ST440/540 Applied Bayesian Analysis

Lab activity for 1/29/2024

A. HOMEWORK AND QUIZ SOLUTIONS

Q1 Give an appropriate family of distributions for the following quantities. Justify your choice.

(a) Distance traveled when a new car first has a mechanical problem.
The support is positive real numbers, so a gamma would be appropriate

(b) The number of states with a female Governor in the year 2028.
Binomial since the support is {0,1,…,50}

(c) The difference between the Nasdaq Index tomorrow versus today.
The support is all real numbers so Gaussian would be appropriate.

Chapter 1, problem 6

The conditional distribution is f(x1|x2) = f(x1,x2)/f(x2). The marginal distribution f(x2) is hard to derive since it requires an integral. However, since we will plot f(x1|x2) only as a function of x1, f(x2) is just a constant that makes the conditional distribution integrate to one. Instead of computing f(x2) using integration, we can just numerically divide the sum of the joint distribution to approximation the conditional distribution. This is what is happening in the normalizing_constant step in (a).

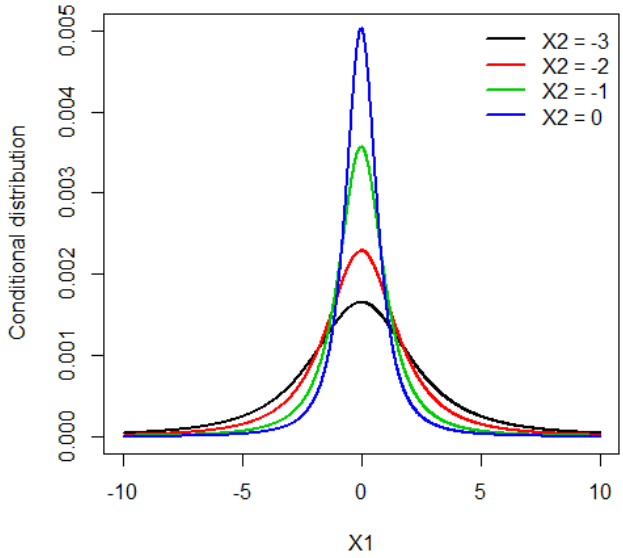
(a) The function below plot f(x1|x2) for x2 = -3, -2, -1, 0. Note that the plots are the same for x2 and -x2.

```
joint <- function(x1,x2){
  (1/(2*pi)) * (1+x1^2+x2^2)^(-3/2)
}

plot(NA,xlab="X1",ylab="Conditional distribution",
     xlim=c(-10,10),ylim=c(0,.005))

x1 <- seq(-10,10,.01)
x2 <- c(-3,-2,-1,0)
for(j in 1:4){
   density    <- joint(x1,x2[j])
   norm_const <- sum(density)
   density    <- density/norm_const
   lines(x1,density,col=j,lwd=2)
}

legend("topright",paste("X2 =",x2),col=1:4,lwd=2,bty="n")
```



(b) They are not independent because the distribution of x1 depends on x2 (e.g., the variance is smaller for x2=0 than other values).
(c) The mean of x1 is zero for all x2, therefore there is not a linear relationship between the mean of x1 and x2. This is a classic example of variables that are dependent but uncorrelated because their relationship can't be captured by the mean only.

Chapter 1, problem 9

https://www4.stat.ncsu.edu/~bjreich/BSMdata/C1#C1p9

Last problem: If 70% of a population is vaccinated, and the hospitalization rate is 5 times higher for an unvaccinated person than a vaccinated person, what is the probability that a person is vaccinated given they are hospitalized?
Let V = vaccinated and H = hospitalized, then the problem says Prob(V)=0.7, Prob(H|not V) = 5p and Prob(H|V) = p where p is the (unknown) probability of hospitalization for a vaccinated person. For Bayes rule we will need Prob(H) = Prob(H|V)Prob(V) + Prob(H|not V)(1-Prob(V)) = p*0.7+5*p*0.3 = p*2.2. Bayes Rule is then

$$\text{Prob}(V|H) = \text{Prob}(H|V)\text{Prob}(V)/\text{Prob}(H) = p*0.7/(2.2*p) = 0.70/0.85 = 32\%.$$

B. DISCUSSION QUESTIONS

(1) We'll use these results throughout:

(a) If Y|p~Binomial(n,p) and p ~ Beta(a,b), then p|Y ~ beta(Y+a,n-Y+b)

