Vanderbilt University
Political Science Department
Fall 2024

Stats 1

(PSCI 8356)
Professor Jim Bisbee

MIDTERM EXAMINATION: OCTOBER 08, 2024, 9:30 a.m. - 10:45 a.m.

This is your midterm exam for Stats 1. You will need either statistical tables like those appearing in the back of the textbook or access to a statistical software program (i.e., `R`) that can provide similar information. Consultation of Web resources is not permitted. You are allowed use one double-sided piece of paper of your notes.

Be sure to show all your work and to use complete sentences to provide explanations. Point totals for each question are as follows. They sum to 100 points. It is recommended that you read all parts to a question before beginning to answer it.

- Question 1: **[15 points]**
- Question 2: **[15 points]**
- Question 3: **[20 points]**
- Question 4: **[20 points]**
- Question 5: **[30 points]**

1. Show that:

   (a) $V[Y + c] = V[Y]$ **[5 points]**

   (b) $V[aY] = a^2V[Y]$ **[5 points]**

   (c) $Cov[X, Y] = E[XY] - E[X]E[Y]$ **[5 points]**

2. Imagine the Democratic Party organizes a convention to reinvent itself and invites 19 former presidential candidates: Bennet, Biden, Booker, Bullock, Buttigieg, Castro, Delaney, Gabbard, Harris, Klobuchar, Messam, O'Rourke, Ryan, Sanders, Sestak, Steyer, Warren, Williamson, and Yang.

   Seeing this, a renowned streaming company decides to create a special that features a conversation between two of these members to capture the different perspectives at the convention. To ensure the pairing is equally likely, they will randomly select two people from the list.

   (a) How many events are in the sample space? **[3 points]**

   (b) Define event A to be: select Warren and Sanders. What is the probability of this event? **[3 points]**

   (c) Define event B to be: select Warren as one of the two. What is the probability of this event? **[5 points]**

   (d) Are A and B mutually exclusive? **[2 points]**

   (e) Are A and B independent? **[2 points]**

3. Tennessee has nine representatives to the House. Define X as the random variable representing the number of republican representatives from Tennessee at any one time. Suppose we knew that $f(x)$ is the following:

x	0	1	2	3	4	5	6	7	8	9
$f(x)$	.01	.01	.02	.02	.05	.25	.3	.2	.1	.04

Table 1: Probability Distribution of x and $f(x)$

   (a) What is the support of X? **[1 point]**

   (b) What is $f(2)$? **[2 points]**

   (c) What is $F(2)$? **[3 points]**

   (d) What is the probability that the number of Republican representatives is fewer than or equal to 2 or more than or equal to 7? **[5 points]**

   (e) What is the expected number of Republican representatives? **[3 points]**

   (f) What is the variance of the number of Republican representatives? **[3 points]**

   (g) Suppose a (strange) new electoral rule passes that says that the number of representatives elected in each district will double, and the way the extra seat in each district will be filled is that every one person elected will bring along a second person from his or her party. What is the new expected number of Republican representatives? How do you know? **[3 points]**

4. Let X and Y be continuous random variables with the joint probability density function (PDF) given by:

$$f(x,y) = \begin{cases} 6(1-x)(1-y) & \text{for } 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

   (a) Find the marginal PDFs of X and Y. **[10 points]**

   (b) Determine if X and Y are independent by checking whether the joint PDF is the product of the marginal PDFs. **[5 point]**

   (c) Calculate the expected value of the joint probability distribution. **[5 points]**

5. Any attempt to assess voter turnout with survey data is made more complicated by the challenge of social desirability bias: survey respondents tend to over-report whether they voted. For example, 58.3 percent of voting-age Americans told the Census Bureau that they voted in the 2020 presidential election. But voting statistics show that only 55.5 percent did. (This gap is even greater in years without presidential elections.)

   You are conducting a post-election survey with which you'd like to get the best possible estimate of aggregate voter turnout. You have the choice between two different survey modes. The first is a traditional telephone poll using professional interviewers ("live"). The second is a telephone poll using the interactive voice response ("IVR") technique, in which questions are asked by a recorded voice and participants enter their responses via their telephone keypad. Previous research has found that live and IVR survey modes have different advantages and disadvantages:

   - The *advantage* of IVR is that–since it is more removed from the social context of a conversation between two people–it is less subject to overreporting of the vote than surveys with live interviewers.
   - The *disadvantage* of IVR is that it is subject to more error than surveys with live interviewers.

   Your goal is to estimate p, the proportion of the voting-age population that turned out to vote. We can consider the proportion of respondents reporting they voted in the live and IVR survey modes as two different estimates of p. Call these estimates $\widehat{p}_L = \frac{Y_L}{n_L}$ and $\widehat{p}_I = \frac{Y_I}{n_I}$, respectively.

   (a) The advantages and disadvantages of the live and IVR modes correspond with two properties of estimators we've discussed in class. In a few sentences, describe these two properties, and say how we would expect $\widehat{p}_L$ and $\widehat{p}_I$ to compare regarding these two properties. **[10 points]**

   Assume it is the case that:

$$E\left(\widehat{p}_L\right) = ap + b; \;\; VAR\left(\widehat{p}_L\right) = \frac{a^2}{n} p\left(1-p\right) \;\; \text{and}$$
$$E\left(\widehat{p}_I\right) = cp + d; \;\; VAR\left(\widehat{p}_I\right) = \frac{c^2}{n} p\left(1-p\right).$$

   Further, assume that $a < c, b > d$, and $\frac{b-d}{c-a} > 1$.

(b) What is $B(\widehat{p}_L)$? What is $B(\widehat{p}_I)$? Show that $B(\widehat{p}_L) > B(\widehat{p}_I)$. **[5 points]**

(c) Show (trivially) that $VAR(\widehat{p}_I) > VAR(\widehat{p}_L)$. **[5 points]**

(d) We have a situation where one of our potential estimators suffers from greater bias, while the other is subject to more error. In class, we learned of a criterion often used to measure the tradeoff between bias and error. What is this criterion called and what is its formula? **[5 points]**

(e) Now consider a case where $a = .01, b = .1; c = .02, d = .05$ and $p = \frac{1}{2}$. According to the criterion you identified in part (d), which is the better estimator? **[5 points]**