First Midterm Examination

Econ 103, Statistics for Economists

September 28th, 2015

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: _____

Signature: _____

Student ID #: _____ Recitation #: _____

Question:	1	2	3	4	5	6	7	Total
Points:	30	25	10	10	15	20	30	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, twenty-five points will be deducted from your final score. In addition, two points will be deducted for each page on which you do not write your name and student ID.

1. Indicate whether each of the following statements is True or False. If False, provide a one sentence explanation. If True, no explanation is needed.

[3] (a) The sample covariance between x and *itself* equals the sample variance of x.

Solution: True.

[3] (b) The sample variance of the z-scores always equals zero.

Solution: False: it equals one. It is the sample *mean* of the z-scores that equals zero.

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

Solution: False: we would expect the mean to be greater than the median, as the mean is influenced by outliers in the (positive) tail.

[3] (d) In a large population that is approximately bell-shaped, roughly 68% of the z-scores will be between -2 and 2.

Solution: False: roughly 95% will be between -2 and 2, while roughly 68% will be between -1 and 1.

3 (e) For any events A and B, $P(A \cap B) = P(B \cap A)$.

Solution: True

3 (f) 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 $P(A|B)$.

3 (g) A random variable is neither random, nor a variable.

Solution: True

3 (h) The expected value of a Bernoulli(p) random variable equals 1/2.

Solution: False: it equals p.

3 (i) If $X \sim$ Bernoulli(p) then $E[X^2] = p$.

Solution: True

3 (j) If X is a discrete RV with pmf $p(x)$, then $\sum_{\text{all } x} xp(x) = 1$.

Solution: False: this expression gives the *expected value* of X which need not be one. The correct statement is $\sum_{\text{all } x} p(x) = 1$.

2. Five friends want to lose weight so they sign up for a five-week program of fitness classes. At the end of the program, they note down how many weeks of classes they each attended, x, and the change in their weight in kilograms, y. The data appear in the table below. Since y denotes a *change* in weight, $y_i = -1$ means that person i lost one kilogram.

Note: *If you get an early part to this question incorrect and have to use the results of this part later on, we will take account of this in grading and assign partial credit accordingly. We have left space in the table for your calculations.*

x	y
3	0
3	-1
4	0
5	-2
5	-2

4 (a) Calculate $\bar{x}$ and $\bar{y}$.

Solution: $\bar{x} = (3+3+4+5+5)/5 = 4$, $\bar{y} = (0-1+0-2-2)/5 = -1$

6 (b) Calculate the sample variances, s_x^2 and s_y^2.

Solution: For parts (b) and (c) it is useful to fill in the table as follows:

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
-1	1	-1
-1	0	0
0	1	0
1	-1	-1
1	-1	-1

We have $s_x^2 = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2$ and $s_y^2 = \frac{1}{n-1}\sum_i^n (y_i - \bar{y})^2$

$$s_x^2 = (1/4)(1+1+0+1+1) = 1$$
$$s_y^2 = (1/4)(1+0+1+1+1) = 1$$

4 (c) Calculate the sample covariance between x and y.

Solution: Using the values from the table in solution to (b),

$$s_{xy} = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})(y_i - \bar{y})$$
$$= (1/4)(-1+0+0-1-1) = -3/4$$

6 (d) Calculate the slope and intercept of a linear regression that uses the dataset for the five friends, given above, to predict weight change, y, from number of weeks of classes attended, x.

Solution: The regression slope is $b = s_{xy}/s_x^2 = -3/4$ and the intercept is

$$a = \bar{y} - b\bar{x} = -1 - (-3/4) \times 4 = 2$$

5 (e) Each of the five friends in our dataset above attended *more* than two weeks of fitness classes. Suppose we wanted to predict the change in weight for someone *not in our dataset* who attended only two weeks of classes. Based on the regression results from the preceding part, how much would you predict that this person's weight would change?

