Assignment 3

Harry Bendekgey
Math 153: Bayesian Statistic

February 22 2018

1. Let p denote the proportion of registered voters in a large city who are in favor of a certain proposition. Suppose that the value of p is unknown, and two statisticians A and B assign to p the following different prior pdf?s: A's is Beta(2,1) and B's is Beta(4,1). In a random sample of 1000 registered voters from the city, it is found that 710 are in favor of the proposition.

a. Find the posterior distribution that each statistician assigns to p.

We have shown in class that the posterior distribution is defined by Beta($\alpha_0 + x, \beta_0 + n - x$). Thus, the distribution for satisticians A and B are Beta$(712, 291)$ and Beta$(714, 291)$ respectively.

b. Find the Bayes estimate for each statistician based on the squared error loss function.

The expected value of a Beta distribution is given by $\frac{\alpha}{\psi}$. Thus the Bayes estimates for statisticians A and B are $\frac{712}{1003}$ and $\frac{714}{1005}$ respectively.

c. Show that after the opinions of 1000 registered voters in the random sample had been obtained, the Bayes estimates for the two statisticians could not possibly differ by more than .002, regardless of the number in the sample who were in favor of the proposition.

Let x be defined as the number of respondents who were in favor of the proposition. The Bayes estimates for satisticians A and B will therefore be $\frac{x+2}{1003}$ and $\frac{x+4}{1005}$ respectively. Thus their difference is given by $\frac{x+4}{1005} - \frac{x+2}{1003} = \left(\frac{4}{1005} - \frac{2}{1003}\right) + \left(\frac{x}{1005} - \frac{x}{1003}\right)$. The first term is between 0 and .002. The second term is equal to $\frac{-2x}{1005 \cdot 1003}$ meaning it's bounded from below by $\frac{-2000}{1005 \cdot 1003}$ which is greater than -0.002. Thus it's bounded between -0.002 and 0. Thus the absolute difference between the estimators, which is the absolute value of the sum of these two terms, cannot be any bigger than 0.002.

2. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[0, \theta]$, where the value of the parameter θ is unknown. Recall the pdf of this distribution is $f(x) = \frac{1}{\theta}\mathbf{1}(x \in [0, \theta])$. Suppose that the prior distribution of θ is the Pareto distribution with parameters $x_0 (> 0)$ and

$\alpha(> 0)$, as defined below.

$$f(\theta|x_o, \alpha) = \frac{\alpha x_0^\alpha}{\theta^{\alpha+1}} \mathbf{1}(\theta \geq x_0)$$

a. If the value of θ is to be estimated by using the squared error loss function, what is the Bayes estimator of θ?

Using Bayes Rule, we know that the posterior is proportional to the likelihood times the prior. the likelihood of seeing some data $x_1, \ldots, x_n$ is given by $\frac{1}{\theta^n}\mathbf{1}(\theta \geq x_{(n)})$. Thus we get:

$$P(\theta|X) \propto \frac{1}{\theta^n}\mathbf{1}(\theta \geq x_{(n)}) \cdot \frac{\alpha x_0^\alpha}{\theta^{\alpha+1}}\mathbf{1}(\theta \geq x_0) = \frac{\alpha x_0^\alpha}{\theta^{\alpha+n+1}}\mathbf{1}(\theta \geq x_0, x_{(n)})$$

Throwing away constants, we can see that this is another Pareto distirbution with $\alpha = \alpha + n$ and $x_0 = \max(x_0, x_{(n)})$ We calculate the expected value of the Pareto distribution:

$$\int_{x_0}^{\infty} \frac{\alpha x_0^\alpha}{\theta^\alpha} \partial\theta = \frac{\alpha x_0^\alpha}{\theta^{\alpha-1}(1-\alpha)}\Big|_{x_0}^{\infty} = 0 - \frac{\alpha x_0^\alpha}{x_0^{\alpha-1}(1-\alpha)} = \frac{\alpha x_0}{\alpha - 1}$$

Plugging in our posterior values for α, x_0, we get:

$$\frac{(\alpha + n)\max(x_0, x_{(n)})}{\alpha + n - 1}$$

b. Argue that the frequentist's estimator, the MLE, is given as $\hat{\theta}_{MLE} = \max\{X_i\}$. Note that you can't use calculus to maximize this function, as the likelihood is not continuous. Draw it to see the location of it's maximum.

