Final Examination

Econ 103, Statistics for Economists

December 19th, 2013

You will 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:	20	30	30	60	30	30	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. (20 points) A startup is developing apps using three different operating systems: Microsoft Windows, Mac OSX, and Linux. On the first trial, apps compiled under Linux crash 10% of the time, compared to 20% of the time for Mac OSX and 30% of the time for Windows. Of the ten computers at the startup six run Linux, three run Mac OSX and one runs Windows. Sarah works at the startup and was randomly assigned one of these computers. Her app crashed on the first trial. Given this information, what is the probability that she was assigned a Windows machine? (Let C be the event that Sarah's app crashes, W that she was assigned Windows, M Mac OSX and L Linux.)

2. Let X and Z be random variables such that $E[XZ] = 0$, $E[X] = E[Z] = 0$, and $Var(X) = Var(Z) = \sigma^2$. Define $Y = \alpha + X + Z$ where α is an unknown constant.

   (a) (3 points) Calculate $E[Y]$.

   (b) (3 points) Calculate $Cov(X, Z)$.

   (c) (3 points) Calculate $Var(Y)$.

(d) (5 points) Suppose that we do not observe X or Z but we *do* observe Y. If we use Y as an estimator of α, what is our mean-squared error (MSE)?

(e) (8 points) Now suppose that we observe *both* Y and X but not Z and use the difference $Y - X$ to estimate α. Compare this estimator to Y from the previous part in terms of bias, MSE and, if applicable, efficiency. Which should we prefer?

(f) (8 points) As in the previous part, suppose that we observe *both* Y and X but not Z. Now, however, suppose we want to estimate α^2 rather than α. Is $(Y - X)^2$ an unbiased estimator? If not, calculate the bias and explain its direction.

3. Let $X \sim N(-1, 1)$ independently of $Y \sim N(1, 1)$.

   (a) (3 points) What R command would you use to calculate the probability that X takes on a positive value? Approximately what result would you get?

   (b) (3 points) What R command would you use to calculate the probability that Y takes on a positive value? Approximately what result would you get?

   (c) (6 points) Suppose I generate a random variable Z using the following steps. First, I make one draw each from X and Y. Then I independently draw Q, a Bernoulli(1/2) random variable. If $Q = 1$, then I set Z equal to the draw from X. Otherwise I set Z equal to the draw from Y. Thus $Z = Q \times X + (1 - Q) \times Y$. Write an R function called `draw.z` that simulates one draw from the distribution of Z.

(d) (4 points) Continuing from the previous part, write R code to carry out a Monte Carlo simulation with 10000 replications to calculate the probability that Z takes on a positive value.

(e) (8 points) Using your answers to parts (a) and (b) above, approximately what result would you get if you ran the code from the previous part? Prove your answer.

(f) (6 points) Continuing from the previous three parts, suppose I make a draw from Z. It is a positive number. Calculate the probability that Q took on the value 1.

4. This question is based on a dataset containing the results of the tae kwon do event in the 2004 Athens Olympics. (In case this event is unfamiliar to you, my dictionary defines tae kwon do as "a modern Korean martial art similar to karate.") 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 an R data table called `taekwondo`. Each row corresponds to a *single bout* in the competition. The columns are as follows:

`class`	weight class of the 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

Here are the first few rows of the dataset:

```
> head(taekwondo)
      class red.id blue.id   round winner
1 under 58kg   5816    5818 last 16   Blue
2 under 58kg   5817    5824 last 16   Blue
3 under 58kg   5819    5825 last 16    Red
4 under 58kg   5820    5822 last 16    Red
5 under 58kg   5821    5827 last 16    Red
6 under 58kg   5828    5823 last 16    Red
                  method red.points blue.points
1                 Points          9           5
2                 Points          3           5
3                 Points         15          16
4                 Points         14          15
5                 Points         13          12
6 Referee Stopped Contest         NA          NA
```

Name: ____________________ Student ID #: ____________________

(a) (4 points) For the rest of the question we'll restrict attention to the "last 16" round of the competition. This ensures 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`.

(b) (6 points) 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`.

(c) (10 points) As it happens there are 32 bouts in `last16`, 8 bouts for each weight class times 4 weight classes, 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. Use the "refined" interval. Do your results suggest that wearing one color versus the other conveys a competitive advantage? Explain.

(d) (10 points) 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 using the refined test. (Again, this is based on the approximation provided by the CLT.) Approximately what is your p-value for this test? Explain your results.