Solution: We simply plug in the value of x into our regression to get the prediction:

$$\hat{y} = a + bx = 2 - (3/4) \times 2 = 2 - 3/2 = 0.5\text{kg}$$

10 3. Use what you know about summation notation to prove the following equality:

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}$$

Solution:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1^n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i - \bar{y}\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} \bar{x}\bar{y} \\
&= \sum_{i=1}^{n} x_i y_i - \bar{x}(n\bar{y}) - \bar{y}(n\bar{x}) + n\bar{x}\bar{y} \\
&= \left(\sum_{i=1}^{n} x_i y_i\right) - n\bar{x}\bar{y}
\end{aligned}$$

4. Suppose I flip a fair coin and roll a single fair die at the same time. Define the events
   A = the coin comes up tails
   B = the die shows a 3 *or* 5
   C = the die shows an *odd* number

3 (a) Calculate $P(B|C)$.

Name: _____ Student ID #: _____

Solution:
$$P(B|C) = P(B \cap C)/P(C) = (1/3)/(1/2) = 2/3$$

3 (b) Calculate $P(A \cap B)$.

Solution: Since the dice roll and coin flip are independent, we have $P(A \cap B) = P(A)P(B) = (1/2) \times (1/3) = 1/6$. You could also draw out the table with all 12 basic outcomes, all of which are equally likely, and count how many are in both A and B. This will give you $2/12 = 1/6$

4 (c) Calculate $P(A \cup B)$.

Solution: $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 1/2 + 1/3 - 1/6 = 3/6 + 2/6 - 1/6 = 4/6 = 2/3$. Again, you could also draw out the table with all 12 basic outcomes, all of which are equally likely, and count how many are in either A, B or both A and B. This will give you $8/12 = 2/3$

5. This question asks you to prove some results about probability that we did not cover in class. Let A, B and C be three arbitrary events.

5 (a) Prove *Boole's Inequality*: $P(A \cup B) \leq P(A) + P(B)$.

Solution: By the Addition Rule $P(A \cup B) = P(A) + P(B) - P(A \cap B)$. The result follows since $P(A \cap B) \geq 0$ by the first axiom of probability.

5 (b) Does Boole's Inequality also apply when there are three events, i.e. is it true that $P(A \cup B \cup C) \leq P(A) + P(B) + P(C)$? Explain why or why not.

Solution: Yes. Let $D = A \cup B$. Then by the previous part

$$P(A \cup B \cup C) = P(D \cup C) \leq P(D) + P(C)$$

Applying the previous part a *second time*,

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

The result follows by combining these.

5 (c) Prove *Bonferroni's Inequality*: $P(A \cap B) \geq P(A) + P(B) - 1$.

Solution: Rearranging the Addition Rule, $P(A \cap B) = P(A) + P(B) - P(A \cup B)$. The result follows since $P(A \cup B)$ is at most one by the first axiom of probability.

6. New England Patriots quarterback Tom Brady has a lucky coin, which he flips five times before every game. The coin is fair, so it has a probability $1/2$ of coming up heads on any given flip.

[5] (a) What is the probability that Tom gets the same outcome for all 5 flips of his lucky coin? Express your answer as a fraction.

Solution: There are two possibilities here: Tom can either get 5 heads, or he can get 5 tails. Both have the same probability, $\left(\frac{1}{2}\right)^5$. As they are mutually exclusive, we can add the probability of each possibility to get:

$$Pr(5H \cup 5T) = P(5H) + P(5T) = \left(\frac{1}{2}\right)^5 + \left(\frac{1}{2}\right)^5 = \frac{1}{32} + \frac{1}{32} = \frac{1}{16}$$

[15] (b) Tom is a superstitious person. If (and only if) he gets 5 heads or 5 tails he will cheat by deflating the balls used in the game. If Tom cheats, the Patriots win with probability $\frac{3}{4}$. But if Tom does not cheat, they win with probability $\frac{2}{5}$. Given that the Patriots won the game, what is the probability that Tom cheated?

