Gov 2002: Problem Set 5

Spring 2021

Problem Set Instructions

This problem set is due on March 17, 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. The subprime loan dataset can be found at `gov2002.mattblackwell.org/data/subprime.csv`.

Question 1

(a) Let X be the random variable indicating whether Xavier votes for the incumbent party in the upcoming general election. Let Y indicate whether his neighbor Yolanda votes for the incumbent party in the upcoming general election. Show that if Xavier and Yolanda's vote choices are uncorrelated, then they are independent.

(b) If a set of random variables are *jointly* independent, it always implies that they are *pairwise* independent. However the opposite is not always true: a set of random variables may be pairwise independent, but not jointly. Verify for the following sets of random variables whether they are just pairwise independent, jointly independent or neither.

(i) $X \sim \text{Bern}(1/2)$, $Y \sim \text{Bern}(1/2)$, $Z = \max(X, Y)$

(ii) $X = 1\{\text{fair dice rolls a 3}\}$, $Y = 1\{\text{second fair dice rolls a 4}\}$, $Z = 1\{\text{sum of die is 7}\}$.

Question 2

You are a policy researcher trying to unpack what happened in the 2007-2008 U.S. foreclosure crisis. You are particularly interested in why certain people got subprime loans, which are directly linked to a higher risk of foreclosure. You have narrowed your research to the Cape Coral-Fort Myers (Florida) area, the area of the United States most devastated by the foreclosure crisis.

The dataset `subprime.csv` contains *all* home lending transactions in Cape Coral and Fort Myers. They contain information on each loan applicant and give information on whether that applicant received a subprime loan (`high.rate`) as well as on the amount of the loan

(`loan.amount`). They also contain basic demographic information such as race, gender, and income. For the remainder of the problem, we will treat this dataset as the entire, true population. Unless otherwise stated, you can assume that the samples we will draw are large enough to use the Normal approximation.

For this question, consider the variable `loan.amount` in the subprime data. You would like to calculate the standard deviation of the loan amounts because you are interested in exploring inequality in the sizes of loans that are given to various minority groups.

The usual estimator for the population standard deviation is $S_1 = [\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2]^{\frac{1}{2}}$. However one of your co-authors, based on a cursory reading of a few pages in Wikipedia, proposes an alternative estimator: $S_2 = [\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2]^{\frac{1}{2}}$. The co-author argues that S_2 is in some ways better performing than S_1 although he is vague about the details.

In order to test his contention, you decide to run a simulation study to compare the performance of the two proposed estimators on the subprime data. As in the first problem, the full subprime dataset is your complete population of interest.

(a) Write a function in `R` which implements your co-author's proposed estimator, S_2. Name your function `sd.alt`, and then check that `sd.alt(1:100)` returns the answer `28.86607`. Note that the S_1 estimator is already implemented in `R` as the `sd()` function.

(b) Set the seed to 111, and then draw 5000 random samples of size $n = 15$ from the population. For each sample, calculate and store each result of the two estimators, S_1 and S_2. Then, for the two simulated sampling distributions, report the following three quantities in a table with two columns, one for each estimator:

(i) the average estimate of the population standard deviation
(ii) the bias of your estimates
(iii) the variance of your estimates

(c) Now repeat part (b), this time with a larger sample size, $n = 100$. What do you notice now about the bias and variance of each of the estimators compared to one another, and compared with the $n = 15$ simulations you conducted in part (b)?

(d) A third co-author prefers not to program and suggests using the following estimator: $S_3 = [\frac{1}{5}\sum_{i=1}^{5}(X_i - \bar{X}_5)^2]^{\frac{1}{2}}$ where $\bar{X}_5 = \frac{\sum_{i=1}^{5} X_i}{5}$ - the mean of the first five observations. Calculate the estimate of standard deviation using S_3. Is this a good estimator? Use the characteristics of estimators discussed in lecture (bias, variance, consistency) to describe why or why not.

(e) You want to determine which estimator is "better." After watching lecture, you recommend using the mean squared error to decide. Calculate the MSE for each estimator (S_1 and S_2 and S_3) and discuss what these values tell you.

Question 3

A popular descriptive measure for a data set is the *histogram*. There are several different choices for how to bin the data and how to scale the vertical axis. In a *vertical histogram* for data $y_1, ..., y_n$, the height of the bar for a bin $(a, b]$ is

$$H_{a,b} = \frac{1}{n(b-a)} \sum_{i=1} 1(y_i \in (a, b])$$

where $1(y_i \in (a, b])$ is the indicator variables for $y_i \in (a, b]$. This means that the area of a bar is the proportion of observations that fall into the corresponding bin, so the total area of the histogram is 1. If the data are i.i.d. draws from a continuous PDF, the histogram can then be regarded as an *estimator* for the PDF.

Suppose that $y_1, ..., y_n$ are the observed values of i.i.d. continuous random variables with CDF F and PDF f.

(a) Find the expectation of $H_{a,b}$ in terms of F

(b) Show that the answer to (a) is approximately $f(a)$ if $b - a$ is small.

Note the definition of the derivative of any arbitrary function $g(x)$ with respect to x is the function $g'(x)$:

$$g'(x) = \lim_{h \to 0} \frac{g(x+h) - f(x)}{h}$$

(c) Find the variance of $H_{a,b}$.