Midterm Examination

Econ 103, Statistics for Economists

June 5th, 2017

You will have 90 minutes to complete this exam. Graphing calculators, notes, and textbooks are not permitted.

I pledge that, in taking and preparing for this exam, I have abided by the University of Pennsylvania's Code of Academic Integrity. I am aware that any violations of the code will result in a failing grade for this course.

Name: _____

Signature: _____

Question:	1	2	3	4	5	6	7	Total
Points:	15	24	20	12	28	14	17	130
Score:								

Instructions: Answer all questions in the space provided, continuing on the back of the page if you run out of space. Show your work for full credit but be aware that writing down irrelevant information will not gain you points. Be sure to sign the academic integrity statement above and to write your name on *each page* in the space provided. Make sure that you have all pages of the exam before starting.

Warning: If you continue writing after we call time, even if this is only to fill in your name, twenty points will be deducted from your final score.

1. This question refers to the following dataset, containing 13 observations:

$$-3 \quad -3 \quad -2 \quad -1 \quad -1 \quad -1 \quad -1 \quad 0 \quad 0 \quad 1 \quad 4 \quad 7 \quad 13$$

[3] (a) Calculate the median of this dataset.

Solution: Since this dataset contains an odd number of observations, the median is simply the middle observation when the data are listed in rank order (as they are here). Thus, the median is -1.

[3] (b) Calculate the mean of this dataset.

Solution: The sum of the observations is $\sum_{i=1}^{n} x_i = 13$ so the sample mean is $\frac{1}{n}\sum_{i=1}^{n} x_i = 13/13 = 1$.

[4] (c) Suppose that it turned out there was a mistake recording the dataset: the observation listed as 13 should actually be 130. How would the mean and median change?

Solution: The median is unchanged since -1 is still the middle observation. The sum of the observations, however, has changed from 13 to 130, so the new mean is $130/13 = 10$.

[5] (d) Let f be a strictly increasing function, that is $x_1 < x_2 \Rightarrow f(x_1) < f(x_2)$. Suppose I apply f to the *original dataset* so that instead of $-3, -3, \ldots, 7, 13$ the data become $f(-3), f(-3), \ldots, f(7), f(13)$. What is the median of the transformed data? Explain your answer.

Solution: The key point here is that the function f *preserves rank orderings*. Thus,

$$f(-3) < f(-2) < f(-1) < f(0) < f(1) < f(4) < f(7) < f(13)$$

After the transformation, the order of the observations stays the same: only the *values* change. It follows that, since the median of the un-transformed data was -1, the median of the transformed data is $f(-1)$.

2. Mark each statement as True or False, and provide a brief explanation. If you do not justify your answer, you will only get partial points.

[4] (a) In large populations that are approximately bell-shaped, roughly 95% of observations will lie within one standard deviation of the mean.

Solution: FALSE: roughly 68% of observations will lie within one standard deviation of the mean. Another way to correct this is to say that roughly 95% of observations will lie within *two* standard deviations of the mean.

[4] (b) If the correlation between x and y is positive, then it must be smaller than the covariance between x and y.

Solution: FALSE: $r_{xy} = s_{xy}/(s_x s_y)$ so if, for example, s_x and s_y are both less than one, the correlation will be *larger* than the covariance.

[4] (c) If a variable has a positive skewness, then we would generally expect its mean to be *greater than* its median.

Solution: TRUE, as the mean is influenced by outliers in the (positive) tail.

[4] (d) For any events A and B, $P(B) = P(A \cap B)/P(B|A)$.

Solution: FALSE: this gives $P(A)$ rather than $P(B)$. To correct the equality, replace $P(B|A)$ by $P(A|B)$.

[4] (e) Out of a sample of 100 active astronauts, the NASA needs to choose 5 for the Mars mission. The total possible number of different teams that we can have is P_5^{100}.

Solution: FALSE: The order doesn't matter. So the correct answer is C_5^{100}.

[4] (f) If X is a random variable, the CDF $F(x_0)$ of X gives the probability that X exceeds a specified threshold x_0.

Solution: FALSE: it gives the probability that X *does not* exceed x_0, namely $P(X \leq x_0)$.