Solution: First write down all the probabilities given in the question statement. Let C be the event that Tom Brady cheats C and W be the event that the Patriots win. (Using this notation C^c is the event that Tom *doesn't* cheat.) From Part (c), we know that $P(C) = 1/16$ and therefore $P(C^c) = 1 - P(C) = 15/16$. We are told in the problem statment that $P(W|C) = 3/4$ while $P(W|C^c) = 2/5$ and are asked to find $P(C|W)$. By Bayes' Rule:

$$P(\mathrm{C}|\mathrm{W}) = \frac{P(\mathrm{C})P(\mathrm{W}|\mathrm{C})}{P(\mathrm{W})}$$

To calculate the denominator, we use the law of total probability:

$$\begin{aligned} P(\mathrm{W}) &= P(\mathrm{W}|\mathrm{C})P(\mathrm{C}) + P(\mathrm{W}|C^c)P(C^c) \\ &= 3/4 \times 1/16 + 2/5 \times 15/16 \\ &= 3/64 + 3/8 = 3/64 + 24/64 = 27/64 \end{aligned}$$

Now putting all of these pieces together, we get:

$$P(\mathrm{C}|\mathrm{W}) = \frac{P(\mathrm{C})P(\mathrm{W}|\mathrm{C})}{P(\mathrm{W})} = \frac{3/64}{27/64} = \frac{3}{27} = \frac{1}{9}$$

7. This question asks you for the R commands needed to carry out tasks similar to those in the tutorials. The first two parts refer to the dataframe `jobsearch` that we created in R Tutorial 1. The entire dataframe is printed below.

```
##   location salary              title hours
## 1 New York  70000     Office Manager    50
## 2  Chicago  80000 Research Assistant    56
## 3   Boston  60000            Analyst    65
## 4   Boston  50000     Office Manager    40
## 5 New York  45000            Analyst    50
```

4 (a) Create a new dataframe called 'highpay' by taking a subset of the 'jobsearch' dataframe to include only jobs which pay more than $55,000.

Solution: `highpay = jobsearch[salary > 55000]` If `jobsearch` is not a `data.table` but a `data.frame`, we must instead use: `highpay = subset(jobsearch, salary > 50000)`

4 (b) Write code to calculate the number of observations in your new dataframe, `highpay`, and call this number `n`.

Solution: `n = nrow(highpay)`

The next four parts of this question concerns the dataframe `survey` from R Tutorial 2. This dataframe has observations of six variables. The variable names, and the first six observations are printed below:

```
##      sex credits eye.color handedness height handspan
## 1   Male       5     Brown        1.0     67     20.0
## 2 Female       5     Brown        0.4     63     19.5
## 3 Female      NA     Brown        0.6     62     19.0
## 4 Female       5     Brown        0.6     65     19.5
## 5 Female       4     Brown        1.0     62     18.5
## 6 Female      NA     Brown        1.0     68     18.5
```

Name: _____ Student ID #: _____

4 (c) Write the command that prints the first six rows of `survey`.

Solution: `head(survey)`

4 (d) Write the command to draw a scatter plot with handedness on the horizontal axis and height on the vertical axis.

Solution: `plot(survey$handedness, survey$height)`

4 (e) Create a dataframe called `numerical` that contains only the numerical variables in our dataframe above.

Solution: There are four numerical variables in our data: credits, handedness, height and handspan. There are various ways to construct the required dataframe. One possibility is `numerical <- survey[,c(2, 4, 5, 6)]` Another is `numerical <- data.frame(survey$credits, survey$handedness, survey$height, survey$handspan)`

The final part of this question does not refer to any of the datasets from above: instead it asks to you demonstrate your knowledge of R functions.

10 (f) Write an R function called `myIQR` that calculates the interquartile range of a vector of data. Your function should take only one argument: the vector of data `x`. You may assume that `x` does not contain any missing values. You may use any R functions you like in your answer *except* the built-in function for calculating an interquartile range, namely `IQR`.

Solution: Many possibilities. Here is one:

```
myIQR <- function(x){
  Q75 <- quantile(x, 0.75)
  Q25 <- quantile(x, 0.25)
  out <- Q75 - Q25
  return(out)
}
```

In fact, this is very close to exactly what the built-in `IQR` function in R does:

```
IQR = function (x, na.rm = FALSE, type = 7) {
diff(quantile(as.numeric(x), c(0.25, 0.75),
              na.rm = na.rm, names = FALSE, type = type))
    }
```

Name: _____ Student ID #: _____