# 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.)

**Solution:** We need to calculate $P(W|C)$. By Bayes' Rule

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

By the law of total probability,

$$\begin{aligned} P(C) &= P(C|W)P(W) + P(C|M)P(M) + P(C|L)P(L) \\ &= 0.3 \times 0.1 + 0.2 \times 0.3 + 0.1 \times 0.6 \\ &= 0.03 + 0.06 + 0.06 \\ &= 0.15 \end{aligned}$$

Hence,

$$P(W|C) = \frac{P(C|W)P(W)}{P(C)} = \frac{0.3 \times 0.1}{0.15} = 1/5 = 0.2$$

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 $\alpha$ is an unknown constant.

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

   **Solution:** By the Linearity of Expectation: $E[Y] = \alpha + E[X] + E[Z] = \alpha$

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

   **Solution:** By the Shortcut Formula for Covariance: $Cov(X, Z) = E(XZ) - E(X)E(Z) = 0$

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

Name: _____ Student ID #: _____

**Solution:** Since $Cov(XZ) = 0$ and $\alpha$ is a constant, $Var(Y) = Var(X + Z) = Var(X) + Var(Z) = 2\sigma^2$

(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 $\alpha$, what is our mean-squared error (MSE)?

**Solution:** As we calculated above $E[Y] = \alpha$ which means that $Y$ is an unbiased estimator of $\alpha$. Hence, the mean-squared error of this estimator simply equals its variance: $2\sigma^2$, as calculated above.

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

**Solution:** Since $Y-X = \alpha+Z$, $E[Y-X] = \alpha$ and $Var(Y-X) = Var(Z) = \sigma^2$. Hence $Y-X$ is, like $Y$, an unbiased estimator of $\alpha$. Thus, its mean-squared error simply equals its variance: $\sigma^2$. The mean-squared error of $Y - X$ is half that of $Y$. Since both estimators are unbiased, we can ask which is more efficient. Since $Y - X$ has half the variance of $Y$, it is *twice* as efficient. If we observe $X$, it's a *much* better idea to use $Y - X$ rather than $Y$ to estimate $\alpha$ since our estimate will be much less variable.

(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 $\alpha^2$ rather than $\alpha$. Is $(Y - X)^2$ an unbiased estimator? If not, calculate the bias and explain its direction.

**Solution:** By the Linearity of Expectation and the Shortcut Formula for Variance:

$$\begin{aligned} E\left[(Y-X)^2\right] &= E\left[(\alpha + Z^2)\right] = E[\alpha^2 + 2\alpha Z + Z^2] \\ &= \alpha^2 + 2\alpha E[Z] + E[Z^2] \\ &= \alpha^2 + 0 + \left[Var(Z) + E(Z)^2\right] \\ &= \alpha^2 + \sigma^2 \end{aligned}$$

Hence $E\left[(Y-X)^2 - \alpha^2\right] = \sigma^2$ so this estimator is biased. Because variances cannot be negative, the bias is positive. On average, this estimator gives values that are too large.

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

Name: _____ Student ID #: _____

(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?

**Solution:** $P(X > 0) = P(X + 1 > 1) =$ `1 - pnorm(1)` $\approx 0.16$ by the symmetry of the normal distribution and the fact that approximately 68% of the probability density of a standard normal lies in the interval $[-1, 1]$.

(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?

**Solution:** $P(Y > 0) = P(Y - 1 > -1) =$ `1 - pnorm(-1)` $\approx 0.84$ by the symmetry of the normal distribution and the fact that approximately 68% of the probability density of a standard normal lies in the interval $[-1, 1]$.

(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 $Z$.

