Final Examination

Econ 103, Statistics for Economists

May 1st, 2018

You have 120 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	Total
Points:	25	25	30	20	60	40	200
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. Someone in San Jose California, a city with 1 million residents, has stolen a car. The incident was captured on video, providing a detailed description of the thief. One out of every 10,000 people in San Jose fits this description. The police spot Amy walking down the street; she meets every detail of the description, so they arrest her. The only evidence against Amy is that she fits the description. The prosecutor argues as follows:

   > It is *highly* unlikely (far beyond a reasonable doubt) that an innocent person would fit this description. Hence, it is highly unlikely that Amy is innocent.

   In parts (b)–(d) below, let I denote the event that someone from San Jose is innocent of the theft, and D be the event that she fits the description from the video.

   $\boxed{5}$ (a) How many of the residents of San Jose fit the description from the video?

   $\boxed{5}$ (b) Calculate $P(I \cap D)$.

   $\boxed{5}$ (c) Calculate $P(D|I)$.

   $\boxed{5}$ (d) Calculate $P(I|D)$.

   $\boxed{5}$ (e) In light of your responses to parts (c) and (d), evaluate the prosecutor's argument. Should the jury vote to convict Amy? Explain briefly.

2. In this question you will determine the same probability two different ways: first by using the rules for calculating probabilities from class, and then by Monte Carlo simulation.

$\boxed{10}$ (a) There is an urn containing six balls, four of which are red. You make two random draws from the urn *without replacement*. What is the probability that both of the balls you draw are red?

$\boxed{15}$ (b) Write R code to check your calculation from part (a) via Monte Carlo simulation using 10,000 simulation replications. For simplicity, I suggest that you create a vector called `urn` containing ones and zeros, where the ones represent the four red balls and the zeros represent the two other balls. You can then make random draws without replacement from `urn`.

3. Let X be a continuous RV with support set $[0, 1]$ and pdf $f(x) = \alpha x^{\alpha-1}$ where $\alpha > 0$.

$\boxed{10}$ (a) Calculate the CDF of X.

$\boxed{10}$ (b) Calculate $E[X]$.

$\boxed{10}$ (c) Calculate $\text{Var}(X)$. You do not have to simplify your answer.

4. Let $X_1, \ldots, X_n \sim$ iid $\mathrm{N}(\mu, \sigma^2)$ where σ^2 is known. In class we derived a $(1-\alpha) \times 100\%$ confidence interval for μ that takes the form Estimator $\pm$ ME. This is called a *two-sided* confidence interval, because it has two finite endpoints. It is also possible to construct *one-sided* confidence intervals, although we did not consider these in Econ 103. An *upper one-sided confidence interval* for a some parameter θ is defined as a range $[\text{LCL}, +\infty)$ constructed from the sample data such that $P(\text{LCL} \leq \theta) = 1 - \alpha$. In other words, the interval $[\text{LCL}, +\infty)$ covers θ with probability $1-\alpha$. In this problem you will construct an upper one-sided confidence interval for μ based on $X_1, \ldots, X_n$.

$\boxed{5}$ (a) What is the sampling distribution of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$? Be sure to specify the values of any and all parameters of its distribution.

$\boxed{5}$ (b) Continuing from the preceding part, write down the line of R code we would use to find the value of c such that $P\big(\sqrt{n}(\bar{X}_n - \mu)/\sigma \leq c\big) = 1 - \alpha$.

$\boxed{10}$ (c) Using the expression from (b), derive the formula for LCL such that $[\text{LCL}, +\infty)$ is an upper one-sided confidence interval for μ with confidence level $(1-\alpha)$.

