Midterm Examination I

Econ 103, Statistics for Economists

February 11, 2013

You will have 70 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: ____________________

Student ID #: ____________________

Signature: ____________________

Question:	1	2	3	4	5	6	7	8	9	Total
Points:	10	10	15	20	20	10	15	20	20	140
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 and student ID number 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, fifteen points will be deducted from your final score. In addition, one point will be deducted for each page on which you do not write your name and student ID.

1. (10 points) A study conducted at the University of Missouri-Columbia ("Mizzou") found that students who had a fake ID were more likely to engage in underage drinking while at college. Suppose we were to give fake IDs to a randomly selected group of Mizzou freshmen. Based only on the evidence given in this question, do you think this would substantially increase the fraction of these students who drink underage? Explain why or why not.

Solution: There are many possible correct answers: the point is to use what we've learned in this course about observational data and confounding to construct an argument. Here is one possibility:

> No – the causality probably runs in reverse: students who plan to do a lot of underage drinking obtain fake IDs *expressly for this purpose.* In contrast, students who do not plan to engage in underage drinking have no need of a fake ID and hence wouldn't bother to obtain them. If we gave fake IDs to teetotalers, they'd never use them. If we gave them to hardened alcoholics, they'd already have one. The only students whom this experiment might impact are those who are constrained in their underage drinking by an inability to obtain a fake ID. Given the many opportunities to drink underage on or near a college campus that *do not* require a fake ID, it is very unlikely that this constraint binds for many students.

2. (10 points) Let A and B be two mutually exclusive events, where both $P(A)$ and $P(B)$ are strictly greater than zero. Are A and B independent? Explain your answer and provide a proof. You may use any results we derived in class without proving them but be sure to name the results you use for full credit.

Solution: For any two events, the Addition Rule gives,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

but for mutually exclusive events,

$$P(A \cup B) = P(A) + P(B)$$

by the third Axiom of Probability. Combining these, $P(A \cap B) = 0$. Two events A and B are independent if and only if $P(A \cap B) = P(A)P(B)$. But we know that $P(A) > 0$ and $P(B) > 0$ so, $P(A)P(B) > 0$. Since $P(A \cap B) = 0$ it follows that A and B cannot be independent. The intuition is as follows. If A and B are mutually exclusive, then the fact that one of them has occurred *completely rules out* the possibility that the other has occurred. In contrast, if two events are independent, knowing that one has occurred gives us *no* information about whether the other will. Thus A and B can't be independent (except in the trivial case where one or both of them has zero probability, which was ruled out in the question statement).

3. Suppose that we want to predict y from x using the linear regression model $\widehat{y} = a + bx$.

(a) (5 points) Write down (but do not solve) the optimization problem needed to calculate a and b from a dataset $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Solution:

$$\min_{a,b} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

(b) (10 points) Using your answer to part (a), prove that the regression line goes through the means of the data, that is $\bar{y} = a + b\bar{x}$.

Solution: Differentiating the above expression with respect to a gives the first order condition:

$$-2\sum_{i=1}^{n}(y_i - a - bx_i) = 0$$

Rearranging and dividing both sides by n,

$$\begin{aligned} \sum_{i=1}^{n}(y_i - a - bx_i) &= 0 \\ \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a - b\sum_{i=1}^{n} &= 0 \\ \sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i &= 0 \\ \bar{y} - a - b\bar{x} &= 0 \end{aligned}$$

Therefore, $\bar{y} = a + b\bar{x}$.

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

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

(a) (5 points) 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.

(b) (5 points) 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 $\frac{1}{n}\sum_{i=1}^{n} x_i = 13/13 = 1$.

(c) (5 points) 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 $130/13 = 10$.

(d) (5 points) 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 $f(-1)$.

5. Consider a dataset with n observations on a variable x: $x_1, \ldots, x_n$. Define a new variable, y, as follows: for each x_i set $y_i = c + dx_i$ where c and d are constants and $d \neq 0$.

(a) (5 points) How is s_y^2 related to s_x^2? Prove your answer.

Solution:

$$\begin{aligned} s_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(c+dx_i-(c+d\bar{x})\right)^2 \\ &= \frac{d^2}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = d^2 s_x^2 \end{aligned}$$

(b) (5 points) How is s_{xy} related to s_x^2? Prove your answer.

Solution:

$$\begin{aligned} s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})\left(c+dx_i-(c+d\bar{x})\right) \\ &= \frac{d}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 = ds_x^2 \end{aligned}$$

(c) (5 points) What is the sample correlation between x and y? Prove your answer.

Solution: The correlation is exactly one because y is simply a linear transformation of x. From above, we know that the covariance between x and y is ds_x^2 and the variance of y is $d^2 s_x^2$. Therefore,

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{ds_x^2}{s_x\sqrt{d^2 s_x^2}} = \frac{ds_x^2}{|d|s_x^2} = \pm 1$$

(d) (5 points) Suppose we were to carry out linear regression to predict y from x, namely $\hat{y} = a + bx$. What values would we find for a and b? Prove your answer. (You may use the regression formulas from class without proving them.)

