1. *(24 pts) Describe the differences between Bayesian and classical inference. Include a discussion on confidence and credible intervals.*

Interval Estimates

Dr. Hoff describes Bayesian inference as, "the process of inductive learning via Bayes' rule" (which is different from the view Gelman and Shalizi (2013) take on Bayesian inference). Using Bayes' rule, we can make probablistic statements about parameters, after incorporating information from the sample. Dr. Hoff says that probabilities are synonomous with information. Bayesian inference subsets the prior information about the parameters into the space of the observed data, and concludes with updated information about the parameters in that subseted parameter space.

In Bayesian inference, we have to find a reasonable prior distribution, complete with it's parameters and know the distribution of the data conditioned on the parameters. In comparison to classical inference, where we do not need to know or guess prior distribution. A benefit of Bayesian inference is that we can make credible intervals for the parameters. The credibility of a credible interval is the probability that the parameter is in the interval.

To make inference about the values for the parameter that are plausible in classical inference, we make confidence intervals. The confidence of a confidence interval is the long run percentage of confidence intervals that are expected to contain the parameter. This means that if I take many, many samples and create corresponding confidence intervals for each sample, that in the long run, the confidence level is the percentage of confidence intervals that are expected to contain the true parameter.

Testing and Modeling

Gelman and Shelizi (2013) discuss classical statistics as "falsification". In classical inference, we make null and alternative hypotheses about parameters or models, and based on the probability of the observed result or more extreme under the null determine the strength of evidence against the null. In classical inference, we can not say that the null hypothesis is true, only the strength of evidence against the null. For example, we could have two hypothesis tests, (1) $H_o : \theta = 0$ and $H_a : \theta \neq 0$ and (2) $H_o : \theta = 0.5$ and $H_a : \theta \neq 0.5$. If both tests results in a high pvalue and suppose that in classical inference we could accept the null, we would conclude both that $\theta = 0$ and $\theta = 0.5$, but θ can only take on one value. In classical hypothesis testing, we can only say there was no evidence that another model or parameter was better than that assumed. This is versus probabilistic statements *for* models we can make using Bayesian inference. Gelman and Shalizi (2013) also discuss the importance of model checking in Bayesian inference, as seeing *where* the model may be less accurate versus having the end point goal of finding a posterior distribution. To me it seems that model checking would be important in both classical and bayesian inference.

2. *(1 pt) Do you consider yourself a Bayesian or classical Statistician?*

   For now, I like the idea of Bayesian inference because it is intuitive and realistic in how many people think. For example, it's hard to think about what sort of assurance we should put in a long run accuracy (confidence level) in a single case. Both classical and bayesian inference seem to have some subjectivity (sample size and priors).

3. *Assume you are hired by Bridger Bowl to compute the probability that an MSU student either skis (or snowboards).*

   (a) *(10 pts) If binary data are collected from 300 students, what is the sampling model for this research question? Please name the distribution and write out the corresponding sampling distribution.*

   WHERE $X_i = 1$ IF SKI/SNOWBOARD 0 ELSE

   Let x = the number of students that either ski or snowboard, then x $\sim$ BIN(300, θ).

   f(x|θ) = $\binom{300}{x}$ $\theta^x(1-\theta)^{300-x}$ $\cdot$I(x)$_{[0,1,\ldots,300]}$

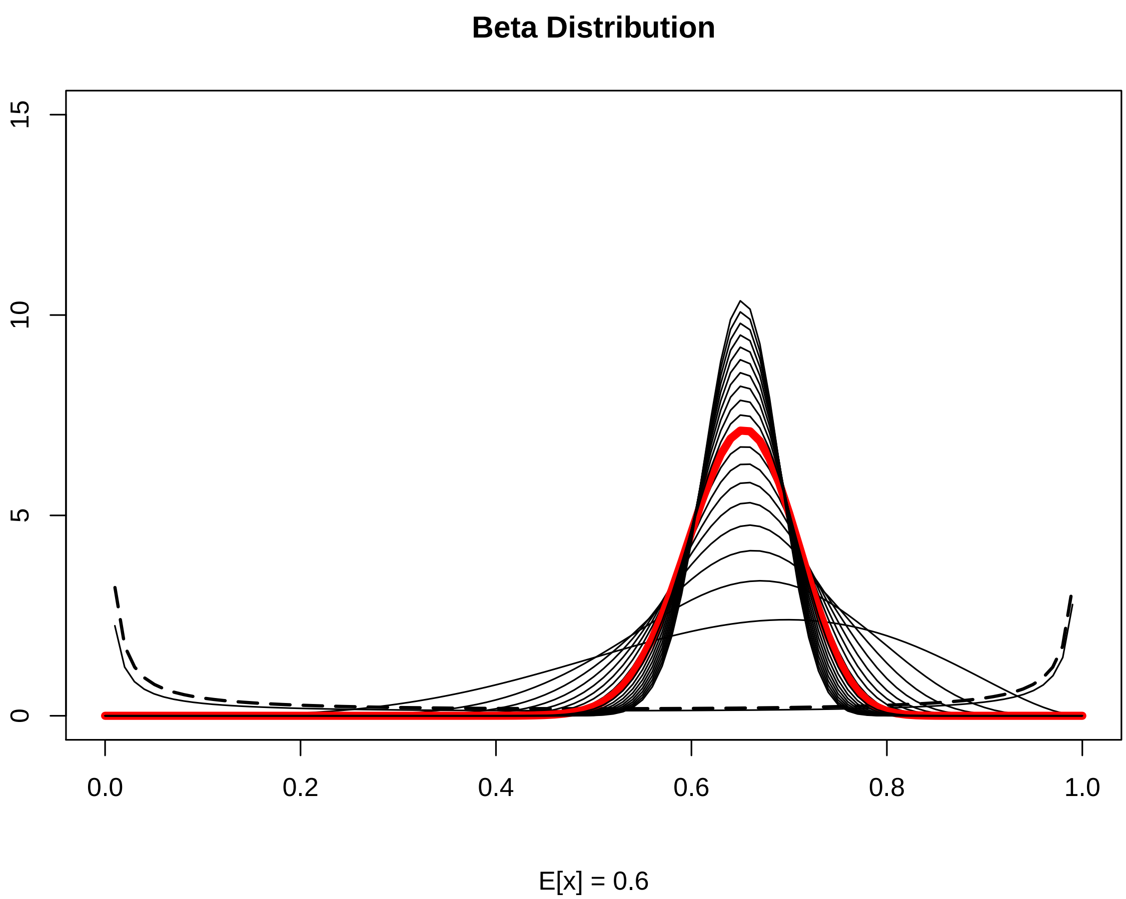
   (b) *(10 pts) Use a prior distribution from the Beta distribution and create a plot/histogram from this distribution. Why did you choose the α and β values for this prior distribution? Hint: R code*

