First Midterm Examination

Econ 103, Statistics for Economists

February 19th, 2019

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:	20	15	20	20	20	20	25	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. Let m be a constant and $x_1, \ldots, x_n$ be an observed dataset.

[10] (a) Show that $\sum_{i=1}^{n}(x_i - m)^2 = \sum_{i=1}^{n} x_i^2 - 2m\sum_{i=1}^{n} x_i + nm^2$.

Solution:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - m)^2 &= \sum_{i=1}^{n}(x_i^2 - 2mx_i + m^2)\\ &= \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2mx_i + \sum_{i=1}^{n} m^2\\ &= \sum_{i=1}^{n} x_i^2 - 2m\sum_{i=1}^{n} x_i + nm^2\end{aligned}$$

[10] (b) Using the preceding part, show that $\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.

Solution: Solving this requires two observations. First, note that $\bar{x}$ is a *constant*, i.e. that it does not have an index of summation. Second, note that $\sum_{i=1}^{n} x_i = n\bar{x}$. Hence, taking $m = \bar{x}$ in the formula from the preceding part,

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2\\ &= \sum_{i=1}^{n} x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2\\ &= \sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\\ &= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2\end{aligned}$$

[15] 2. Given observations $x_1, x_2, \ldots, x_n$, what value of a minimizes $\frac{1}{n}\sum_{i=1}^{n}(x_i^2 - a)^2$? Explain.

Solution: Differentiating with respect to a, the first order condition is

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

Name: _____ Student ID #: _____

Re-arranging, and solving for a, we obtain:

$$a = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

20 3. Suppose I flip a fair coin three times. Let A be the event that I get *at least one head*, and B be the event that I get *exactly two heads*.

(a) Calculate $P(B|A)$.

Solution: This experiment has $2 \times 2 \times 2 = 8$ equally likely basic outcomes. Of these, 7 have at least one head:

HHH THH HTH HHT HTT THT TTH

Hence, $P(A) = 7/8$. Among the 7 basic outcomes with at least one head, 3 have exactly two heads, namely:

THH HTH HHT

Thus, $P(B \cap A) = 3/8$. By the definition of conditional probability, the desired probability is $P(B|A) = P(B \cap A)/P(A) = (3/8)/(7/8) = 3/7$.

(b) Are the events A and B independent? Justify your answer.

Solution: Of the 8 equally likely basic outcomes, 3 have exactly two heads, namely HHT, HTH, and THH. Thus, $P(B) = 3/8$. From the preceding part, however, we know that $P(B|A) = 3/7$. Since knowing that A has occurred changes the probability that B will occur, A and B are *not independent*. Another way to argue this is by pointing out that $P(A) \times P(B) = \frac{7}{8} \times \frac{3}{8} = \frac{21}{64}$ whereas $P(A \cap B) = 3/8$.

20 4. Bob is a randomly chosen resident of Peoria, a city in which 3% of people use cocaine. Bob tests positive for cocaine in a drug test that correctly identifies users 95% of the time and correctly identifies non-users 90% of the time. Calculate the probability that Bob is a cocaine user. (In your calculations, let U be the event that Bob is a cocaine user and T be the event that he tests positive.)

Name: _____ Student ID #: _____

Solution: By the law of total probability:

$$P(T) = P(T|U)P(U) + P(T|U^C)P(U^C) = \frac{95}{100} \times \frac{3}{100} + \frac{10}{100} \times \frac{97}{100}$$
$$= (285 + 970)/(100 \times 100) = 1255/(100 \times 100)$$

Therefore, by Bayes' rule

$$P(U|T) = \frac{P(T|U)P(U)}{P(T)} = \frac{285}{1255} \approx 0.23$$

5. Let X be a RV with support set $\{-1, 1\}$ and $p(-1) = 1/2$.

[2] (a) Write out the pmf of X.

Solution: $p(-1) = 1/2$ and $p(1) = 1/2$.

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

Solution: $E[X] = -1 \times 1/2 + 1 \times 1/2 = 0$

[5] (c) Write out the CDF $F(x_0)$ of X.

Solution:

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

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

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

[2] (e) Calculate $Var(X)$.

Solution: By the shortcut formula, $Var(X) = E[X^2] - E[X]^2 = 1 - 0^2 = 1$.

[5] (f) Calculate $E\left[\frac{X}{X^2+1}\right]$

Solution:

$$E\left[\frac{X}{X^2+1}\right] = \frac{-1}{(-1)^2+1} \times \frac{1}{2} + \frac{1}{1^2+1} \times \frac{1}{2} = -1/2 \times 1/2 + 1/2 \times 1/2 = 0.$$

6. In each of the following parts, write down the result that would appear in the R console if you were to run the indicated lines of code.

