Gov 2002: Problem Set 3

Spring 2021

Problem Set Instructions

This problem set is due on February 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.

Question 1 (25 points)

One common assumption that nearly all Presidential election forecasts[1] make is assuming that the number of voters who turn out to vote in a particular geography (and vote choice amongst those who turn out) follows a binomial distribution .

Consider the 2016 United States Presidential primary election in Erie County, Pennsylvania where there is a voting-eligible population of 200,000. Let D be the total number of voters who turn out for the Democratic primary and R be the turnout for the Republican primary (assume no voter can vote in both).

Of the Democratic primary voters, let D_{bernie} and D_{hillary} be the number of voters who vote for Bernie Sanders and Hillary Clinton respectively, and of the Republican primary voters, let R_{trump} and R_{cruz} be the number of voters who vote Donald Trump and Ted Cruz respectively.

Let's assume that $D, R, D_{\text{bernie}}, D_{\text{hillary}}, R_{\text{trump}}, R_{\text{cruz}}$ all follow the binomial distributions (with different success probabilities).

Label each of the following statements as true or false and provide some reasoning for your answer.

a. The number of Trump voters is unconditionally independent from the number of Bernie voters in Erie County.
b. The number of Trump voters is conditionally independent from the number of Cruz voters given the total Republican turnout.
c. The expectation of the total turnout across both parties is equal to the sum of the expectations of the total number of Dem voters and total number of GOP voters.

[1] For a seminal paper on probabilistic election forecasting, see "Dynamic Bayesian Forecasting of Presidential Elections in the States" (Linzer).

d. The variance of the total turnout is equal to the sum of the variances of the Dem turnout and Republican turnout.
e. The Binomial distribution assumptions would be correct if members of each household decided whether and how they were going to vote together.

Question 2 (25 points)

A common approximation for a Binomial random variable $X \sim \text{Bin}(n, p)$, such as any of the turnout counts in the previous problem, is a Poisson random variable $Y \sim \text{Pois}(\lambda)$ where $\lambda = n \cdot p$.

(a) Holding n fixed, would Y better approximate X if $p = \frac{1}{10}$ or if $p = \frac{1}{2}$? Explain your reasoning.

(b) Holding p fixed, would Y better approximate X if $n = 1{,}000$ or if $n = 100{,}000$? Explain your reasoning.

Hint: To evaluate how well a random variable approximates another random variable, compare their expectations to each other and their variances to each other.

Hint: The difference and ratio are two ways to compare quantities.

(c) In `R`, plot histogram of draws from $\text{Bin}(10, 0.7)$ and a histogram of draws from $\text{Bin}(1000, 0.7)$. Now a plot histograms of draws from their Poisson approximations. You should have four histograms in total. Compare each Binomial histogram plot to its Poisson histogram plot. Which Binomial distribution has a better Poisson approximation?

Question 3 (20 points)

Consider two independent discrete random variables, X and Y, each with two values in their support (not necessarily the same two values).

Write down probability mass functions (PMFs) for X and Y such that:

(1) $100 \cdot E[Y] < E[X]$

and,

(2) Y is greater than X with probability at least 0.99 (i.e., $\Pr(Y > X) \geq 0.99$).

There's a large number of possible solutions here. You should show that these two conditions hold for your case.

Hints:

- The PMF is a function that defines the probability that a random variable takes on every value in it's support. These probabilities must sum to 1. E.g., Let Z be a third random variable with PMF: $\Pr(Z = 1) = 0.5$ and $\Pr(Z = 5) = 0.5$. The support of Z is $\{1, 5\}$.

- Independence implies that $\Pr(X = x, Y = y) = \Pr(X = x) \Pr(Y = y)$

Question 4 (30 points)

A group of 50 people are comparing their birthdays (assume their birthdays are independent, equally likely, are not February 29, etc.).

1. Write down:
   (a) The number of pairs of people we can construct from a group of 50 people
   (b) The probability that both people in a pair share a birthday
   (c) The expected number of pairs of people with the same birthday out of a group of 50 people
2. (a) Let A_1 be an indicator variable that takes on the value of 1 if at least two people were born on Jan 1 and 0 otherwise. Find $\Pr(A_1 = 1)$, i.e., the probability that at least two people were born on Jan 1. You may find it easier to first consider the probability of the complement of this event, $\Pr(A_1 = 0)$.
   (b) Now consider analogous indicator variables for all 365 days of the year. Find the expected number of days in the year on which at least two people were born.
3. Now suppose a group of n people are comparing their birthdays
   (a) Find the expected number of pairs of people with the same birthday as a function of n (your answer to part 1 is useful here).
   (b) Let X_n be the random variable for the number of pairs of people with the same birthday out of a grop of n. What is the value of n such that the expected value is greater than 1. You can do this by showing the expected value on a graph with n on the x-axis and $E[X_n]$ on the y-axis, or by solving for the value of n such that $E[X_n] > 1$ (the quadratic formula will be useful).