   The hinted R code used values of 0.1 for both α and β. Using those values, that means we would expect half of the students to ski or snowboard and that the distribution of the proportion that ski or snowboard would be similar to the plot below.

   First, I would guess that more than half of MSU students have skiied or snowboarded before. I would guess this is around 65%. However, there are many distributions that would yield an expected 65% of skiiers or snowboarders. Below are plotted several with α and β combinations, all with an expected value of 0.65, with α and β ranging from 0.1 to 100 and from 0.0534 to 53.85, respectively. I came up with the α and β values by making an equally spaced sequence from 0.1 to 100 of length 20 α, solving for β that would make the expected proportion equal to 0.65, and then plotting each of these beta distributions.

   The distribution that I think visually is most reasonable is shown in red, with $\alpha = 47.42$ and $\beta =$ 25.53, although there is not much of a difference in the distributions with parameters near to these values. The distribution I chose suggests that on average, 65% of MSU students ski/snowboard, and it is most likely that between about 50% and 80% of MSU students ski/snowboard.

   The BETA(0.1,0.1) is plotted with the dashed line. The expected value of any beta distribution with $\alpha = \beta$ is symmetric about $\frac{1}{2}$. You might choose the parameters on the prior beta distribution to be equal if you believe that the probability distribution of θ taking on each of the possible values is symmetric about $\frac{1}{2}$. By setting the prior parameters both equal to 0.1, you are assuming it is most likely that either there is no chance of students skiing or snowboarding, or there is a high chance of students skiing or snowboarding.


Beta Distribution
15
10
5
0
0.0
0.2
0.4
0.6
0.8
1.0
E[x] = 0.6

Why to choose α and β: It is very rare that we know the parameter (here θ, the true probability of a student snowboarding or skiing). We know that for probabilities, the beta distribution, which has domain (0,1) is often reasonable. However, to use the beta distribution to model the probability of θ taking on certain values, we have to make more assumptions about the probability distribution of θ, we have to set the parameters of the beta distribution. In class we discussed that α can be interpretted as the expected number of students that ski or snowboard, and β is the expected total number of students, which is a place to start. However, many beta distributions have expectations of $\frac{1}{2}$, and just because we may expect the probability of a person skiing or snowboarding to be $\frac{1}{2}$, that does not tell us about the probability of θ falling in any interval. It is not clear to me about how to set these parameters or how to account for the added uncertainty with making assumptions about the prior parameters. Note that as we discussed in class, the BETA(1,1) is the UNIF(0,1) distribution which would mean essentially that the prior is uninformative about which intervals are likely for θ.