5 (a)
```
x <- c(-1, 5, 2, -4, 8)
x[c(1,2)]
```

Solution: `-1 5`

5 (b)
```
w <- c(4, 5, 6)
z <- c(3, 2, 1)
rbind(z, w)
```

Solution:
```
3 2 1
4 5 6
```

5 (c)
```
M <- cbind(c(1, 2, 3), c(4, 5, 6))
M[2,]
```

Solution: `2 5`

5 (d)
```
person <- c("Alice", "Bob", "Cari", "Dan")
year_of_birth <- c(1985, 1992, 1985, 1997)
df <- data.frame(person, year_of_birth)
subset(df, year_of_birth == 1985)
```

Solution:
```
person    year_of_birth
Alice     1985
Cari      1985
```

7. This question is based on a dataset called `brexit.csv` that is available on my website at `http://ditraglia.com/econ103/brexit.csv`. Here are the first few rows:

```
                Area     Region Pct_Leave mean_hourly_pay2005
```

Name: _____ Student ID #: _____

1	Hartlepool	North East	69.57	10.89
2	Middlesbrough	North East	65.48	10.02
3	Redcar and Cleveland	North East	66.19	11.45
4	Stockton-on-Tees	North East	61.73	12.15
5	Darlington	North East	56.18	11.03
6	Halton	North West	57.42	10.50

The dataset contains results from the 2016 UK Brexit referendum, in which British voters were asked whether they wished to "leave" or "remain" in the European Union. Each row contains information for a single voting area (effectively a precinct). The column `Area` is a character vector containing the name of the area, while `Region` is a factor indicating the region in which this area is located. The remaining columns are numeric vectors: `Pct_Leave` gives percentage of voters in an area who voted to *leave* the European Union ($0 = 0\%$ and $100 = 100\%$), while `mean_hourly_pay2005` gives the mean hourly pay of the area in 2005 measured in pounds sterling (GBP). There are no missing values.

3 (a) Write R code to load `brexit.csv` from my website and store it as a dataframe called `brexit`.

Solution: `brexit <- read.csv('http://ditraglia.com/econ103/brexit.csv')`

(b) Write R code to display the first six rows of the dataframe `brexit`.

Solution: `head(brexit)` or alternatively `head(brexit, n = 6)`

3 (c) Write R code to make a histogram of mean hourly pay in 2005 across areas. You do not have to add a title or label the axes.

Solution:

```
hist(brexit$mean_hourly_pay2005)
```

3 (d) Write R code to carry out the following steps: (i) run a regression using mean hourly pay in 2005 to predict the percentage voting leave, (ii) store the result in an object called `reg`, (iii) display the slope and intercept of `reg`.

Solution:

```
reg <- lm(Pct_Leave ~ mean_hourly_pay2005, brexit)
coef(reg)
```

5 (e) The results of the code you wrote in the preceding part are as follows:

```
(Intercept) mean_hourly_pay2005
       76.2                -1.7
```

Suppose we consider two areas. In the first, mean hourly pay equals 20 GBP; in the second it equals 10. Based on the regression results, how would we predict that the percentage voting "leave" would differ between these areas? Your answer should *not involve any R code*.

Solution: We would predict that `Pct_Leave` would be about 17 *lower* in the area with the higher hourly pay. In other words, voters are more likely to vote leave in *poorer* areas.

5 (f) I ran the following line of R code:

```
sd(brexit$Pct_Leave) / sd(brexit$mean_hourly_pay2005)
```

and got a result of approximately 3.5. Based on this and the regression results from the preceding part, what is the approximate correlation between `Pct_Leave` and `mean_hourly_pay2005`? Your answer should *not involve any R code*.

Solution: Recall that:

$$b = \frac{s_{xy}}{s_x^2} = \frac{r s_x s_y}{s_x^2} = r\left(\frac{s_y}{s_x}\right)$$

We know that $b = -1.7$ and $s_y/s_x = 3.5$. Hence, $r = -1.7/3.5 \approx -0.49$.

3 (g) Write a line of R code that uses `reg` to predict the percentage of voters that we would expect to vote remain in four hypothetical areas with mean hourly pay equal to 5, 10, 15, and 20.

Solution:

```
predict(reg, newdata = data.frame(mean_hourly_pay2005 = c(5, 10, 15, 20)))
```

3 (h) Write a line of R code to make a side-by-side boxplots of the percentage voting leave, broken down by Region. You do not have to add a title or axis labels.

Solution:

```
boxplot(Pct_Leave ~ Region, brexit)
```