Gov 2002: Problem Set 4

Spring 2023

Problem Set Instructions

This problem set is due on **February 22, 11:59 pm** Eastern time. Please upload a PDF of your solutions to Gradescope. We will accept hand-written solutions but we strongly advise you to typeset your answers in Rmarkdown. Please list the names of other students you worked with on this problem set.

Question 1 (24 points)

Suppose you're interested in studying the distribution of political ideology in the US, a random variable that we'll call X. Individuals are placed on a continuous one-dimensional ideology scale that varies from -1 to 1, where lower score are more liberal. Instead of having to do sampling to estimate the distribution, Keith Poole[1] comes to you in a dream and tells you the unnormalized distribution of this random variable, which is as follows:

$$f(x) = \begin{cases} c(\frac{1}{2}x + 1) & -1 \le x \le 1, \\ 0 & \text{else} \end{cases}$$

where c is a normalizing constant[2].

a. Find the value of c that would make $f(x)$ a valid probability distribution function, $f_X(x)$.

b. Calculate $E[X]$.

c. Find the cumulative density function of X.

d. Calculate $P(X > 0)$.

Question 2 (20 points)

Let's say you wanted to verify your answers to the above questions using simulation techniques. Since X does not follow a standard probability distribution e.g. a normal, exponential, or uniform distribution, there is no function built in to R to generate random samples from it

---

[1]One of the creators of NOMINATE, a method for estimating Congressional ideology.

[2]Note that this distribution is completely made up and is not based on actual data.

like `rnorm()`, `rexp()`, or `runif()`. Fortunately, we can generate samples from any probability distribution for which there is a closed-form CDF using what is called "inversion sampling" or "inverse transform sampling."

To do this, we take advantage of two things: 1) the fact that all computing languages can generate pseudo-random numbers from the uniform distribution over the interval $[0, 1]$ and 2) the fact that the CDF of X (and other random variables generally) is a function that maps the support of X to probabilities between zero and one.

a. First, find the inverse of CDF F_X which you found in Question 1. This function, denoted F_X^{-1} or Q, is called the quantile function. The quantile function of a random variable specifies the value of a random variable such that the probability of the variable being less than or equal to that value equals the given quantile.

b. Generate 10,000 draws from a standard uniform distribution. These values can be thought of as probabilities. Plug these values into the function Q that you just derived. By doing so, you generate a sample of size 10,000 from the random variable X as specified in Question 1.

c. Verify your answers to Questions 1b and 1d by estimating $E[X]$ and $P(X > 0)$. Be sure to set your seed to `02138`.

Question 3 (20 points)

Suppose $X \sim \text{Pois}(\lambda)$, where λ is fixed but unknown.

An estimator is a function of the data and the **bias** of an estimator, $f(X)$, is defined as $E[f(X)] - \theta$, where θ is the **estimand** (an unknown quantity we would like to estimate from the observable data).

For instance our estimand could be λ, and we know by the properties of a Poisson random variable that the bias of the estimator, $f(X) = X$, is $E[X] - \lambda = \lambda - \lambda = 0$. We call an estimator with 0 bias an **unbiased** estimator.

For this question, suppose that our estimand is λ^3 rather than λ.

(a) Show that X^3 is *not* an unbiased estimator of λ^3 and specify the bias as a function of λ.

Hints:

1. You may use the following result: if $X \sim \text{Pois}(\lambda)$, then $E[X \cdot g(X)] = \lambda E[g(X + 1)]$ for any function $g(\cdot)$.

2. You may use the result for $E[X^2]$ derived in lecture and section, (i.e., no need to derive it again).

(b) Suppose $\lambda = 5$. Use $150,000$ simulations to validate your result to part (a). That is, calculate the bias of you estimator from both the simulations and the analytical results. Print your results in the format below:

```
set.seed(02138)

# Simulation result

# Analytical result
```

Question 4: Fisher's method in forecasting (30 points)

In this problem, we're going to explore a real-world example of Fisher's "lady tasting tea" experiment from lecture: election forecasters – who have, for better or worse, become a big part of politics in the United States and elsewhere. Be sure to set your seed: `set.seed(02138)`.

(a) Suppose that Bob has correctly predicted six of the last eight election outcomes. What is the probability that someone randomly flipping a coin each of the same elections would have experienced the same success as Bob? Use `R` to compute your answer analytically (i.e. not by simulation).

Forecasting has become so popular that riffraff are flooding the market. These "uniform amateurs" predict the vote share for each state in the U.S. presidential election by drawing a uniform random variable between 0 and 1, independently across states. You are deciding whether or not to hire a forecaster, Nate, to to forecast each of the 50 state election winners in the 2024 presidential general election based on the performance of his 2020 election forecast, but you are worried that Nate might be one of these amateurs. When you ask him to justify his 2020 forecasts, he says "my highest predicted [Democratic] vote share was 0.8 which is very unlikely if I were a uniform amateur." Let's evaluate his claim.

(b) Suppose Nate is a uniform amateur and let X be the maximum of the 51 uniform vote share draws (include D.C.). Derive the CDF and PDF of X. Use these to calculate the probability of Nate's highest Democratic vote share being 0.8 or less if he were a uniform amateur.

(c) To be on the lookout for more uniform amateurs, it's helpful to know what highest vote share we should expect. To that end, calculate $E[X]$.

Question 5 (6 points)

Let $Z \sim \mathcal{N}(0, 1)$. Express the random variable $Y \sim \mathcal{N}(1, 4)$ as a simple function in terms of Z. Make sure to check that your Y has the correct mean and variance.