First Midterm Examination

Econ 103, Statistics for Economists

February 13th, 2018

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	Total
Points:	40	30	25	15	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, a point will be deducted for each page on which you do not write your name and student ID.

1. Write down the answer to each of the following. No explanation is needed.

4 (a) What is the formula for the sample variance of $x_1, \ldots, x_n$?

Solution: $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

4 (b) What is the formula for the sample covariance between $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$?

Solution: $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

4 (c) Using the rule of thumb for skewness, how would we expect the mean and median of a right-skewed dataset to compare to one another?

Solution: We would expect the mean to be *larger* than the median.

4 (d) If x is measured in meters, what are the units of s_x^2?

Solution: Square meters

4 (e) TRUE or FALSE: $s_{xy} \geq r_{xy}$.

Solution: FALSE: $s_{xy} = r_{xy}s_x s_y$ so if $s_x s_y < 1$ we have $s_{xy} < r_{xy}$.

4 (f) Let A and B be two events. Write down the expression for $P(A \cup B)$.

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

4 (g) Let A and B be two events. What is the definition of $P(B|A)$?

Solution: $P(B|A) = P(B \cap A)/P(A)$.

4 (h) Let A and B be two events. Write down Bayes' rule for calculating $P(B|A)$.

Solution: $P(B|A) = P(A|B)P(B)/P(A)$

4 (i) TRUE or FALSE: if events A and B are mutually exclusive they are independent.

Name: _____ Student ID #: _____

Solution: FALSE: if A and B are mutually exclusive, the fact that one occurred implies that the other *cannot have occurred*.

4 (j) TRUE or FALSE: $P(A \cap B) \leq P(A \cup B)$.

Solution: TRUE: $(A \cap B) \subseteq (A \cup B)$ so $P(A \cap B) \leq P(A \cup B)$ by the logical consequence rule.

The following question is based on a problem from your assigned homework. The reasoning and solutions to parts (a)–(d) are identical to the corresponding questions from your homework, although I have re-worded the question for clarity. Part (e) is new.

2. Let $x_1, \ldots, x_n$ be a sample of n observations and define $y_i = c + dx_i$ where c and d are constants and $d \neq 0$. Let $\bar{x}$ be the sample mean and s_x^2 be the sample variance of x. Analogously, let $\bar{y}$ be the sample mean and s_y^2 be the sample variance of y. Show your work in each of the following parts and note that your answers may involve c and d.

5 (a) Express $\bar{y}$ in terms of $\bar{x}$.

Solution:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n}(c + dx_i) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} c + d\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = c + d\bar{x}$$

5 (b) Express s_y^2 in terms of s_x^2.

Solution:

$$s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}[(c + dx_i) - (c + d\bar{x})]^2 = \frac{1}{n-1}\sum_{i=1}^{n}[d(x_i - \bar{x})]^2 = d^2 s_x^2$$

5 (c) Express s_y in terms of s_x.

Solution: $s_y = \sqrt{s_y^2} = |d| \times s_x$

5 (d) Let z_i^x be the sample z-score of x_i and z_i^y be the sample z-score of y_i. Briefly explain how z_i^y is related to z_i^x. Does the relationship depend on c or d? If so, how?

Solution: They are identical as long as d is positive, but the sign will flip if d is negative:

$$z_i^y = \frac{y_i - \bar{y}}{s_y} = \frac{(c + dx_i) - (c + d\bar{x})}{|d| \times s_x} = \frac{d}{|d|}\left(\frac{x_i - \bar{x}}{s_x}\right) = \text{sign}(d) \times z_i^x$$

10 (e) Calculate the sample correlation r_{xy} between x and y. Does the answer depend on c or d? If so, how?

Solution: The correlation equals 1 if d is positive and -1 if d is negative:

$$\begin{aligned} r_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\frac{d}{|d|}\left(\frac{x_i - \bar{x}}{s_x}\right) \\ &= \frac{d}{|d|}\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)^2\right] = \frac{d}{|d| \times s_x^2}\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right] \\ &= \frac{d \times s_x^2}{|d| \times s_x^2} = \frac{d}{|d|} \end{aligned}$$

3. Suppose that 1 out of N coins is defective and I chose my coin at random. Whereas a regular coin is equally likely to come up heads or tails, a defective coin *always* comes up heads. I choose a coin at random and flip it ten times: I get ten heads. Let D be the event that my coin is defective and A be the event that I get ten heads in ten flips of my coin. Be sure to explain your reasoning in each of the following parts.

