STAT 5444G
Homework Assignment 2
Colburn Hassman-906006816

Feb 28, 2021

[2] 1. The shape of the beta distribution depends on the parameters α and β. Plot each of the following densities in R

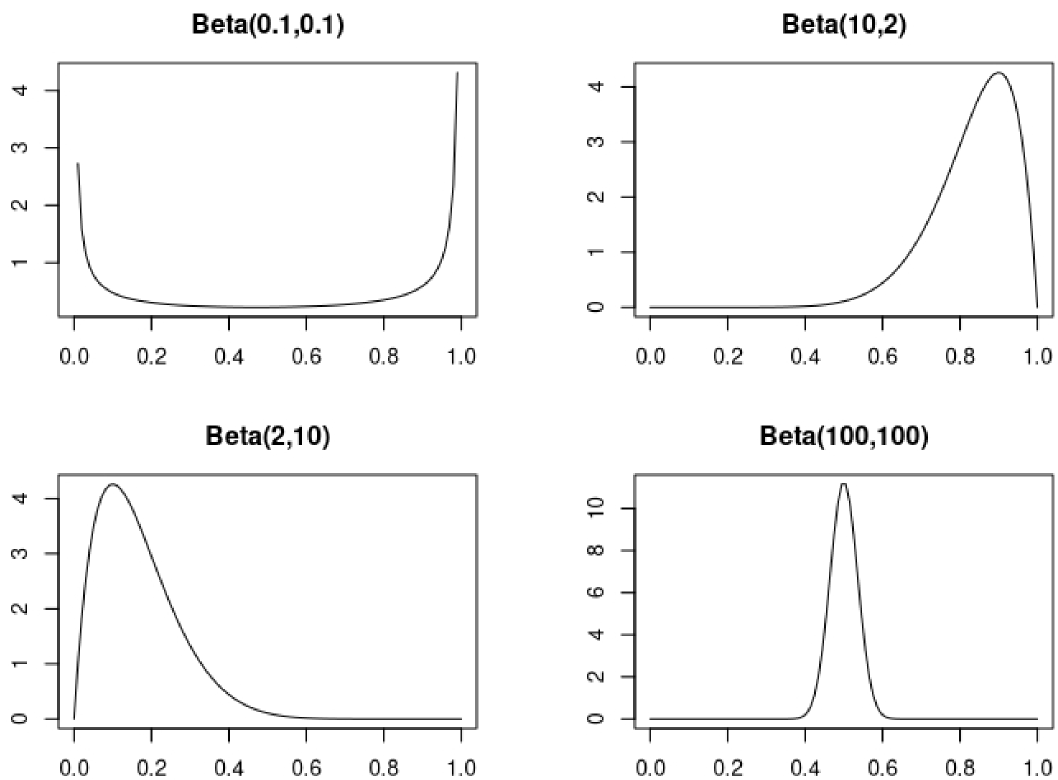

Figure 1: Selected Beta Distributions

[2] 2.

It is known that the uniform distribution is a special case of the beta distribution.

A. What are the numeric values for the uniform distribution?

B. What is the equivalent prior sample size for a U(0, 1) prior?

C. What is the equivalent prior sample size for a beta(10,15)prior?

A. The Beta Distribution that is equivalent to a uniform distribution is Beta(1,1)

B. The equivalent prior sample size for a Beta distribution is $\alpha + \beta - 2$. Given $U(0,1) = Beta(1,1)$, the equivalent sample size is 0. Uniform is a non-informative prior, it is as if you did not sample as all.

