

**Ground rules:** You must work alone on this quiz. Do not collaborate with anyone, either within or outside the class, to obtain answers or even hints. Questions of clarification should be directed to the instructor; in addition, the use of the Internet should be avoided. This quiz contains 5 questions, each of equal weight (12 points each), making the quiz worth 60 points.

1. I am presently part of a committee (one of 12 members) that reviews papers written by PhD students in Statistics and Biostatistics from all over the US. Each year (right at about this time!), we review about 100 papers in a “competition-style” format and give awards for the top 20 papers. Two of the criteria that we use to evaluate the merit of a paper are **methodological advance** and the (biomedical) **importance** of the problem. There are other criteria, but we will not consider these. Each of these criteria is judged by assigning a numerical rating 1-4; 1 = Fair, 2 = Good, 3 = Very good, and 4 = Excellent. For notational purposes, I will denote

$$\begin{aligned} Y_1 &= \text{methodological advance score} \\ Y_2 &= \text{importance score.} \end{aligned}$$

From my involvement on the committee over the last two years, I have compiled the following bivariate distribution for  $Y_1$  and  $Y_2$ . Treat this as the “correct” model to answer the questions below.

	$y_2 = 1$	$y_2 = 2$	$y_2 = 3$	$y_2 = 4$
$y_1 = 1$	0.04	0.06	0.10	0.06
$y_1 = 2$	0.02	0.08	0.12	0.08
$y_1 = 3$	0.04	0.04	0.14	0.10
$y_1 = 4$	0.01	0.03	0.05	0.03

- Find both marginal distributions.
- What is the probability that I rate a paper as “Excellent” in terms of methodological advance? in terms of importance? in terms of both criteria?
- Find  $p_{Y_1|Y_2}(y_1|y_2 = 4)$ , the conditional distribution for the methodological advance scores among those papers that received an excellent rating for importance.
- Compute the covariance of  $Y_1$  and  $Y_2$ . Interpret its value.

2. Credit card companies commonly target college students as a cohort for new credit cards. One company would like to study (probabilistically) the joint behaviour of the following two random variables:

$$\begin{aligned} Y_1 &= \text{proportion of students who are approved for Card 1} \\ Y_2 &= \text{proportion of students who are approved for Card 2.} \end{aligned}$$

The joint probability density function (pdf) for  $Y_1$  and  $Y_2$  is given by

$$f_{Y_1, Y_2}(y_1, y_2) = \begin{cases} k(1 - y_1)(1 - y_2), & 0 < y_1 < 1, 0 < y_2 < 1 \\ 0, & \text{otherwise.} \end{cases}$$

- (a) Sketch a graph of the support set in two dimensional space. Put  $y_1$  on the horizontal axis and  $y_2$  on the vertical axis. Describe in words what  $f_{Y_1, Y_2}(y_1, y_2)$  represents.
- (b) Show that  $k = 4$ .
- (c) Compute  $P(Y_1 > Y_2)$ ; that is, the probability that the proportion of students approved for Card 1 exceeds that of Card 2. Sketch a picture of the set that you are integrating over.
- (d) Compute the marginal densities of  $Y_1$  and  $Y_2$ .
- (e) Are  $Y_1$  and  $Y_2$  independent? Explain why or why not with mathematical justification.

3. Exit polling is still used as a means to gather data on voter preference despite widespread criticisms. There are many reasons that erroneous conclusions may be drawn from the data, but nearly all of them are linked to pollster incompetence and bad sampling design. What is worse is that when the “true results” are known, those results that contradict the early polling data cause some pundits to conclude voter fraud or other illegal activities have occurred. In my opinion, the reason for discrepancies is obvious: bad data give wrong conclusions. In the light of this, consider taking a more mathematical approach in the 2008 election and suppose that we decide to model

$$\begin{aligned} Y_1 &= \text{proportion of voters who vote Democrat} \\ Y_2 &= \text{proportion of voters who vote Republican} \end{aligned}$$

for Vermont, say, (who leans Democrat) with the following joint pdf

$$f_{Y_1, Y_2}(y_1, y_2) = \begin{cases} 120y_1^3y_2, & 0 < y_1 < 1, 0 < y_2 < 1, y_1 + y_2 \leq 1 \\ 0, & \text{otherwise.} \end{cases}$$