As I pointed out before, likelihood is given by: $\frac{1}{\theta^n}\mathbf{1}(\theta \geq x_{(n)})$. This is a decreasing function within the bounds of the indicator function. Thus we want the smallest number in the indicator, or $x_(n)$, because this is the point that maximizes likelihood.

c. Use the fact that the CDF of the maximum, $F_n(x) = P(\max\{X_i\} < x) = P(\text{all } X_i < x) = F(x)^n$ to derive the pdf of the MLE (just differentiate). Notice that $\frac{x}{\theta}$ lives on [0,1], and thus state the (named) distribution of $\frac{\max\{X_i\}}{\theta}$. Use known facts of this distribution to create an unbiased estimator (that I actually got wrong in lecture, if only slightly). (This feels very frequentist, but the trick to avoid integrating is classic Bayesian.)

First we note that $F(x) = \frac{x}{\theta}$. Thus $F_n(x) = \left(\frac{x}{\theta}\right)^n$. If we differentiate it to get its pdf, we get: $p_n(x) = \frac{n}{\theta}\left(\frac{x}{\theta}\right)^{n-1}$. We recognize this as the kernel of a Beta distribution for $\frac{x}{\theta}$ with $\alpha = n, \beta = 1$. Thus $\frac{x_{(n)}}{\theta} \sim$ Beta $(n, 1)$.

We know therefore that $E\left[\frac{x_{(n)}}{\theta}\right] = \frac{n}{n+1}$. θ is just a constant, so we rearrange to get:

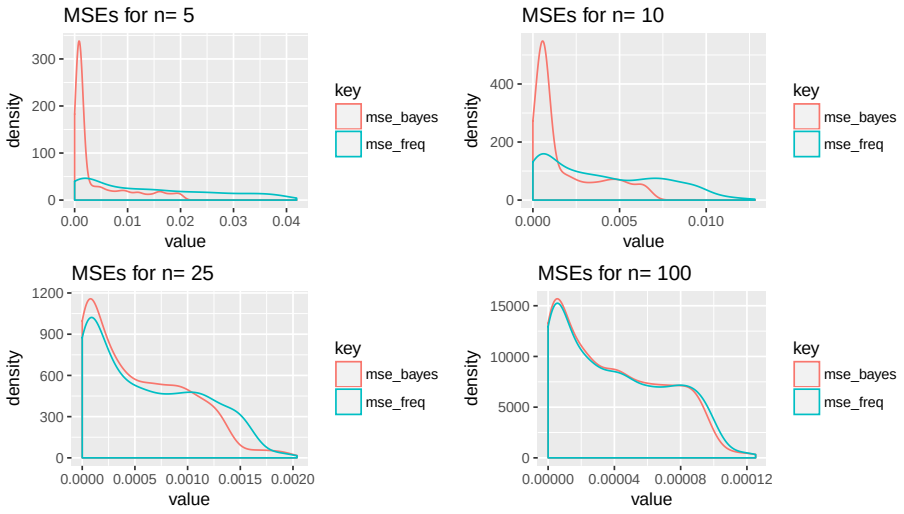
$$E\left[\frac{n+1}{n} x_{(n)}\right] = \theta$$

This is our unbiased estimator.

d. Pick a reasonable prior, and compare the behavior of the two estimators with respect to MSE.
The MLE estimator is $\frac{n+1}{n}x_{(n)}$ which we note is a case of the Bayesian estimator, $\frac{(\alpha+n)\max(x_0,x_{(n)})}{\alpha+n-1}$, namely $\alpha = 1, x_0 = 0$. Let's say however that we're certain $\theta \geq 0.9$, and we feel pretty confident it's between that and ~ 1.1. Then let's set a slightly higher α, perhaps 3. Let's test:

```
compare_mse <- function(n,a=3,x0=0.9,theta=1) {
  mse_freq <- c()
  mse_bayes <- c()
  for (i in 1:1000) {
    x <- max(runif(n,min=0,max=1))
    mse_freq[i] <- ((n+1)/n * x - theta)^2
    mse_bayes[i] <- ((n+a)/(n+a-1) * max(x,x0) - theta)^2
  }
  df <- tidyr::gather(data.frame(mse_freq,mse_bayes))
  return (ggplot(df, aes(x=value, color=key)) +
            geom_density() +
            scale_x_continuous(limits = c(0,mean(mse_freq) * 1.5))+
            labs(title=paste("MSEs for n=", n)))
}
```

```
ggarrange(compare_mse(5),compare_mse(10),
          compare_mse(25),compare_mse(100))
```