5. This question is taken from your homework. It is based on a dataset containing the results of the tae kwon do event in the 2004 Athens Olympics. The competition is a tournament consisting of a number of bouts. In each bout, a pair of competitors fight each other, points are awarded, and a winner is declared by the judges. In accordance with Olympic regulations, one of the competitors in each bout is *randomly chosen* to wear blue body protectors. The other wears red body protectors. This question investigates whether wearing one color or the other gives an advantage in the competition. The data are stored in a data table called `taekwondo`. Each row corresponds to a *single bout*:

`red.id`	competitor id number for the fighter who wore red
`blue.id`	competitor id number for the fighter who wore blue
`round`	round of the tournament (i.e. semifinals, finals, etc.)
`winner`	color worn by the fighter who won the bout
`method`	method of win (i.e. points, knockout, etc.)
`red.points`	number of points awarded to the fighter who wore red
`blue.points`	number of points awarded to the fighter who wore blue

There are no missing values in the dataset. Here are the first few rows:

```
   red.id blue.id   round winner  method   red.points blue.points
1:   5816    5818 last 16   Blue  Points            9           5
2:   5817    5824 last 16   Blue  Points            3           5
3:   5819    5825 last 16    Red  Points           15          16
4:   5820    5822 last 16    Red  Points           14          15
5:   5821    5827 last 16    Red  Points           13          12
6:   5828    5823 last 16    Red  Knockout          7           3
```

4 (a) We'll restrict attention to the "last 16" round of the competition to ensure that each row contains a *unique* pair of fighters. Write R code to extract only those rows of `taekwondo` for which the value in the column `round` is "last 16" and store the result in a data table called `last16`.

6 (b) To begin, we'll analyze the *proportion* of bouts won by the blue fighter. Write R code to: (i) count the number of elements in the column `winner` of `last16` and store the result in a variable called `n`, and (ii) count the number of bouts won by the blue fighter and store the result in a variable called `n.blue`.

Name: ________________ Student ID #: ________________

10 (c) There are 32 bouts in `last16` of which 19 were won by the blue fighter. Using this information, calculate an approximate 95% confidence interval for the population proportion of bouts won by fighters wearing blue based on the approximation provided by the CLT. Do your results suggest that wearing one color versus the other conveys a competitive advantage? Explain briefly.

10 (d) Now suppose that you wanted to test the null hypothesis that the population proportion of bouts won by fighters wearing blue equals 0.5 against the two-sided alternative. Approximately what is your p-value for this test? Explain your results.

6 (e) For the remainder of the question, we will examine the relative difference in the number of *points* scored by the blue and red fighters in each bout. Write R code accomplish the following: (i) select only those rows of `last16` for which the value in the column `method` is `Points` and store the result in a data table called `last16.points`, (ii) create a vector called `D` whose entries contain the *difference* in the number of points scored by blue versus red (Blue - Red) in each bout with `method` equal to `Points`.

$\boxed{4}$ (f) I calculated the mean of the column `red.points` in `last16.points` and got 10.1. Similarly, I calculated the mean of the column `blue.points` and got 11.7. If I were to run the command `mean(D)` at the R console what result would I get?

$\boxed{10}$ (g) I entered the command `var(D)` at the R console and got 25. Next I entered `var(last16.points$red.points)` and `var(last16.points$blue.points)` and got 17 and 31, respectively. Calculate the sample correlation between the columns `red.points` and `blue.points` of the data table `last16.points`.

$\boxed{10}$ (h) To test the null hypothesis that red and blue fighters are awarded, on average, the same number of points against the two-sided alternative, should we use a test for independent samples or matched pairs data? Explain briefly and then carry out the appropriate test at the 5% level based on the CLT. To answer, you will need the fact that there are 29 rows in the data table `last16.points`. Be sure to report: (i) the test statistic, (ii) the decision rule, and (iii) the result of the test.