C. The equivalent prior samples size equals $\alpha + \beta - 2$, thus for Beta(10,15) it is 23.

During the severe floods in the Midwest in 2008, Iowa City and Coralville in Johnson County, Iowa, were hit hard and hundreds of homes, businesses, churches, and university buildings were destroyed. Less than a year later, a vote was held on a proposal to impose a local sales tax of one cent on the dollar to pay for flood-prevention and flood-mitigation projects. A few days before the actual vote, a local newspaper reported in its online edition:

"The outcome of Tuesday's local-option sales tax election in Johnson County appears too close to call, based on results from a Gazette Communications poll of voters.The telephone survey of 320 registered voters in Johnson County, conducted April 27–29, shows 40% in favor of the 4-year 1% sales tax. ."

A member of a local organization called "Ax the Tax" claims that this means that under half of all registered voters in the county support the local-option sales tax. She would like to use the sample survey data of the newspaper to test the two hypotheses:

$$H_0 : \pi \geq 0.5 \tag{1}$$

$$H_1 : \pi < 0.5 \tag{2}$$

where π represents the proportion of all Johnson County registered voters who support the sales tax. Let us practice on frequentist approach and Bayesian approach using this real problem.

[5] 3. Frequentist Approach:

A. Use calculus to derive the maximum likelihood estimator of the population proportion π is $\hat{\pi} = \frac{y}{n}$

B. Calculate the point estimate (MLE) of the population proportion, π, using the given information

C. Calculate a 95% confidence interval for the population proportion, π, and Interpret the 95% confidence interval in the context.

D. Test the claim at 5% significance level and state your decision

E. What is the interpretation ofthe p-value of the test you did in (D.).

A. Recall the probability mass function (PMF) for a binomial distribution.

$$p(y|n) = \binom{n}{y} \pi^y (1-\pi)^{n-y} \tag{3}$$

(**i**) Take the log of the PMF

$$log(\pi) = log\binom{n}{y} ylog(\pi) + (n-y)log(1-\pi) \tag{4}$$

(ii) Take the derivative of the log

$$\frac{dl(\pi)}{d\pi} = \frac{y}{n} - \frac{n-y}{1-n} \tag{5}$$

(iii) Set the derivative equal to 0 to find local minima/maxima

$$\frac{y}{n} - \frac{n-y}{1-n} = 0 \tag{6}$$

(iv) MLE estimate is thus equal to:

$$\hat{\pi} = \frac{y}{n} \tag{7}$$

(v) Ensure estimate is a maxima (negative second derivative).

$$\begin{aligned} \frac{d^2 l(\pi)}{d\pi^2} &= -\frac{y}{\pi^2} - \frac{n-y}{(1-\pi)^2} \\ &\frac{-n^3}{y(n-y)} < 0 \end{aligned} \tag{8}$$

B. In this example, the MLE estimate is:

$$\hat{\pi} = \frac{y}{n} = \frac{128}{320} = 0.40 \tag{9}$$

C. I utilize R to calculate the 95% confidence interval for *pi*.

```
binom.test(128, 320, p = 0.5, conf.level = 0.95)

95 percent confidence interval:
    0.3459083 0.4559608
```

Confidence intervals acknowledge that the MLE estimate is very unlikely to be the true population proportion. We can interpret this confidence interval as we calculate the distribution of proportion estimates for all possible samples of size n, and 95% of those estimates lie between 0.346 and 0.456.

D. Again, I utilize R, and test the hypothesis $H_0 : \pi \geq 0.5$, $H_1 : \pi < 0.5$

```
binom.test(128, 320, p = 0.5, alternative = "less",
           conf.level = 0.95)

p-value = 0.0002059
alternative hypothesis: true probability of success is
less than 0.5
```

E. The null hypothesis was that the true proportion π was greater than or equal to 0.5. *p-values* are based off the concept of repeated sampling and assume that the null hypothesis is true. The *p-value* is the probability, given H_0, of drawing a random sample with greater or equal evidence against the H_0 as we already have. Given the small *p-values*, the data that we have is *inconsistent* with our hypothesis that $\pi \geq 0.5$, which supports "Ax the Tax" claims that less than half of the voters support the tax.