**Solution:**

```
draw.z <- function(){
  x <- rnorm(1, mean = -1, sd = 1)
  y <- rnorm(1, mean = 1, sd = 1)
  q <- rbinom(1, size = 1, prob = 0.5)
  z <- q * x + (1 - q) * y
  return(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.

**Solution:**

```
sims <- replicate(10000, draw.z())
sum(sims > 0)/length(sims)
```

(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.

**Solution:** By the Law of Total Probability:

$$\begin{aligned} P(Z>0) &= P(Z>0|Q=1)P(Q=1) + P(Z>0|Q=0) \\ &= P(Z>0|Q=1)\times 1/2 + P(Z>0|Q=0)\times 1/2 \\ &= P(X>0)\times 1/2 + P(Y>0)\times 1/2 \\ &\approx 1/2\times[0.16+0.84] = 0.5 \end{aligned}$$

In fact this answer is exact rather than approximate which we can show via the symmetry of the normal distribution if we plug in the `pnorm` commands rather than their approximate values.

(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.

**Solution:** By Bayes' Rule,

$$\begin{aligned} P(Q=1|Z>0) &= P(Z>0|Q=1)P(Q=1)/P(Z>0) \\ &= P(X>0)P(Q=1)/P(Z>0) \\ &\approx 0.16\times 0.5/0.5 = 0.16 \end{aligned}$$

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 dataframe 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
```

(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 dataframe called `last16`.

**Solution:**

```
last16 <- subset(taekwondo, round == "last 16")
```

(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`.

**Solution:**

```
n <- length(last16$winner)
n.blue <- sum(last16$winner == '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.

**Solution:**

$$\begin{aligned} \tilde{p} &= (19+2)/(32+4) = 21/36 \approx 0.583 \\ \tilde{SE}(\tilde{p}) &= \sqrt{\tilde{p}(1-\tilde{p})/(n+4)} \\ &= \sqrt{\left(\frac{21}{36} \times \frac{15}{36}\right)/36} \approx 0.082 \end{aligned}$$

Hence, the CI is approximately $0.583 \pm 2 \times 0.082$ or roughly $(0.42, 0.75)$. We do not find convincing evidence that either color conveys an advantage. If we absolutely had to guess, we would say that blue might convey a slight advantage but our results are perfectly consistent with the reverse as well: the difference between the estimated proportion and 0.5 could easily be nothing more than sampling variability.

(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.

**Solution:** The test statistic is:

$$T = \frac{\widehat{p} - 0.5}{\sqrt{0.5^2/n}} = \frac{19/32 - 0.5}{\sqrt{0.25/32}} \approx 1.06$$

If the test statistic were *exactly* one, the p-value for a two-sided test would be `2 * (1 - pnorm(1))`$\approx 2 \times 0.16 = 0.32$. The test statistic here is slightly larger

than one, so the p-value should be slighly smaller than 0.32. This is a very large p-value: we would *fail* to reject the null at any of the standard significance levels (i.e. 10%, 5%, 1%). We have not found convincing evidence that wearing either color conveys a competitive advantage.

(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 dataframe 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.

**Solution:**