4 (a) What is $P(A|D)$?

Solution: $P(A|D) = 1$ since a defective coin always comes up heads.

4 (b) What is $P(A|D^c)$?

Solution: $P(A|D^c) = 1/2^{10} = 1/1024$ since a regular coin is equally likely to come up heads or tails, and each flip is independent.

4 (c) Expressed as a function of N, what is $P(D)$?

Solution: $P(D) = 1/N$ since 1 out of N coins is defective and I chose a coin at random.

[4] (d) Expressed in terms of N, what is $P(D^c)$?

Solution: By the complement rule $P(D^c) = 1 - P(D) = 1 - 1/N = (N-1)/N$.

[4] (e) Expressed in terms of N, what is $P(A)$?

Solution: By the law of total probability:

$$\begin{aligned} P(A) &= P(A|D) \times P(D) + P(A|D^c) \times P(D^c) \\ &= 1/N \times 1 + (N-1)/N \times 1/1024 = (1024 + N - 1)/(N \times 1024) \end{aligned}$$

[5] (f) Expressed in terms of N, what is $P(D|A)$?

Solution: By Bayes' rule

$$P(D|A) = \frac{P(A|D)P(D)}{P(A)} = \frac{1/N}{(1024 + N - 1)/(N \times 1024)} = \frac{1024}{1024 + N - 1}$$

4. To answer this question, you will need the following fact: in R the function `factorial` is used instead of an exclamation point "!" to calculate a factorial. For example, `factorial(3)` would return a result of `6`.

[5] (a) Write down the formula for $\binom{n}{k}$ in terms of factorials. To be clear: this part is *not* asking you for any R code.

Solution: $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

[10] (b) Using your formula from the preceding part, write an R function called `mycombn` that calculates combinations. Your function should take two inputs, `n` and `k` and return a single output: $\binom{n}{k}$.

Solution:

```
mycombn = function(n, k) {
  numerator = factorial(n)
  denominator = factorial(k) * factorial(n - k)
  return(numerator / denominator)
}
```

5. This question is based on a dataset called `pickup.csv` containing the model year, mileage, price (in US dollars), and make of 46 pickup trucks listed for sale on Craigslist in Austin Texas. Here are the first few rows of the dataset:

```
   year  miles price  make
1: 2008  17638 14995   GMC
2: 2003 174000  8500 Dodge
3: 2001   1500  9998 Dodge
4: 2007  22422 23950   GMC
5: 2007  34815 19980   GMC
6: 1997 167000  5000   GMC
```

In any R code that you write to answer this question, you may assume that the `data.table` package has already be installed and loaded.

[3] (a) The data are stored at `http://jgscott.github.io/teaching/data/pickup.csv`. Using this url, write out the line of R code you would use to download the file `pickup.csv` and store it in a data.table called `pickup`.

Solution:

```
pickup = fread('http://jgscott.github.io/teaching/data/pickup.csv')
```

[3] (b) Here is a table that contains average price for each make of truck:

```
    make avg_price
1:   GMC  7996.208
2: Dodge  6554.200
3:  Ford  8867.917
```

Write down the line of R code you would use to generate these results.

Solution:

```
pickup[, .(avg_price = mean(price)), by = make]
```

[3] (c) Here are the results of a linear regression that uses `miles` to predict `price`:

```
Coefficients:
(Intercept)        miles
 14419.3762      -0.0643
```

Write down the line of R code you would use to generate these results.

Solution:

```
pickup[, lm(price ~ miles)]
```

Name: _____ Student ID #: _____

6 (d) In the regression results from above, what are the units of the intercept? What are the units of the slope?

Solution: The intercept is measured in dollars, and the slope is measured in dollars per mile.

5 (e) Suppose a truck with zero miles were listed on Craigslist. Based on the regression results from above, what price would we predict for this truck?

Solution: Around 14,419 dollars.

5 (f) Consider two pickup trucks: truck A has 10,000 more miles than truck B. Based on the regression results from above, which truck would we predict has the higher price? How much higher?

Solution: We would predict truck A to have a *lower* price than truck B by $0.0643 \times 10,000 = 643$ dollars.

5 (g) The sample mean of `price` is approximately `7900` dollars. Based on the regression results from above, approximately what is the sample mean of `miles`?

Solution: We know that $\bar{y} = a + b\bar{x}$, $\bar{y} = 7900$, $a = 14419$ and $b = -0.0643$. Thus, $\bar{x} = (\bar{y} - a)/b \approx 101{,}400$ miles.

Name: _____ Student ID #: _____