Homework 5

Kalvin Goode, Amil Khan

Due on November 18, 2018 at 11:59 pm

Note: If you are working with a partner, please submit only one homework per group with both names and whether you are taking the course for graduate credit or not. Submit your Rmarkdown (.Rmd) and the compiled pdf on Gauchospace.

Problem 1

Frequentist Coverage of The Bayesian Posterior Interval. In quiz 1 we explored the importance and difficulty of well-calibrated prior distributions by examining the calibration of subjective intervals. Suppose that $y_1, .., y_n$ is an IID sample from a $Normal(\mu, 1)$. We wish to estimate μ.

Part a.)

For Bayesian inference, we will assume the prior distribution $\mu \sim Normal(0, \frac{1}{\kappa_0})$ for all parts below. State the posterior distribution of μ given $y_1, .., y_n$, and the 95% quantile-based posterior credible interval for μ.

Solution.

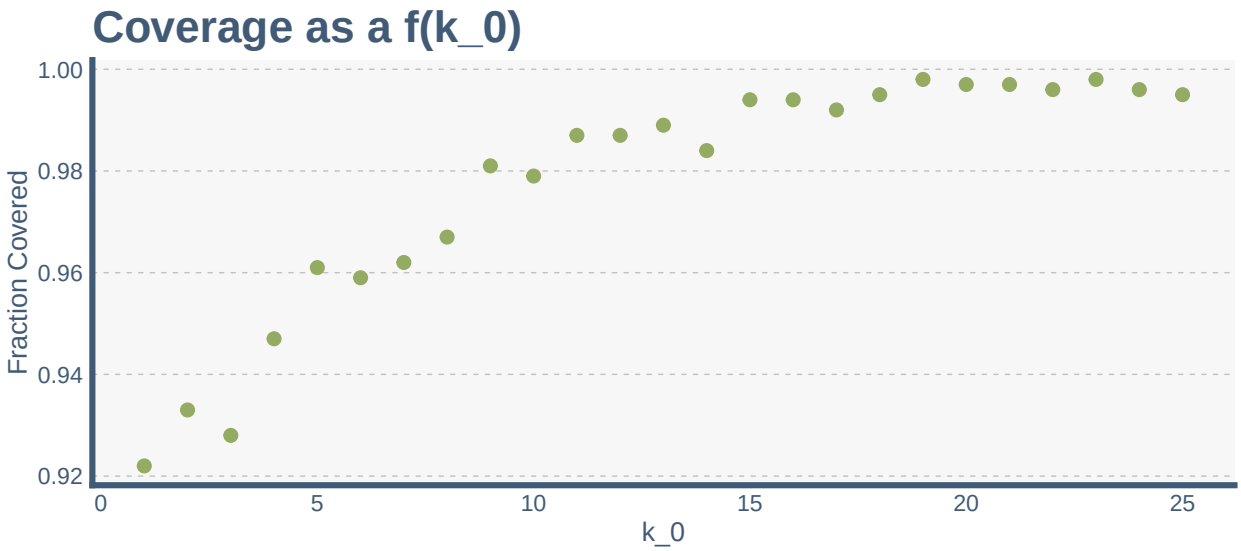
Prior Distribution: $\mu \sim Normal(0, \frac{1}{\kappa_0})$

Posterior Distribution: $\mu \mid y_1, \ldots, y_n \sim \mathrm{N}\left(0, \frac{1}{\kappa_n}\right)$

Quantile-based Posterior Credible Interval for $\mu : \bar{\mu} \pm 1.96\sqrt{\frac{\bar{\mu}(1-\bar{\mu})}{n}}$

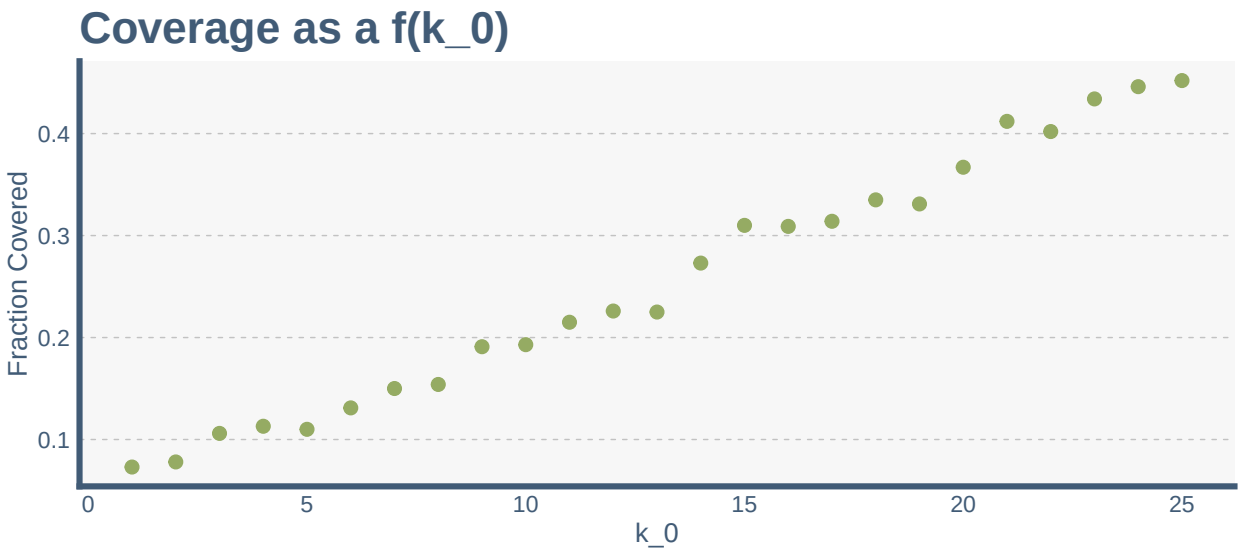
Part b.)