(b) p ~ beta(1,1) is equivalent to p ~ Uniform(0,1)

Say 1,000 high school students are randomly selected to enter a tutorial program. It is known that 70% of the population from which they are drawn graduate from high school. After the program, it is found that 725 of the 1,000 students graduate high school. We then want to test the hypotheses

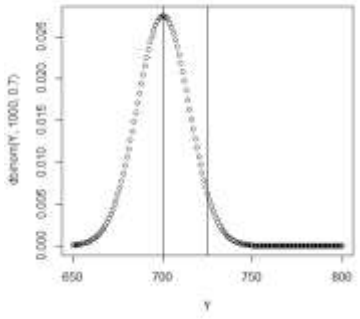
Ho: the graduation rate for students in the program is less than or equal to 70%

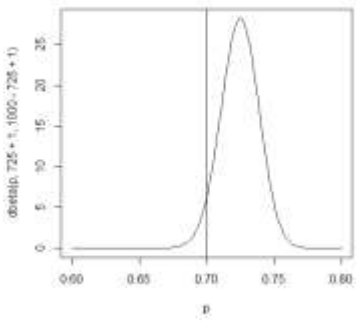
Ha: the graduation rate for students in the program is greater than 70%

Can we conclude the program is effective? Here are some plots/stats that may be useful:

```
> Y <- seq(650,800,1)
> plot(Y,dbinom(Y,1000,0.7))
> abline(v=700)
> abline(v=725)
> 1-pbinom(724,1000,0.7)
[1] 0.04459259
```

```
> p <- seq(0.6,0.8,0.001)
> plot(p,dbeta(p,725+1,1000-725+1),type="l")
> abline(v=0.7)
> pbeta(0.7,725+1,1000-725+1)
[1] 0.04272651
```





(a) How would a frequentist test of these hypotheses? Can we conclude the program was effective? Explain the results as if you're presenting them to a non-statistician.

Let n=1000 and Y be the number of students that graduate. We assume Y|theta ~ Binomial(n,theta). Under Ho, theta=0.7 and P(Y>=725) = 0.044 is the p-value. Since the p-value is less than 0.05, we reject Ho and conclude the program is effective.

(b) How would a Bayesian test these hypotheses? Can we conclude the program was effective? Explain the results as if you're are presenting them to a non-statistician.

Likelihood: Since Y is an integer between 0 and n we assume it is distributed Y|theta ~ Binomial(n,theta)

Prior: Since theta is a probability (real number between 0 and 1) we set prior theta ~ Beta(a,b). To make the prior uninformative we set a=b=1.

Posterior: The posterior is then theta|Y ~ beta(Y+a,n-Y+b). The code above computes P(Ha|Y) = P(theta<.7|Y) = 0.04, so the probability that the program is effective is 96%.

(c) Define a p-value and posterior probability of Ho, and describe how they are different.

The p-value is the probability, assuming null is true, of observing data more extreme than we observed. The posterior probability of the null is just what it says, P(Ho|Y)=P(theta<0.7|Y). The p-value quantifies uncertainty through Y|theta and the posterior probability of the null through theta|Y.

(2) Say we presented the results in (1) to the school board but they did not feel the study is large enough to be definitive. So, the next school year you enroll an additional 1,000 students and record that 745 graduated from high school.

(a) Describe how you would conduct a Bayesian analysis of these data. Give the likelihood, prior and posterior and describe how you would summarize the results.

Option 1: We pool the data from the two years as Y = 725+745 = 1470 and n=2000 and then do the same analysis as in 1b, i.e., theta|Y ~ beta(Y+1,n-Y+1) = beta(1470+1,2000-1470+1) and compute posterior probability of Ho.

Option 2: Treat the posterior from the first 1000 as the prior for the second 1000 and then compute posterior probability of Ho. After the first 1000 we have theta following a beta distribution with a=725+1 and b = 1000-725+1. With this prior, likelihood Y|theta ~ Binomial(1000,theta) and Y=745, the final posterior is theta ~ beta(745+a,1000+b) = beta(745+725+1,1000+1000+1).

It turns out both options are equivalent!!!

(b) What assumptions you are making and how might you justify them?

We are assuming the success probability is the same in both years. This could be tested by comparing data across years.

(3) In last year's midterm exam, we reanalyzed the data from the initial Moderna COVID vaccine trial. The data are in the table below.