[8] 4. Bayesian Approach:

A. Under a uniform (0,1) prior distribution

- **(i)** Show that the posterior distribution of the population proportion π,is Beta(129,193)
- **(ii)** Calculate the posterior mean, posterior mode, posterior median, and posterior variance of the population proportion, π.
- **(iii)** Calculate a 95% equal-tail posterior credible set and interpret it.
- **(iv)** Calculate $P(\pi \geq 0.5|y)$and $P(\pi < 0.5|y)$ and what is your conclusion?[i.e. is there significant evidence in support of hypothesis $H_a : \pi < 0.5$?]

B. Under a Beta(20,45) prior distribution

- **(i)** Use calculus to show that the posterior distribution of the population proportion pi,is Beta(148,237).
- **(ii)** Plot the likelihood function, prior distribution, and posterior distribution by R function triplot.
- **(iii)** Calculate a 95% equal-tail posterior credible set and interpret it.Which credible set is wider, under uniform or beta prior? why?
- **(iv)** Calculate $P(\pi \geq 0.5|y)$and $P(\pi < 0.5|y)$ and what is your conclusion?[i.e. is there significant evidence in support of hypothesis $H_a : \pi < 0.5$]

For both of these question, I use the methodology for calculating a posterior distribution from a $Beta(\alpha, \beta)$ prior and a dataset with y successes in a sample size n, as described on Slide 35 in Unit 4.

$$\begin{aligned} prior &= Beta(\alpha_{post}, \beta_{post}) \\ \alpha_{post} &= \alpha + y \\ \beta_{post} &= \beta + n - y \end{aligned} \tag{10}$$

A. Uniform Distribution prior

Recall that the Uniform Distribution is equivalent to a Beta(1,1) distribution.

- **(i)** Prove the Posterior distribution is Beta (129, 193):

$$\begin{aligned} \alpha_{post} &= \alpha_{prior} + y = 1 + 128 = 129 \\ \beta_{post} &= \beta_{prior} + n - y = 1 + 320 - 128 = 193 \end{aligned} \tag{11}$$

(ii) Posterior Mean, Posterior Mode, Poster Median, and Posterior variance

$$\begin{aligned}
E(\pi|y) &= \frac{\alpha+y}{\alpha+\beta+n} = \frac{1+128}{1+1+320} = 0.4006 \\
Mode(\pi|y) &= \frac{\alpha+y-1}{\beta+n-y-2} = \frac{1+128-1}{1+320-128-2} = 0.6702 \\
Median(\pi|y) &= \frac{\alpha+y}{\alpha+\beta+n-y} = \frac{1+128}{1+1+320-128} = 0.6649 \\
Variance(\pi|y) &= \frac{(\alpha+y)(\beta+n-y)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} \\
&= \frac{(1+128)(1+320-128)}{(1+1+320)^2(1+1+320+1)} = 0.000746
\end{aligned} \tag{12}$$

(iii) The 95% equal-tail posterior credible set can be calculated using R:

```
qbeta(c(0.025, 0.975), 129, 193)

[1] 0.3478036 0.4546081
```

Bayesian posterior credible sets have a very straightforward interpretation: with 95% probability, the true population proportion lines between 0.3478 and 0.4546.

(iv) $P(\pi \geq 0.5|y)$ and $P(\pi < 0.5|y)$ can be calculated using R. Here is $P(\pi \geq 0.5|y)$:

```
pbeta(.5, 129 , 193 , lower.tail = FALSE)

[1] 0.0001697076
```

Here is $P(\pi < 0.5|y)$, which can also be calculated as $1 - P(\pi \geq 0.5|y)$:

```
pbeta(.5, 129 , 193 , lower.tail = TRUE)

[1] 0.9998303
```

The probability of *pi* being equal or greater than 0.5 is very small. There is significant evidence to support $\pi < 0.5$.