Now we will evaluate the frequentist coverage of the credible interval on simulated data. Generate 1000 data sets where the true value of $\mu = 0$ and $n = 10$. For each dataset, compute the posterior 95% interval and see if it covers the true value of $\mu = 0$. Compute the frequentist coverage as the fraction of these 1000 posterior 95% credible intervals that contain $\mu = 0$. Do this for each value of $\kappa_0 = 1, 2, ..., 25$. Plot the coverage as a function of κ_0.



Part c.)

Repeat the previous part but now generate data assuming the true $\mu = 1$.



Part d.)

Explain the differences between the two plots. For what values of κ_0 do you see closer to nominal coverage (i.e. 95%)? For what values does your posterior interval tend to overcover (the interval covers the true value more than 95% of the time)? Undercover (the interval covers the true value less than 95% of the time)? Why does this make sense?

Solution.

In the first plot, we see our coverage for the true value $\mu = 0$ plateaus around $\kappa_0 = 15$. Values below $\kappa_0 = 5$ cover the true value less than 95% of the time, while values above $\kappa_0 = 6$ overcover the true value. Conversely, for the true value $\mu = 1$, we see that our coverage of the true value increases as we increase the number of κ_0.

Problem 2

Modeling Election Outcomes. On November 4, 2014 residents of Kansas voted to elect a member of the United States Senate to represent the state. After the primaries, there were four major contenders in the race: 1) Republican incumbent Pat Roberts, 2) Democrat Chad Taylor, 3) Independent Greg Orman, and 4) Liberatarian Randall Batson.

For this problem we will reference polling data that can be found here:

http://en.wikipedia.org/wiki/United_States_Senate_election_in_Kansas,_2014#Polling_3

In mid-August 2014 a SurveyUSA poll of 560 people found the following vote preferences:

Pat Roberts	Chad Taylor	Greg Orman	Randall Batson	Undecided
37%	32%	20%	4%	7%

Ignoring the "undecided votes", the maximum likelihood estimate for the true vote shares of each candidate assuming, assuming a multinomial distribution over the 4 candidates, is simply the fraction of people.

Part a.)

Assume that you first interview the 7% of undecided voters. They claim they are equally likely to vote for any of the four candidates. Before reviewing the other survey data, you decide to use this information to construct a prior distribution for the true vote shares of the four candidates. What is the prior distribution and what are it's parameters (think pseudocounts)? Given the survey data above and the prior, specify the posterior distribution for the vote shares of the four candidates and the parameters of this distribution.

Solution.

```
## Our assumption for prior distribution is
## Dirichlet( 9.8 , 9.8 , 9.8 , 9.8 ).
## Then, the posterior distribution is Dirichlet( 217 , 189 , 121.8 , 32.2 ).
```

Part b.)

On September 3, 2014 Democratic nominee Chad Taylor withdrew from the race. Assume that amongst those who said they would vote for Taylor in the August survey, 70% of them changed their vote to Orman, 20% to the Libertarian, Baston, and the remaining 10% for Roberts. The above information should be used construct a new prior distribution for the 3-candidate race again assuming that the undecided voters from the August poll will now vote equally among the remaining three candidates. Calculate the new posterior distribution over the vote shares for the 3 remaining candidates. Use Monte Carlo to find the posterior probability that more people in Kansas support Pat Roberts than Greg Orman.

Solution.

```
## Our assumption for prior distribution is
## Dirichlet( 27.72 , 135.24 , 45.64 ).
## The posterior distribution is Dirichlet( 234.92 , 247.24 , 68.04 ).
## In posterior distribution, the probability that more people in
## Kansas support Pat Roberts than Greg Orman is  0.27466 .
```

Part c.)

From October 22-26, 2014 SurveyUSA released a poll of 623 found the following preferences:

Pat Roberts	Chad Taylor	Greg Orman	Randall Batson	Undecided
42%	–	44%	4%	10%

Use the posterior from the previous part as the prior for this new survey. Compute a new posterior given the new survey data above. Assume that the population consists of 100,000 eligible voters. However, not all eligible voters actually vote. In fact, roughly between 30-50% and eligible voters actually turn out in a midterm election https://www.fairvote.org/voter_turnout#voter_turnout_101. You express your uncertainty by assuming that the fraction of eligible voters who actually turn out is a Beta(40, 60) random variable. Assuming a random sample of eligible voters actually turn out, generate 10000 samples from the posterior predictive distribution. Use Monte Carlo to answer the following questions.

Part C.1

Greg Orman's team believes that if they can get at least 20000 votes they will win the election. What is the posterior predictive probability that Greg Orman receives at least 20000 votes *and* wins the election?

Solution.

```
## The posterior pridictive probility predicts that Greg Orman receives
## at least 20000 votes and wins the election is 0.1423
```

Part C.2

Both leading candidates fear that the third party vote is taking away potential supporters. What is the posterior predictive probability that the difference between Greg Orman and Pat Roberts is smaller than then vote total for Randall Batson.

Solution.

```
## The posterior predictive probility predicts that the difference between Greg Orman
## and Pat Roberts is smaller than then vote total for Randall Batson is 0.6628
```

Part d.)

Discuss the assumptions that were made to generate your predictions. If you think some assumptions were poor, how might you change the model to improve upon them? This is an open-ended question with no right or wrong answers and will be graded on thoughtfullness and effort only.

Solution.
One assumption is that each person voting for the candidate is identical independent distribution. We can improve model by studying the historical preferred party for any voter and also review historical rate of voting within a region.