# Gov 2002: Problem Set 3

Spring 2023

## Problem Set Instructions

This problem set is due on **February 15, 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 (15 points)

Consider an example of stratified sampling: A certain small town, whose population consists of 100 families, has 30 families with 1 child, 50 families with 2 children, and 20 families with 3 children. The birth rank of one of these children is 1 if the child is the firstborn, 2 if the child is the secondborn, and 3 if the child is the thirdborn.

(a) A random family is chosen (with equal probabilities), and then a random child within that family is chosen (with equal probabilities). Find the PMF and mean of the child's birth rank.

(b) A random child is chosen in the town (with equal probabilities). Find the PMF and mean of the child's birth rank.

(c) Find the variance of the child's birth rank under the set up in (a) and (b) respectively, what do you notice?

## Question 2 (15 points)

Let $n$ be the number of students in a certain graduate program, and c be the number of courses that the department will offer next semester. Suppose that each student chooses their course schedule by randomly choosing 4 courses to take in the department (with all sets of 4 courses equally likely), independent of other students' choices.

(a) Assume for this part that simultaneous enrollment is allowed (i.e., a student can enroll in two or more courses that have overlapping meeting times) and there are no enrollment caps (so students can enroll in whatever courses they want). Find the expected number of pairs of students such that the two students in the pair will take exactly the same set of courses next semester. (HINT: use the fundamental bridge.)

(b) Now suppose instead that simultaneous enrollment is *not* allowed (which explains the low workshop enrollment in said graduate program). Fortunately, there are still no enrollment caps. There are 8 different time slots in which a course can be offered (the time slots are non-overlapping). Suppose that $c = 8k$ for some positive integer $k$, and that there are exactly k courses in each of the 8 time slots next semester. As before, each student chooses their courses randomly, but now their chosen schedule will not be allowed if there are time conflicts between the courses. Find the expected number of students who will be able to take their initially chosen set of courses next semester.

## Question 3 (30 points)

Let $X$ and $Y$ be Pois($\lambda$) r.v.s, and $T = X + Y$.

(a) Using LOTUS, find $\mathbb{E}[2^X]$ and $\mathbb{E}[e^Y]$ (Hint: Recall that the Taylor series for $e^x$ is $e^x = \sum_{k=0}^{\infty} x^k/k! = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \ldots$)

(b) Prove that if $X$ and $Y$ are independent, then $T \sim \text{Pois}(2\lambda)$ (Hint: derive the PMF of $T$ from the PMFs of $X$ and $Y$ and compare this to the PMF of Pois($2\lambda$).)

(c) Formally or informally, show that if $X = Y$, then $T$ cannot follow a Pois($2\lambda$) distribution.

## Question 4 (20 points)

In many problems about modeling count data, it is found that values of zero in the data are far more common than can be explained well using a Poisson model (e.g., the onset of major conflicts between great powers). The Zero-Inflated Poisson distribution is a modification of the Poisson to address this issue, making it easier to handle frequent zero values gracefully. A Zero-Inflated Poisson r.v. $X$ with parameters $p$ and $\lambda$ can be generated as follows:

First, flip a coin with probability of $p$ of Heads. Given that the coin lands Heads, $X = 0$. Given that the coin lands Tails, $X$ is distributed Pois($\lambda$). Note that if $X = 0$ occurs, there are two possible explanations: the coin could have landed Heads (in which case the zero is called a structural zero), or the coin could have landed Tails but the Poisson r.v. turned out to be zero anyway. For example, if $X$ is the number of atom bomb a random country detonated in another country within a year, then $X = 0$ for all non-nuclear powers (this is a structural zero), but a nuclear power could still have $X = 0$ occur.

(a) Find the PMF of a Zero-Inflated Poisson r.v. $X$.
(b) Explain why $X$ has the same distribution as $(1-I)Y$ , where $I \sim \text{Bern}(p)$, $Y \sim Pois(\lambda)$, and $Z \perp\!\!\!\perp Y$.
(c) Find $\mathbb{E}[X]$ (Hint: You can either derive this using the definition of expectation, or the result from part (b))
(d) Find $\mathbb{V}[X]$ (Hint: You can either derive this using the definition of variation, or the result from part (b))

## Question 5 (20 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 $D_{\text{Bernie}}$ and $D_{\text{Hillary}}$ be the number of voters who vote for Bernie Sanders and Hillary Clinton respectively, and of the Republican primary voters, let $R_{\text{Trump}}$ and $R_{\text{Cruz}}$ be the number of voters who vote Donald Trump and Ted Cruz respectively.

Let's assume that $D$, $R$, $D_{\text{Bernie}}$, $D_{\text{Hillary}}$, $R_{\text{Trump}}$, and $R_{\text{Cruz}}$ all follow the binomial distributions (with different success probabilities). Label each of the following statements as true or false and provide some reasoning, formally or informally, 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.
(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.

[1] For a seminal paper on probabilistic election forecasting, see "Dynamic Bayesian Forecasting of Presidential Elections in the States" (Linzer 2013).