This is exactly as we expect. because our prior was "smart". At small sample sizes, our prior helps us get really small squared errors. At large sample

sizes, the frequentist estimator converges.

e. Why do we need to be extra careful in selecting a prior in this situation?
The Pareto distirbution assigns 0 values to certain values of θ. This is dangerous because this makes it possible that our posterior never acknowledges the overwhelming evidence of the data, and we refuse to consider the true answer as possible. In the case of the previous question, if the true value of θ had been less than 0.9, we would've been in a lot of trouble

3. Look at the 2001 article "*Interval Estimation for a Binomial Proportion*" by Brown et.al. (you can get away with mostly looking at the pictures to get a sense of what's going on).

a. See if the Bayesian credible interval for p (don't worry about using the HDI, any one will do) suffers from the same problem.
The key problems of the frequentist confidence interval are

- we must approximate $\hat{p}$ for p because we don't have a pivot statistic, and
- we are using a continuous distribution (the normal) to approximate a discrete distribution

The Bayesian estimator suffers from neither of these problems, because the posterior distribution was always continuous, and we are simply taking quantiles of it, instead of trying to take quantiles of a distribution of the data. (which is not only different than the parameter but not even continuously distributed)

b. Does the credible interval attain it's stated coverage probability? The statement for a confidence interval is that we are using a method that, should we repeat the experiment indefinitely, 95% (for example) of the resulting (95%) intervals will cover the target, in this case p (the above article shows that is not quite true, but this is due to approximations). Is this the same sense in which the credible interval works (repeating experiments)? Or is it in the sense of the coverage probability over a lifetime of experiments (for "appropriately" chosen prior as discussed in Lab 2)?

```
calc_coverage_prob <- function(p,n,a=10,b=10) {
  coverage <- c()
  for (j in 1:10) {
    coverage[j] <- 0
    for (i in 1:10000) {
      x <- rbinom(1,n,p)
      if (p <= qbeta(.975, a + x, b + n -x) &
          p >= qbeta(.025, a + x, b + n -x)) {
            coverage[j] <- coverage[j] + 1
```

```
          }
    }
    coverage[j] <- coverage[j]/10000
  }
  return(coverage)
}
```

Here, we're considering a prior of $\alpha = \beta = 10$. First, let's consider a huge sample size. Because Bayesians and Frequentists are hard to distinguish at large sample sizes, we expect that the frequentist statement holds: if we repeat the experiment indefinitely we will capture it 95% of the time. That is, we expect the coverage probability for $p = 0.5$ to be roughly 95%

```
calc_coverage_prob(0.5, 1000)
```

```
##  [1] 0.955 0.956 0.953 0.953 0.949 0.955 0.949 0.952 0.951 0.952
```

We get what we expect! But what happens at small sample sizes?

```
calc_coverage_prob(0.5, 10)
```

```
##  [1] 1 1 1 1 1 1 1 1 1 1
```

This is actually unsurprising. We're pretty sure at the start that the answer is 0.5 so we capture it in our credible interval every single time.

But even if we buy the definition given of an "appropriately chosen prior", then we know that p will often times be something other than the center of the prior. Consider $p = 0.68$:

```
pbeta(0.68,10,10)
```

```
## [1] 0.95
```

It's at the 95^{th} percentile of our prior, so if it was an appropriately chosen prior, we would expect to have a true value of p as extreme as this or more, 10% of the time. What is our coverage probability under this model:

```
calc_coverage_prob(0.68, 10)
```

```
##  [1] 0.819 0.814 0.814 0.816 0.818 0.811 0.814 0.807 0.805 0.816
```

This isn't quite what we want (or at least, not what a frequentist would want). Our prior was appropriately chosen, but it's off center. If we repeat this experiment over and over again we're going to get poor coverage compared to what we want.

What we do get, however, is the second statement. If the value of p really is distributed according to the prior (whatever that means to us), then after seeing

data the possible values of p will be distributed according to the posterior, and the credible interval will capture it 95% of the time. Thus over our lifetime of science we will be right about our 95% intervals 95% of the time. We can't make the statement about repeating the same experiment over and over again because when we repeat the experiment we don't get to resample a true value of p from our prior distribution.

And we don't really care about what the result would be if we repeated it over and over again because we only did it once. All we really care about is that if we're saying we're 95% confident we're right, well 95% of the times we say that we will be right.