3. Let X be a random variable with support set $\{-1, 1\}$, $p(1) = q$, and $p(-1) = 1 - q$.

[5] (a) Write down the CDF of X.

Solution:

$$F(x_0) = \begin{cases} 0, & x_0 < -1 \\ 1 - q, & -1 \leq x_0 < 1 \\ 1, & x_0 \geq 1 \end{cases}$$

[3] (b) Calculate $E[X]$.

Solution: $E[X] = -1 \times (1 - q) + 1 \times q = 2q - 1$

[3] (c) Calculate $E[X^2]$.

Solution: $E[X^2] = (1 - q) \times (-1)^2 + q \times 1^2 = (1 - q) + q = 1$

[4] (d) Calculate $Var(X)$.

Solution: By the shortcut rule:

$$\begin{aligned} Var(X) &= E[X^2] - E[X]^2 = 1 - (2q - 1)^2 \\ &= 1 - (4q^2 - 4q + 1) \\ &= 4q(1 - q) \end{aligned}$$

Another way to solve this is to notice that we can write X as a linear transformation of a Bernoulli RV. In particular if $Z \sim$ Bernoulli(q) then $X = 2Z - 1$ and hence $Var(X) = 4Var(Z) = 4q(1 - q)$ as we calculated directly.

[5] (e) Let X_1 and X_2 be independent RVs both of which have the same pmf as X, defined in the problem statement. Write out the support set and pmf of $Y = 2(X_1 + X_2)$.

Solution: The support set is $\{-4, 0, 4\}$ and the pmf is

$$\begin{aligned} p(-4) &= P(X_1 = -1 \cap X_2 = -1) = (1 - q)^2 \\ p(0) &= P(X_1 = -1 \cap X_2 = 1) + P(X_1 = 1 \cap X_2 = -1) = 2q(1 - q) \\ p(4) &= P(X_1 = 1 \cap X_2 = 1) = q^2 \end{aligned}$$

4. Three percent of *Tropicana* brand oranges are already rotten when they arrive at the supermarket. In contrast, six percent of *Sunkist* brand oranges arrive rotten. A local supermarket buys forty percent of its oranges from *Tropicana* and the rest from *Sunkist*. Let R be the event that an orange is rotten and T be the event than it is a *Tropicana*.

[4] (a) What is the probability that a randomly chosen orange in the supermarket is a *Tropicana* and is rotten?

Solution: By the multiplication rule:

$$P(R \cap T) = P(R|T)P(T) = 0.03 \times 0.4 = 0.012$$

[4] (b) What is the probability that a randomly chosen orange is rotten?

Solution: By the law of total probability:

$$\begin{aligned} P(R) &= P(R|T)P(T) + P(R|T^c)P(T^c) \\ &= 0.03 \times 0.4 + 0.06 \times 0.6 \\ &= 0.012 + 0.036 \\ &= 0.048 \end{aligned}$$

[4] (c) Suppose we randomly choose an orange from the supermarket and see that it is rotten. What is the probability that it is a *Tropicana*?

Solution: By Bayes' Rule:

$$\begin{aligned} P(T|R) &= \frac{P(R|T)P(T)}{P(R)} \\ &= \frac{0.012}{0.048} = 1/4 = 0.25 \end{aligned}$$

5. Let Y and Z be discrete RVs with the following joint pmf:

		Z		
		0	1	2
Y	0	1/4	3/8	1/8
	1	0	1/8	1/8

[3] (a) Write down the support set and marginal pmf of Y.

Solution: $p_Y(0) = 3/4$, $p_Y(1) = 1/4$ with support set $\{0, 1\}$

[5] (b) Calculate $E(YZ)$.

Solution: $E[YZ] = 1 \times 1 \times 1/8 + 1 \times 2 \times 1/8 = 3/8$.

[5] (c) Is Y and Z independent? (State yes/no and justify your answer.)

Solution: No. $p_{YZ}(0,0) = 1/4$. However, $p_Y(0) \times p_Z(0) = 3/4 \times 1/4 = 3/16$.

[5] (d) Write down the conditional pmf of Y given that $Z = 1$.