6. This question is based on a data table called `face` containing data from a paper investigating whether "inferences of competence based solely on facial appearance predicted the outcomes of U.S. congressional outcomes." Here are the first few rows of the dataset:

```
   year state     Dcomp    Dshare blueState
1: 2002    AK 0.3335918 0.1166021         0
2: 2004    AK 0.4193548 0.4777151         0
3: 2002    AL 0.5510950 0.4048368         0
4: 2004    AL 0.3223141 0.3236140         0
5: 2002    AR 0.2738834 0.5386246         0
6: 2004    AR 0.7360000 0.5582051         0
```

Each row corresponds to a particular U.S. congressional election: `year` gives the year of the election, `state` gives the state, and `Dshare` gives the Democratic vote share. For example, the value of 0.1166021 for `Dshare` in the first row indicates that the Democratic candidate in the 2002 congressional race in Alaska got just under 12% of the vote. The column `blueState` is a dummy variable taking the value 1 if the corresponding state is a "blue state," defined as a state in which most voters have favored the Democratic presidential candidate in past elections. We see from the first six rows of this dataset that Alaska (AK), Alabama (AL), and Arkansas (AR), are *red states*. The column `Dcomp` is a measure of the "facial competence" of the Democratic congressional candidate, constructed as follows. For each congressional race, a group of students were shown *unlabeled* photographs of the Democratic and Republican candidates side-by-side:

Which person is the more competent?

After viewing the photographs for one second, the students were asked which person appeared more competent, based on the photographs. The variable `Dcomp` records the *fraction* of students who thought that the Democratic candidate appeared more competent. For example, the value of 0.3335918 for `Dcomp` in the first row of `face` indicates that just over a third of the students thought that the Democratic candidate from the 2002 Alaska congressional race appeared more competent than his Republican rival. Any student who recognized either photograph was excluded from the calculation, so that `Dcomp` is constructed purely "based on judgements derived from facial appearance in the absence of prior knowledge about the person." To be clear: the students did not know that the people in the photos were congressional candidates, or which was a Democrat.

To answer the following parts you will need to refer to the regression results on the final page of this exam. You may want to tear out that page for ease of reference.

$\boxed{5}$ (a) What is the sample correlation between `Dcomp` and `Dshare`?

$\boxed{5}$ (b) Construct an approximate 95% confidence for the slope in a linear regression in which `Dcomp` *alone* is used to predict `Dshare`. Briefly interpret your results.

$\boxed{5}$ (c) Continuing from the previous part, how accurately does `Dcomp` predict `Dshare`?

$\boxed{5}$ (d) Suppose you wanted to test the null hypothesis that Democratic congressional candidates do equally well, in terms of vote share, in blue and red states against the two-sided alternative. What is the value of your test statistic? Approximately that is the p-value of the test? Briefly interpret your results.

$\boxed{5}$ (e) Is there evidence that the relationship between `Dcomp` and `Dshare` differs in red and blue states? Justify your answer.

$\boxed{15}$ (f) Based on the full set of regression results, do you agree with the claim that "inferences of competence from faces predict election outcomes?" Justify your answer using the tools you have learned in Econ 103. Clear, concise answers will be treated more favorably than long-winded ones.

Regression #1
```
lm(formula = Dshare ~ Dcomp, data = face)
            coef.est coef.se
(Intercept) 0.34     0.03
Dcomp       0.33     0.06
---
n = 118, k = 2
residual sd = 0.13, R-Squared = 0.19
```

Regression #2
```
lm(formula = Dshare ~ blueState, data = face)
            coef.est coef.se
(Intercept) 0.46     0.02
blueState   0.12     0.03
---
n = 118, k = 2
residual sd = 0.14, R-Squared = 0.15
```

Regression #3
```
lm(formula = Dshare ~ blueState + Dcomp, data = face)
            coef.est coef.se
(Intercept) 0.33     0.03
blueState   0.09     0.02
Dcomp       0.28     0.06
---
n = 118, k = 3
residual sd = 0.13, R-Squared = 0.27
```

Regression #4
```
lm(formula = Dshare ~ blueState + Dcomp + blueState:Dcomp, data = face)
                coef.est coef.se
(Intercept)      0.30     0.04
blueState        0.18     0.07
Dcomp            0.35     0.08
blueState:Dcomp -0.18     0.13
---
n = 118, k = 4
residual sd = 0.13, R-Squared = 0.29
```