(e) (6 points) 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.

(f) (4 points) 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?

(g) (10 points) 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`.

(h) (10 points) 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.

5. Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\mu_X, 1)$ independently of $Y_1, \ldots, Y_m \overset{\text{iid}}{\sim} N(\mu_Y, 1)$ and we want to test $H_0\colon \mu_X = \mu_Y$ against the two-sided alternative. Frame the comparison as "$X - Y$" rather than the reverse and let $\bar{X}_n = (\sum_{i=1}^n X_i)/n$ and $\bar{Y}_m = (\sum_{j=1}^m Y_j)/m$.

   (a) (4 points) What is the appropriate test statistic for this problem? What is its sampling distribution under the null hypothesis?

   (b) (4 points) Suppose we choose $\alpha = 0.05$. What is the approximate critical value for our test? What is our decision rule?

   (c) (8 points) Calculate the sampling distribution of your test statistic from part (a) *when the null is false*. Express your answer in terms of n, m, μ_X and μ_Y.

   (d) (14 points) Now suppose that $n + m = 100$ but we're free to choose n. Whatever value we choose for n, we set $m = 100 - n$. (For example, perhaps we're running an experiment with 100 subjects, and are free to choose how many to assign to the control group X.) What value of n *maximizes* the power of our test? Explain.

6. Earlier in the semester, I constructed four regression models to see how well I could predict scores on the first midterm using information available to me *before* you took the exam itself. Specifically, I predicted `midterm1`, a given student's percentage score on the first midterm, using `diagnostic`, the student's percentage score on the math diagnostic test, and `active`, a "dummy" variable that takes on the value 1 if the student was active on Piazza and 0 otherwise. Here are the first few rows of the dataset:

```
> head(m1predict)
  midterm1 diagnostic active
1       54         68      1
2       64         66      1
3       69         57      1
4       60         96      0
5       61         34      0
6       76         58      1
```

All regression results appear in Table 1 on the last page of this exam. You may find it helpful to tear out the page of regression results so you can consult it while answering the following questions.

(a) (5 points) Use the regression results to construct an approximate 95% confidence interval for the difference of mean scores on midterm one between students who were active on Piazza and those who were not (Active - Inactive). Explain your results.

(b) (5 points) Based on the results of Regression 2, is there any evidence that students who do well on the math diagnostic test tend to do better on the first midterm? If so, about how much better? Explain briefly.

Name: ______________________ Student ID #: ______________________

(c) (5 points) Based on the results of Regression 3, is there evidence that, even after controlling for math diagnostic test results, students who are active on Piazza do better on the first midterm? Explain.

(d) (5 points) Sara was inactive on Piazza but got a 90% on the math diagnostic test. Kevin was active but only got a 75% on the diagnostic. Based on Regression 3, who would we predict will earn a higher score on midterm one? How much higher?

(e) (5 points) Do the regression results provide any evidence that the relationship between math diagnostic test results and midterm one scores differs according to whether or not a student was active on Piazza? Explain briefly.

(f) (5 points) Compare the predictive accuracy of the four regression models. How accurate is the most accurate model compared to the least accurate model? Which model would you choose to predict midterm scores and why? Explain briefly.

Table 1: Regression Results

Regression 1:

```
lm(formula = midterm1 ~ active)
            coef.est coef.se
(Intercept) 66.75     2.37
active       9.19     3.55
---
n = 79, k = 2
residual sd = 15.69, R-Squared = 0.08
```

Regression 2:

```
lm(formula = midterm1 ~ diagnostic)
            coef.est coef.se
(Intercept) 47.81     7.72
diagnostic   0.34     0.11
---
n = 79, k = 2
residual sd = 15.45, R-Squared = 0.11
```

Regression 3:

```
lm(formula = midterm1 ~ active + diagnostic)
            coef.est coef.se
(Intercept) 44.16     7.56
active       9.00     3.37
diagnostic   0.33     0.11
---
n = 79, k = 3
residual sd = 14.87, R-Squared = 0.18
```

Regression 4:

```
lm(formula = midterm1 ~ active + diagnostic + active:diagnostic)
                  coef.est coef.se
(Intercept)       45.04     9.41
active             6.62    15.52
diagnostic         0.32     0.13
active:diagnostic  0.04     0.22
---
n = 79, k = 4
residual sd = 14.96, R-Squared = 0.19
```