(c) *(15 pts) Assume 234 of the sample MSU students claim to either ski or snowboard. Compute the posterior distribution, $p(\theta|Y)$ where θ is the probability of MSU student skiing and Y is the observed responses.*

$\theta \sim$ BETA(α, β) and $Y|\theta \sim$ BIN(300,θ)

$p(\theta|Y) = \frac{p(Y|\theta)*p(\theta)}{p(Y)}$:

$p(Y) = \int_\theta \binom{300}{Y} \cdot \theta^Y \cdot (1-\theta)^{300-Y} \cdot \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta$

$= c^* \cdot \int_\theta \theta^{Y+\alpha-1}(1-\theta)^{300-Y+\beta-1} d\theta$

where $c^* = \binom{300}{Y} \cdot \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

then if $c^{**} = c^* \cdot \frac{\Gamma(\alpha+\beta+300)}{\Gamma(Y+\alpha)\Gamma(300-Y+\beta)}$

$p(Y) = c^{**} \cdot 1$

now, $p(\theta|Y) = \frac{p(Y|\theta)*p(\theta)}{p(Y)} = \frac{\binom{300}{Y} \cdot \theta^Y \cdot (1-\theta)^{300-Y} \cdot \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \cdot \theta^{\alpha-1}(1-\theta)^{\beta-1}}{\binom{300}{Y} \cdot \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \cdot \frac{\Gamma(\alpha+\beta+300)}{\Gamma(Y+\alpha)\Gamma(300-Y+\beta)}}$

$= \frac{\theta^{Y+\alpha-1} \cdot (1-\theta)^{300-Y+\beta-1}}{\frac{\Gamma(\alpha+\beta+300)}{\Gamma(Y+\alpha)\Gamma(300-Y+\beta)}}$

$= \frac{\Gamma(Y+\alpha)\Gamma(234-Y+\beta)}{\Gamma(\alpha+\beta+300)} \cdot \theta^{Y+\alpha-1} \cdot (1-\theta)^{300-Y+\beta-1}$

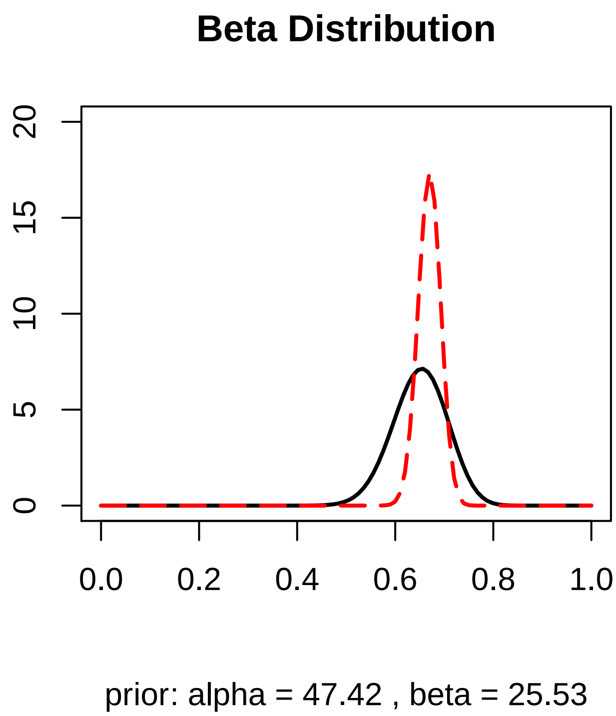
which is the pdf of a BETA$(Y+\alpha,\, 300-Y+\beta)$ distribution.

If the prior of θ is BETA(α, β), the posterior is BETA$(\alpha+x, \alpha+\beta+N-y)$ where N is the total sample size.

$\theta|x \sim BETA(47.42+234, 47.42+25.53+300-234) \implies \theta|x \sim BETA(281.42, 138.95)$

and again recall that $\theta \sim BETA(47.42, 25.53)$.

(d) *(10 pts) Plot the posterior distribution computed in part (c).* The posterior is plotted in black below, and the prior in red. The data change the beliefs about the posterior distribution substantially.



(e) *(10 pts) Compute a 95% credible interval for θ.*

There is a 95% chance that theta is between 0.5378032 and 0.7544716 .

4. *Hoff Exercise 2.1 - Marginal and conditional probability: The social mobility data from Section 2.5 gives a joint probability distribution on* (Y_1, Y_2) = *(father's occupation, son's occupation). Using this joint distribution, calculate the following distribution:*

   (a) *(5 pts) the marginal probability distribution of a father's occupation*

	farm	operatives	craftsmen	sales	professional	father.total
farm	0.018	0.035	0.031	0.008	0.018	0.110
operatives	0.002	0.112	0.064	0.032	0.069	0.279
craftsmen	0.001	0.066	0.094	0.032	0.084	0.277
sales	0.001	0.018	0.019	0.010	0.051	0.099
professional	0.001	0.029	0.032	0.043	0.130	0.235
son.total	0.023	0.260	0.240	0.125	0.352	1.000

	farm	operatives	craftsmen	sales	professional	else
father's marginal pdf	0.11	0.28	0.28	0.10	0.24	0.00

   (b) *(5 pts) the marginal probability distribution of a sons occupation*

	farm	operatives	craftsmen	sales	professional	else
son's marginal pdf	0.02	0.26	0.24	0.12	0.35	0.00

   (c) *(5 pts) the conditional distribution of a son's occupation, given that the father is a farmer*

	farm	operatives	craftsmen	sales	professional	else
son's pdf | father = farmer	0.164	0.318	0.282	0.073	0.164	0

   (d) *(5 pts) the conditional distribution of a father's occupation, given that the son is a farmer*

	farm	operatives	craftsmen	sales	professional	else
father's pdf | son = farmer	0.783	0.087	0.043	0.043	0.043	0