```
last16.points <- subset(last16, method == 'Points')
D <- last16.points$blue.points - last16.points$red.points
```

(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?

**Solution:** $11.7 - 10.1 = 1.6$

(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 dataframe `last16.points`.

**Solution:** Rearranging the formula from class and substituting values from the question statement:

$$\begin{aligned} s_d^2 &= s_x^2 + s_y^2 - 2s_x s_y r_{xy} \\ 2s_x s_y r_{xy} &= s_x^2 + s_y^2 - s_d^2 \\ r_{xy} &= \frac{s_x^2 + s_y^2 - s_d^2}{2s_x s_y} \\ &= \frac{17 + 31 - 25}{2\sqrt{17 \times 31}} = \frac{23}{2 \times \sqrt{527}} \approx 0.5 \end{aligned}$$

(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 dataframe `last16.points`. Be sure to report: (i) the test statistic, (ii) the decision rule, and (iii) the result of the test.

**Solution:** This is matched pairs data: the score earned by the red fighter in a given bout cannot possibly be independent of the score earned by the blue fighter *in the same bout*. The test statistic is

$$T = \frac{\bar{D}}{s_d/\sqrt{n}} = 1.6/(5/\sqrt{29}) \approx 1.7$$

For a 5% test, the decision rule is: Reject $H_0$ if $|T| > 2$. In this case we fail to reject the null hypothesis.

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?

   **Solution:** The test statistic is

   $$T = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{1/n + 1/m}}$$

   and it has a standard normal 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?

   **Solution:** The critical value is approximately 2 so the decision rule is: reject $H_0 \colon \mu_X = \mu_Y$ provided that $|T| > 2$. (It's also fine to write greater than or equals because the sampling distribution is continuous.)

   (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 $n, m, \mu_X$ and $\mu_Y$.

**Solution:** Regardless of the true value of $\mu_X - \mu_Y$,

$$\bar{X}_n - \bar{Y}_m \sim N(\mu_X - \mu_Y, 1/n + 1/m)$$

by the properties of normal distributions. Dividing by $\sqrt{1/n + 1/m}$,

$$T = \frac{\bar{X}_n - \bar{Y}_m}{\sqrt{1/n + 1/m}} \sim N\left(\frac{\mu_X - \mu_Y}{\sqrt{1/n + 1/m}}, 1\right)$$

(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.

**Solution:** To maximize power, we need to make the distribution of our test statistic under the alternative *as far away as possible* from its distribution under the null. As we saw above, both distributions are normal with variance one. They only differ in their means: under the null the mean is zero, while under the alternative the mean is $(\mu_X - \mu_Y)/\sqrt{1/n + 1/m}$. Thus, to maximize power it suffices to make $(\mu_X - \mu_Y)/\sqrt{1/n + 1/m}$ as far away from zero as possible. The numerator isn't under our control but the denominator is. Thus, it suffices to *minimize* $\sqrt{1/n + 1/m}$ subject to the constraint $n+m = 100$. This is equivalent to minimizing $1/n + 1/(n - 100)$ since $\sqrt{x}$ is strictly increasing on $[0, \infty)$. The first order condition is: $-n^{-2} + (100 - n)^{-2} = 0$. Rearranging:

$$\begin{aligned} n^2 &= (100 - n)^2 \\ n^2 &= 100^2 - 200n + n^2 \\ 200n &= 100^2 \\ n &= 50 \end{aligned}$$

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.

**Solution:** Using the results of Regression 1, we see that the difference of means was approximately 9.2 points with a standard error of about 3.6 points, hence $9.2 \pm 7.2$ or equivalently $(2, 16.4)$. Our data suggest that students who are active on Piazza tend to do better on the first midterm.

(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.

**Solution:** An approximate 95% confidence interval for the coefficient on `diagnostic` in Regression 2 is $0.34 \pm 2 \times 0.11 = (0.56, 0.12)$. This is evidence of a positive relationship between math diagnostic test scores and scores on midterm one. For two students who differed by one percentage point in their score on the math diagnostic test, we'd predict that the student with the higher score would score about 1/3 of a point higher on the first midterm.

(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.

**Solution:** Yes. The coefficient on `active` is the difference of intercepts for the two regression lines. In other words, this is the difference in scores on midterm one that we would predict between two students who both earned the same score on the diagnostic test if only one of them was active on Piazza

(Active - Inactive). An approximate 95% confidence interval for this difference is $9 \pm 2 \times 3.4 = (2.2, 15.8)$.

(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?

**Solution:** Since Sara was inactive on Piazza, the regression line we use to predict her midterm score is

$$44.2 + 0.33 \times \texttt{diagnostic} = 44.2 + 0.33 \times 90 \approx 74$$

Since Kevin was active on Piazza, the regression line we use to predict his midterm score is

$$44.2 + 9 + 0.33 \times \texttt{diagnostic} = 53.2 + 0.33 \times 75 \approx 78$$

We would predict that Kevin will do about 4 points better on midterm one.

(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.

**Solution:** To answer this, we look at the results of Regression 4. The coefficient for `active:diagnostic` is the difference of *slopes* for the two lines corresponding to active and inactive students (Active - Inactive). An approximate 95% confidence interval for this difference of slopes is $0.04 \pm 2 \times 0.22 = (-0.4, 0.48)$. We find no evidence of a difference of slopes.

(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.

**Solution:** The most accurate is Regression 3, which predicts to an accuracy of about 14.9 percentage points. The least accurate is Regression 1 which predicts to an accuracy of about 15.7 percentage points. The differences in predictive accuracy between the models aren't especially large in this example: less then one percentage point. Various arguments could be made in favor of any of the four. The point is to say something sensible and demonstrate an understanding of the problem.

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
```

Name: _____ Student ID #: _____