B. Beta (20, 45) prior

(i) Prove the Posterior distribution is Beta (148,237)

$$\begin{aligned}
\alpha_{post} &= \alpha_{prior} + y = 20 + 128 = 148 \\
\beta_{post} &= \beta_{prior} + n - y = 45 + 320 - 128 = 237
\end{aligned} \tag{13}$$

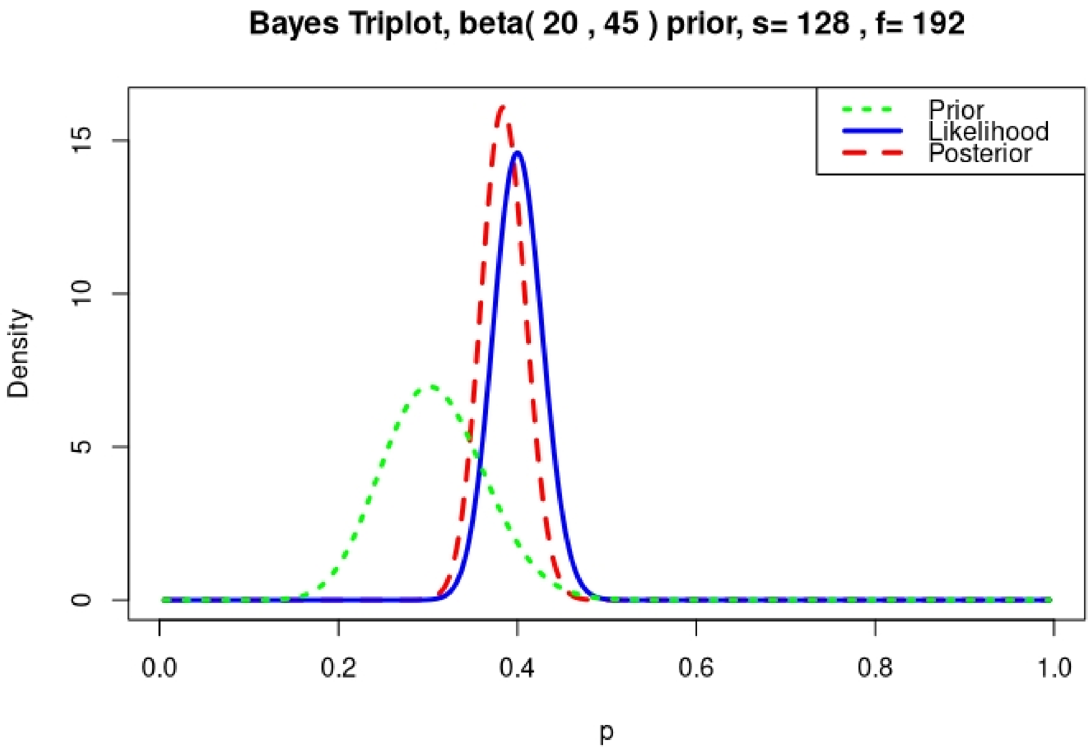

Figure 2: Triplot

(ii) Posterior Mean, Posterior Mode, Poster Median, and Posterior variance

$$\begin{aligned}
E(\pi|y) &= \frac{\alpha+y}{\alpha+\beta+n} = \frac{20+128}{20+45+320} = 0.3844 \\
Mode(\pi|y) &= \frac{\alpha+y-1}{\beta+n-y-2} = \frac{20+128-1}{45+320-128-2} = 0.7106 \\
Median(\pi|y) &= \frac{\alpha+y}{\alpha+\beta+n-y} = \frac{20+128}{20+45+320-128} = 0.6537 \\
Variance(\pi|y) &= \frac{(\alpha+y)(\beta+n-y)}{(\alpha+\beta+n)^2(\alpha+\beta+n+1)} \\
&= \frac{(20+128)(45+320-128)}{(20+45+320)^2(20+45+320+1)} = 0.0006959
\end{aligned} \tag{14}$$

(iii) The Triplot is plotted in Graph 2

(iv) The 95% equal-tail posterior credible set can be calculated using R:

```
qbeta(c(0.025, 0.975), 148 , 237)

[1] 0.3364848 0.4334839
```

With 95% probability, the true population proportion lines between 0.3365 and 0.4335. This credible set is *thinner* than that of the uniform prior, because our prior was informative and thus reduced uncertainty.