Solution:

$$\begin{aligned} b &= \frac{s_{xy}}{s_x^2} = \frac{ds_x^2}{s_x^2} = d \\ a &= \bar{y} - b\bar{x} = (c+d\bar{x}) - d\bar{x} = c \end{aligned}$$

6. (10 points) Let A and B be two events where $P(B) > 0$. Prove that

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

In your answer you may use any of the rules we learned in class except for the one you are being asked to prove. For full credit, provide the name of each rule that you use.

Solution: First, note that $A \cap B$ and $A \cap B^c$ are mutually exclusive and

$$A = (A \cap B) \cup (A \cap B^c)$$

Therefore, by the third axiom of probability,

$$P(A) = P(A \cap B) + P(A \cap B^c)$$

Finally, by the multiplication rule,

$$P(A \cap B) = P(A|B)P(B)$$

and

$$P(A \cap B^c) = P(A|B^c)P(B^c)$$

Combining these,

$$P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$$

7. The Triangle is a neighborhood that once housed a chemical plant but has become a residential area. Two percent of the children in the city live in the Triangle, and fourteen percent of these children test positive for excessive presence of toxic metals in the tissue. For children in the city who do not live in the Triangle, the rate of positive tests is only one percent. Let T be the event that a child lives in the Triangle and M be the event that a child tests positive for excessive presence of toxic metals.

   (a) (5 points) If we randomly select a child who lives in the city, what is the probability that she both lives in the Triangle and tests positive?

   Solution: By the multiplication rule:

   $$P(T \cap M) = P(M|T)P(T) = 0.14 \times 0.02 = 0.0028$$

   (b) (5 points) If we randomly select a child who lives in the city, what is the probability that she tests positive?

Name: ______________________ Student ID #: ______________________

Solution: By the law of total probability

$$\begin{aligned} P(M) &= P(M|T)P(T) + P(M|T^c)P(T^c) \\ &= 0.14 \times 0.02 + 0.01 \times 0.98 \\ &= 0.0028 + 0.0098 \\ &= 0.0126 \end{aligned}$$

(c) (5 points) If we randomly select a child who lives in the city and she tests positive, what is the probability that she lives in the Triangle?

Solution: By Bayes' Rule:

$$\begin{aligned} P(T|M) &= \frac{P(M|T)P(T)}{P(M)} \\ &= \frac{0.0028}{0.0126} = 2/9 \approx 0.22 \end{aligned}$$

8. Let X be a random variable that takes on the values 1, 2, and 3 with equal probability.

(a) (4 points) Define the term *random variable.*

Solution: A random variable is a deterministic function that assigns a real number to every basic outcome in the sample space of a random experiment.

(b) (4 points) What is the support set of X?

Solution: $\{1, 2, 3\}$

(c) (4 points) What is the pmf of X? Write it out as a piecewise function.

Solution:

$$p(x) = \begin{cases} 1/3, & x = 1 \\ 1/3, & x = 2 \\ 1/3, & x = 3 \\ 0, & \text{otherwise} \end{cases}$$

(d) (4 points) What is the CDF of X? Write it out as a piecewise function.

Solution:

$$F(x_0) = \begin{cases} 0, & x_0 < 1 \\ 1/3, & 1 \leq x_0 < 2 \\ 2/3, & 2 \leq x_0 < 3 \\ 1, & x_0 \geq 3 \end{cases}$$

(e) (4 points) Calculate the expected value of X.

Solution:

$$E[X] = 1 \cdot 1/3 + 2 \cdot 1/3 + 3 \cdot 1/3 = 1/3 + 2/3 + 1 = 2$$

9. This question refers to commands from the R statistical package that you have studied in recitation. Suppose I have a dataframe called `survey` with two columns. The first column is `height` and the second is `handspan`. Both contain numeric data. Here are the first few lines of the dataset:

```
  height handspan
1     67     20.0
2     63     19.5
3     62     19.0
4     65     19.5
5     62     18.5
6     68     18.5
```

Solution: Note that there are many possible correct answers to this question since there's more than one way to do things in R. Any set of commands that accomplishes what was asked will be counted as correct.

(a) (4 points) What command would I use to display the column `handspan` only?

Solution: `survey$handspan` would work, as would `survey[,2]`

(b) (4 points) Suppose I wanted to display only those rows from `survey` corresponding to students whose height is greater than 60 inches. What command would I use?

Solution: `survey[height > 60]`

(c) (4 points) Suppose I wanted to display only the 2nd and 9th rows of `survey`. What command would I use?

Solution: `survey[c(2,9),]`

(d) (4 points) When looking at my data, I see that one of the values for handspan is `NA`. What does this mean?

Solution: `NA` is R's way of indicating an missing observation. There is a missing observation for `handspan`.

(e) (4 points) Suppose that, for some strange reason, I wanted to rename the column `height` to `altitude` but leave the name of `handspan` unchanged. What command would I use?

Solution: This isn't covered in the current tutorials, but the proper way to do so for a `data.table` is: `setnames(survey, 'height', 'altitude')` The traditional way to do it with `data.frames` is: `names(survey) <-c(‘altitude', ‘handspan')` is one possibility, and `names(survey)[1] <- ‘altitude'` is another.