+396) = .798$$

iv. *What is the* 95% *credible interval? Interpret.*

```
theta.95 ← c(.025,.975)
(ci ← qbeta(theta.95,1569,396))
```

$$(.780, .816)$$

Given the data, the probability that the true mean is within the interval from .780 to .816 is .95. Also note that the sample proportion $p = .798$ is within this interval.

v. *What is the* 95% *high density interval? Interpret.*

```
inbeta ← rbeta(1E5,1589,396)
hdi.p ← hdi(inbeta,credMass=.95)
```

$$(.782, .818)$$

Given the data, the probability that the true mean is within the interval from .780 to .816 is .95.

vi. *If you were reporting this result in a paper, which interval would you use?*

It doesn't really matter, I could report either one. They are almost the same because posterior ends up being fairly symmetric (due to large sample size).

(b) For the year 2016,

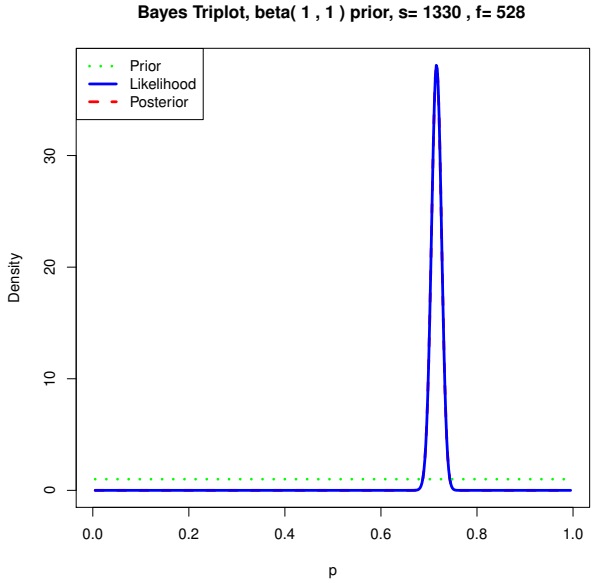
i. *What is the sample proportion to favor a law?*

$$p = 1330/(1330 + 528) = .716$$

ii. *Using data from 2016 as the prior, what is the posterior distribution for the proportion of people who favor gun control? Also plot the prior, likelihood and posterior and comment on the plot.*

```
triplot(c(1,1),c(1330,528),where="topleft")
```

Posterior distribution is beta(1331,529).
Comment: pretty tight and symmetric. Likelihood and posterior are the same because of the flat prior; that is, prior has not influence.



iii. *What is the mean of this distribution?*

$$\text{mean} = 1331/(1331 + 529) = .716$$

iv. *What is the 95% credible interval? Interpret.*

```
theta.95 ← c(.025,.975)
(ci ← qbeta(theta.95,1331,529))
```

$$.695, .736$$

Given the data, the probability that the true proportion is in the interval from .695 to .736 is .95. This interval include out sample proportion, which it should.

v. *What is the* 95% *high density interval? Interpret.*

```
inbeta ← rbeta(1E5,1331,529)
hdi.p ←
```

$$(.695, .736)$$

Given the data, the probability that the true proportion is in the interval from .695 to .736 is .95. This interval include out sample proportion, which it should. It is a bit narrower than the credible intervals (have to go to 4 decimal points to see it.

vi. *If you were reporting this result in a paper, which interval would you use?*

Either one would be fine.