- (a) Sketch a graph of the support set in two dimensional space. Put  $y_1$  on the horizontal axis and  $y_2$  on the vertical axis. Describe in words what  $f_{Y_1, Y_2}(y_1, y_2)$  represents.
- (b) Show that both marginal distributions are beta. What are the parameters?
- (c) Find both conditional distributions. Make sure to note the support each.
- (d) Compute  $P(Y_1 > 0.5 | Y_2 = 0.25)$  and  $P(Y_1 > 0.5)$ . If this is the correct model, use these values to predict how well a third-party candidate will perform in Vermont.
- (e) Are  $Y_1$  and  $Y_2$  independent? Explain why or why not.

4. Phase I clinical trials are experiments designed to find the maximum tolerable dose of a drug in the human population. For a new prospective drug treatment, these trials are an important first step in human testing because drug companies need to determine the efficacy of the drug (at some optimal dose), while maintaining a low risk for side effects. Sometimes side effects (e.g., nausea, cramping, internal bleeding, temporary blindness) can be so severe that a patient is forced to withdraw (or dropout) from the study for safety concerns. Suppose that for a group of patients in a Phase I trial, we record

$$\begin{aligned} Y_1 &= \text{time until onset of side effect} \\ Y_2 &= \text{time until dropout} \end{aligned}$$

and model  $Y_1$  and  $Y_2$  using the following joint pdf:

$$f_{Y_1, Y_2}(y_1, y_2) = \begin{cases} e^{-y_2}, & 0 < y_1 < y_2 < \infty, \\ 0, & \text{otherwise.} \end{cases}$$

Note that by its construction,  $Y_2$  always is larger than (or equal to)  $Y_1$ , so we are not considering those patients who drop out for other reasons (i.e., other than the onset of side effects). In addition, we assume that once a side effect is experienced, it remains until dropout. Both times are measured in months.

- (a) Sketch a graph of the support set in two dimensional space. Put  $y_1$  on the horizontal axis and  $y_2$  on the vertical axis. Describe in words what  $f_{Y_1, Y_2}(y_1, y_2)$  represents.
- (b) Compute  $E(Y_1)$ ,  $E(Y_2)$ , and  $E(Y_2 - Y_1)$ . Interpret, in words, each of these expectations.
- (c) Compute  $\text{Cov}(Y_1, Y_2)$ . Are  $Y_1$  and  $Y_2$  independent?
- (d) Assuming this is the correct model, what is the proportion of subjects that do not dropout before one month?
- (e) Among those subjects that are experiencing side effects at the one month mark (but are still in the study), what is the proportion that will dropout before the second month?

5. Suppose that  $Y_1$ ,  $Y_2$ , and  $Y_3$  are random variables with

$$\begin{array}{lll} E(Y_1) = 5 & E(Y_2) = -3 & E(Y_3) = 1 \\ V(Y_1) = 4 & V(Y_2) = 9 & V(Y_3) = 16 \\ \text{Cov}(Y_1, Y_2) = 0 & \text{Cov}(Y_1, Y_3) = 1 & \text{Cov}(Y_2, Y_3) = -1. \end{array}$$

Define the statistics  $U_1 = Y_1 - Y_2$ ,  $U_2 = Y_2 + Y_3$ , and  $U_3 = 3Y_1 - 2Y_2 + 5Y_3$ .

- (a) Find the mean and variance of  $U_1$ ,  $U_2$ , and  $U_3$ .
- (b) Find  $\text{Cov}(U_1, U_2)$  and  $\text{Cov}(U_2, U_3)$ .
- (c) Find the correlation between  $U_1$  and  $U_2$ .