Gov 2001: Problem Set 3

Due Wednesday, February 24th by 6pm

Important Reminder:

The deadline for choosing a paper to replicate is **Wednesday, February 24, 2014 at 6pm**. On or before February 24, submit the information about your article to the Quiz section on Canvas. We will give you feedback on your paper. Include within your email a brief note (approx. 2-5 sentences) explaining why this paper is a good choice.

Instructions

You should submit your answers and R code to the problems below using the Quizzes section on Canvas.

Likelihood Inference

Problem 1

You are analyzing the number of trips each member of the House of Representatives makes to their home district in a given month. You start by researching the Representatives from Massachusetts and find that Mike Capuano made six trips to his home district in January, that is $y = 6$. You want to know the average rate at which Representatives return home, which is represented by λ. Assume that $Y_i \sim Poisson(\lambda_i)$.

1.A If we assume that $\lambda = 5$, what is the probability of observing this data point?

1.B What is the likelihood, $L(\lambda|y = 6)$, for the above example? Do not assume that $\lambda = 5$.

1.C) Write a function for this likelihood in R. Use the curve() function to plot the likelihood curve over the range of λ from 0 to 15. On the x-axis you should have λ and on the y-axis you should have the value of the likelihood. Be sure you set the limits of your x axis so you can see the whole curve. Submit your plot below.

1.D) Using your plot from part C, provide a rough estimate of value of λ that maximizes the likelihood function. (Note: There is no need to use calculus here; an eye-ball estimate from your graph is fine.)

Bayesian Inference

Problem 2

Continuing with Problem 1, you still want to find the average rate at which Representatives return home, but now you want introduce some prior information. You believe that the prior distribution of λ is given by

$$p(\lambda|\theta, k) = \frac{1}{\Gamma(k)\theta^k}\lambda^{k-1}e^{-\frac{\lambda}{\theta}}$$

. Assume that θ and k are known constants. $\Gamma(n)$ denotes the Gamma function, which for integers is equal to $(n-1)!$ and can be evaluated in R using the function `gamma`.

2.A Make a graph of the density of the prior distribution of λ. Your x-axis should range from 0 to 15. Assume $\theta = 2$ and $k = \frac{1}{2}$.

2.B What is the posterior density of λ, $P(\lambda|y, \theta, k)$? You do not need to simplify this completely, although you can remove any additive or multiplicative constants.

2.C Recreate the plot of the likelihood curve exactly as you did in question 1C. Now add the curve of the posterior density from 2B onto the same exact graph. You should have a figure with one graph and two curves (not two graphs side by side). To do this you'll need to write a function for the posterior density, and you should assume that Assume $\theta = 2$ and $k = \frac{1}{2}$.

2.D Other than the functions having potentially different heights, what do you notice about the maximum of the two curves? What do you think is going on?

2.E What is the probability that λ is between 4 and 6 given our posterior distribution and that $\theta = 2$, $k = \frac{1}{2}$ and $y = 6$? Keep in mind that the posterior function you wrote probably is proportional (not equal) to $P(\lambda|y, \theta, k)$, so to get the probability you'll need to find the normalizing constant, that is, divide by the total density of the posterior across the whole support of λ. (hint: use `integrate`)

Bayes Theorem

Problem 3

You've been contacted by an international election monitoring organization to help them investigate allegations of fraud in a foreign election. Specifically, the organization has that 30% of precincts reported fraudulent election returns and wants your help determining whether unusual returns from another precinct are the result of fraud.
The organization provides you with a few facts:

- The probability of fraud in any given precinct is 30%.
- The probability of getting unusual looking election returns when there is fraud is 99%.
- The probability of getting unusual looking election returns when there is no fraud is 25%.

3.A What is the unconditional probability of having unusual looking election results?

3.B What is the prior probability of fraud occurring at this precinct?

3.C We know that the election returns from this new precinct look unusual. Given this, what is the probability these unusual election returns were actually the result of fraud?

3.D Imagine that the organization provides you with election return and incidence of fraud data from all the other precincts in the country. You are going to use the data to build an inference model (Bayes or likelihood). You first need to decide the probability distribution that generated the election results for the candidates in each precinct. Which distribution would you choose? And what are some of the assumptions you'd be making by choosing this distribution?

Neyman-Pearson Hypothesis Testing

Problem 4

Suppose you are estimating a mean from a population. You sample n independent and identically distributed observations $X_1, \ldots, X_n$ where for all i, $E[X_i] = \mu$ and $Var(X_i) = 9$.

4.A Assuming that n is sufficiently large, what is the approximate distribution, mean and variance of the sampling distribution of your sample mean $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

4.B You want to test the null hypothesis $H_0 : \mu = 0$ against the alternative hypothesis $H_A : \mu \neq 0$ with a sample of 30 observations. Load in the dataset `p4_1.csv` into R, compute your estimated sample mean, test statistic (assuming you know $Var(X_i) = 9$), and p-value for a large sample two-sided z-test of the null. Do you reject the null hypothesis at the $\alpha = .05$ level?

4.C Suppose that the true population μ is 0.1. Given that you find a statistically significant estimate (that is, p-value $< .05$), what is the probability that the point estimate is of the wrong sign (that is, negative)? Note: You can compute this analytically (using the normal CDF in R (`qnorm`) or by simulation

4.D Now suppose that instead of having 30 observations, you have 300 (we still have the true $\mu = .1$). Given that you find a statistically significant estimate (that is, p-value $< .05$), what is the probability that the point estimate is of the wrong sign (that is, negative)? Again, you can compute this analytically or by simulation