Placebo		Vaccine	
Infected	Participants	Infected	Participants
185	14073	11	14134

Let θ_0 be the probability of getting infected under placebo and θ_1 be the probability under vaccine.

(a) Can we say θ_0 = 185/14073 = 0.01315 and θ_1 = 11/14134 = 0. 00078? Why?

No. Because these are sample proportions (statistics) not the true probabilities (parameters).

(b) What priors would you pick for θ_0 and θ_1?

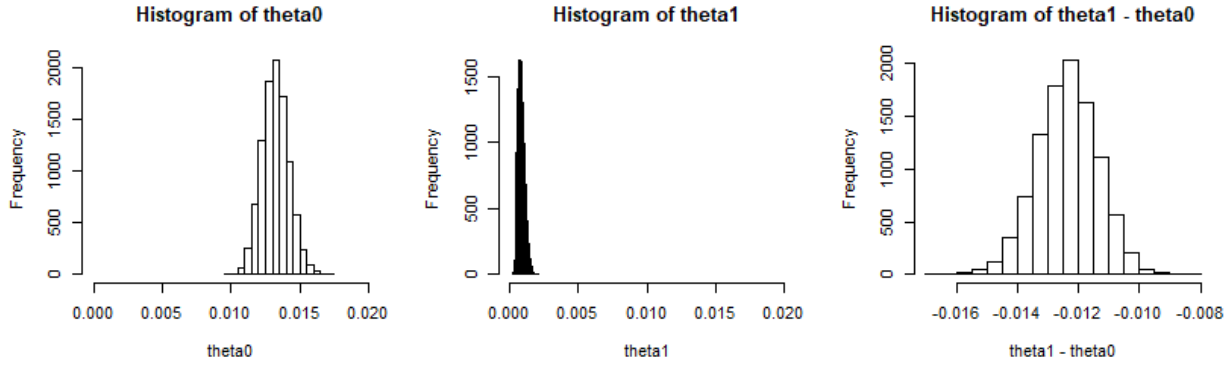
θ_0 ~ Beta(1,1) and θ_1 ~ Beta(1,1) (independent of each other) puts equal mass on all probabilities and is thus a reasonable starting place.

(c) How would you conduct a Bayesian test that the vaccine is effective? Give the likelihood, priors and posterior and how you would summarize the results.

Say n0 = 14073 and n1 = 14134 are the number of observations in each group, and Y0 = 185 and Y1 = 11 are the number that get infected. The likelihood is chosen to be Y_0 |theta0 ~ Binomial(n0,theta0) and Y1 ~ Binomial(n1,theta1) because both Y0 and Y1 are counts bounded by the sample size. With the prior in (b) we have theta0 | Y0 ~ beta(Y0+1,n0-Y0+1) and theta1 | Y1 ~ beta(Y1+1,n1-Y1+1). We summarize the results by computing P(theta1<theta0|Y0,Y1), which is approximated using Monte Carlo sampling below. The probability is 1.0 that the infection probability is lower in the vaccine group than the placebo group.

```
> n0 <- 14073
> n1 <- 14134
> Y0 <- 185
> Y1 <- 11
> S  <- 10000

> theta0 <- rbeta(S,Y0+1,n0-Y0+1)
> theta1 <- rbeta(S,Y1+1,n1-Y1+1)
> hist(theta0,xlim=c(0,0.02))
> hist(theta1,xlim=c(0,0.02))
> hist(theta1-theta0)
> mean(theta1>theta0)
0.0
```



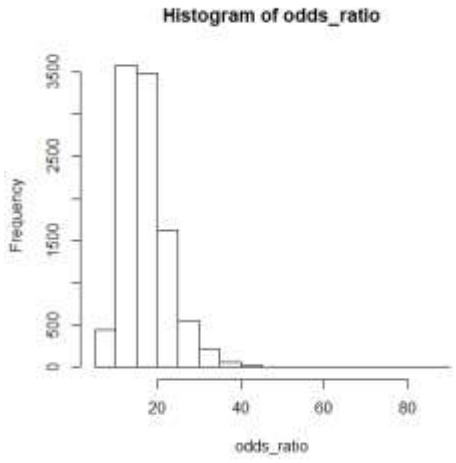
(d) What are some key assumptions you're making and how might you justify them?

Assuming the people in the two groups are comparable, which should be the case if they were randomized into the two groups. A binomial distribution assumes all patients are independent.