Solution: $p_{Y|Z}(0|1) = (3/8)/(4/8) = 3/4$ and $p_{Y|Z}(1|1) = (1/8)/(4/8) = 1/4$.

[5] (e) Calculate $E[4Y + 5|Z = 1]$.

Solution: $E[4Y + 5|Z = 1] = 4E[Y|Z = 1] + 5 = 4(0 \times 3/4 + 1 \times 1/4) + 5 = 6$.

[5] (f) Calculate $E[E[Y|Z]]$. (*Hint: the law of iterated expectations*)

Solution: By the law of iterated expectation, $E[E[Y|Z]] = E[Y] = 0 \times 3/4 + 1 \times 1/4 = 1/4$.

6. Let X and Y be two random variables with covariance σ_{XY} and variances σ_X^2 and σ_Y^2 where $E[X] = E[Y] = 0$. Define $\beta = \sigma_{XY}/\sigma_X^2$ and $Z = Y - \beta X$.

[3] (a) Is β a random variable or a constant? Explain briefly.

Solution: A constant: it's a function of parameters, which are constants.

[5] (b) Calculate $E[Z]$.

Solution: $E[Z] = E[Y - \beta X] = E[Y] - \beta E[X] = 0$

[6] (c) Calculate $Cov(X, Z)$.

Solution: Since both X and Z have zero mean,

$$\begin{aligned} Cov(X,Z) &= E[XZ] = E\left[X\left(Y - \beta X\right)\right] \\ &= E[XY] - \beta E[X^2] = Cov(X,Y) - \beta Var(X) \\ &= \sigma_{XY} - \frac{\sigma_{XY}}{\sigma_X^2} \cdot \sigma_X^2 = 0 \end{aligned}$$

using the fact that $E[X] = E[Y] = 0$ and the definition of β.

7. This question concerns an R dataframe called `tips` containing data collected by a waiter. This data includes the amount of money he received as tips and the characteristics of the tables he served at the restaurant. Here are the first few rows of the dataframe:

```
  total_bill  tip    sex smoker day   time size
1      16.99 1.01 Female     No Sun Dinner    2
2      10.34 1.66   Male     No Sun Dinner    3
3      21.01 3.50   Male     No Sun Dinner    3
4      23.68 3.31   Male     No Sun Dinner    2
5      24.59 3.61 Female     No Sun Dinner    4
6      25.29 4.71   Male     No Sun Dinner    4
```

Each row corresponds to a particular table that this waiter served and there are no missing values. The first two columns are measured in US dollars: `total_bill` gives the total bill while `tip` gives the amount of the tip. To be clear: `total_bill` *does not include the tip*. The next four columns are categorical: `sex` is either `Female` or `Male` indicating the sex of the person in the party who paid the bill, `smoker` is either `Yes` or `No` indicating whether there were any smokers in the party, `day` indicates the day of the week when this party came to the restaurant (`Thurs`, `Fri`, `Sat`, or `Sun`), and `time` indicates whether the meal served was `Lunch` or `Dinner`. The final column, `size`, is a count of the number of diners in the party.

3 (a) What R command did I use to display the first few rows of the `tips` dataframe above?

Solution: `head(tips)`

3 (b) Write a line of R code to make a scatterplot with `total_bill` on the x-axis and `tip` on the y-axis.

Solution: `plot(tip ~ total_bill, tips)`

3 (c) Write a line of R code that will create a column called `percent` containing the tips left by each table in `tips` as a *percentage* of the total bill. Express the values as percentage points rather than decimals. For example, if a table left a tip of \$10 on a \$50 bill, the corresponding element of `percent` should be 20.

Solution: `percent <- 100 * tips$tip / tips$total_bill`

3 (d) Write a line of R code that will create a new dataframe called `smokers` containing only those rows of `tips` corresponding to tables with smokers.

Solution: `smokers <- subset(tips, smoker == "Yes")`

5 (e) Write R code to calculate the mean of `percent` broken down by `sex` and `smoker`. You can do this in one command or several: either is fine.

Solution: `as.table(by(percent, tips[,c("sex", "smoker")], mean))`