(v) $P(\pi \geq 0.5|y)$ and $P(\pi < 0.5|y)$ can be calculated using R. Here is $P(\pi \geq 0.5|y)$:

```
pbeta(.5, 148 , 237 , lower.tail = FALSE)

[1] 2.53955e-06
```

Here is $P(\pi < 0.5|y)$, which can also be calculated as $1 - P(\pi \geq 0.5|y)$:

```
pbeta(.5, 148 , 237 , lower.tail = TRUE)

[1] 0.9999975
```

The probability that π is more than 0.5 is very small, less than 1 in 1000, given our prior distribution and the survey results. The probability that π is less than 0.5 is very high, so it is appropriate for "Ax the Tax" to make such claims.

[2] 5. What are your conclusions, your observations, or your comments from all the analyses above in 1 and 2?

I take away a couple core things from the above problem. The first is that it is preferable to have an informative prior. My opinion is, if you are able to make any defensible estimation of the prior you should do so, because it adds value to the model. The second thing I take away is the value of having large sample sizes, which reduce model uncertainty. This is the same as in the frequentist framework. Additionally, I have realized that the effect that your prior estimates have on your posterior distribution is a function of how large your same size is. Finally, I have realized how important the Beta distribution is in a binomial framework. It is very easy to work with.

[2] 6. Using a uniform prior for the population proportion π and a random sample of $n = 320$ voters, $y = 128$ support the sales tax, suppose that the newspaper plans on taking a new survey of 25 voters, $n^\star = 25$. Let $y^\star$ denote the number in this new sample who support the sales tax.

A. Find the posterior predictive probability that $y^\star = 6$.

B. Find the 90% posterior predictive interval for $y^\star$. Hint: find the predictive probabilities for each of the possible values of $y^\star$and ordering them from largest probability to smallest probability. Then add the most probable values of $y^\star$ into your probability set one at a time until the total probability exceeds 0.90 for the first time.

Given our uniform prior, the correct distribution is Beta(129, 193). I use **R** and the package *LearnBayes* to answer these questions.

A. $P(y^\star = 6)$:

```
pbetap(c(129, 193), 25, 6)
```

```
[1] 0.0466728
```

The probability that 6 of the 25 respondents say they against the tax is 4.67%.

B. I employ the methodology suggested in the Hint, using R to calculate all probabilities for $y^\star$ from 0 to 25, ordering them, and looking at the cumulative probabilities.

```
l = pbetap(c(129, 193), 25, 0:25)
df <- data.frame("y" = 0:25, "Prob"=l)
df <- df[order(-df$Prob), ]
df$CumSum <- cumsum(df$Prob)
```

The 90% probability set ranges from 6 to 13. Table 1 includes the full results.

Table 1: $P(y^\star)$

$y^\star$	$P(y^\star)$	Cum. Sum.
10	0.155	0.155
9	0.146	0.301
11	0.142	0.444
8	0.118	0.562
12	0.113	0.674
7	0.081	0.755
13	0.078	0.833
6	0.047	0.880
14	0.046	0.926
15	0.024	0.950
5	0.022	0.972
16	0.011	0.983
4	0.008	0.991
17	0.004	0.995
3	0.002	0.998
18	0.001	0.999
2	0.001	0.999
19	0.000	1.000
20	0.000	1.000
1	0.000	1.000
21	0.000	1.000
0	0.000	1.000
22	0.000	1.000
23	0.000	1.000
24	0.000	1.000
25	0.000	1.000