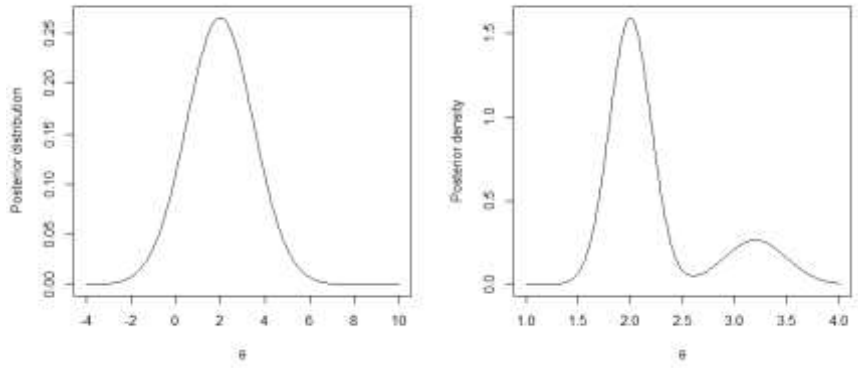
(4) Communicating results from studies such as in (3) is difficult because the probabilities are so small. Therefore, you often hear statements like “the odds of contracting the virus are X times higher if you are unvaccinated compared to vaccinated.” (the odds of an event are the probability it occurs divided by the probability it does not occur.) Write R code to use the data from (3) to compute a point estimate, 95% credible interval and plot of the posterior distribution of the odds ratio X.

The posterior mean and 95% credible interval are 17.1 and (9.4,30.9) and the histogram is below.

```
> theta0 <- rbeta(S,Y0+1,n0-Y0+1)
> theta1 <- rbeta(S,Y1+1,n1-Y1+1)
> odds0  <- theta0/(1-theta0)
> odds1  <- theta1/(1-theta1)
> odds_ratio <- odds0/odds1
> hist(odds_ratio)
> mean(odds_ratio)
[1] 17.16515
> quantile(odds_ratio,c(0.025,0.975))
    2.5%     97.5%
 9.359827 30.875484
```



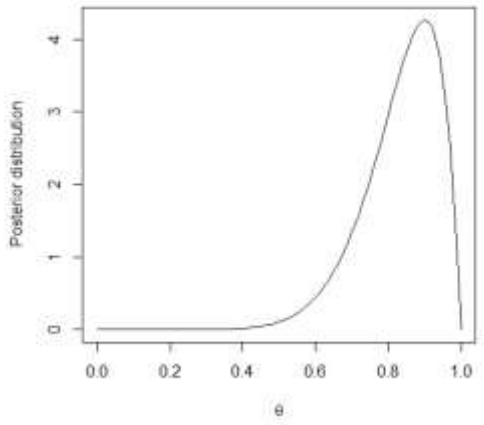
(5) How would you summarize the posterior distributions below in a table?



Left: This appears to be approximately Gaussian so a mean and variance will suffice.
Right: This is very complicated and so it's probably better to show the plot.

(6) Say we observe Y=9 successes in n=10 trials and use a uniform Beta(1,1) prior for the success probability so the posterior is Beta(Y+a,n-Y+b) = Beta(10,2).

```
> theta <- seq(0,1,0.01)
> plot(theta,dbeta(theta,10,2),type="l",xlab=expression(theta),ylab="Posterior distribution")
```



(a) Give R code to compute an equal tailed 90% interval.

```
> qbeta(0.05,10,2)
[1] 0.6356405
> qbeta(0.95,10,2)
[1] 0.9666808
```

(b) The highest posterior density interval "searches for the smallest interval that contains the proper probability." Write (or at least sketch out) R code to compute this interval.

```
> p   <- seq(0,0.1,length=100)
> lo  <- rep(0,100)
> hi  <- rep(0,100)
> for(i in 1:100){
+  lo[i] <- qbeta(p[i],10,2)
+  hi[i] <- qbeta(1-(0.1-p[i]),10,2)
+ }
> width    <- hi-lo
> shortest <- which.min(width)
> p[shortest]
[1] 0.09292929
> lo[shortest]
[1] 0.683717
> hi[shortest]
[1] 0.988255
```

(c) Which interval to you expect to have this highest lower bound? That is, if the first is (L_{ET}, U_{ET}) and the second is (L_{HPD}, U_{HPD}), do you expect L_{ET}>L_{HPD}?

The HPD interval has higher lower bound. Because the distribution is left-skewed, the HPD includes more of the right side of the distribution.