# Final Examination

## Econ 103, Statistics for Economists

May 4th, 2015

**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: ______________________________

Signature: ______________________________

Student ID #: ______________ Recitation #: ______________

| Question: | 1 | 2 | 3 | 4 | 5 | Total |
|---|---|---|---|---|---|---|
| Points: | 30 | 35 | 20 | 35 | 80 | 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. Mark each statement as TRUE or FALSE. If FALSE, provide a one sentence explanation.

$\boxed{3}$ (a) All else equal, a 99% confidence interval is narrower than a 95% interval.

**Solution:** FALSE: it is *wider.*

$\boxed{3}$ (b) A p-value $> 0.05$ implies that we would reject the null hypothesis with $\alpha = 0.05$.

**Solution:** FALSE: we would reject if the p-value is *less* than 0.05.

$\boxed{3}$ (c) If you reject $H_0$ with $\alpha = 0.05$, you would also have rejected with $\alpha = 0.1$.

**Solution:** TRUE

$\boxed{3}$ (d) If zero lies outside a 90% confidence interval for $\mu$, this implies that we would reject $H_0\colon \mu = 0$ with $\alpha = 0.1$ against the two-sided alternative.

**Solution:** TRUE

$\boxed{3}$ (e) If $\alpha$ is the Type I error rate for a hypothesis test, $1 - \alpha$ is the power of that test.

**Solution:** FALSE: power is one minus the Type *II* error rate.

$\boxed{3}$ (f) For any *mutually exclusive* events $A$ and $B$ we have $P(A \cup B) = P(A)P(B)$.

**Solution:** FALSE: $P(A \cup B) = P(A) + P(B)$ under this assumption.

$\boxed{3}$ (g) For any two events $A$ and $B$, $P(A|B)/P(B|A) = P(A)/P(B)$.

**Solution:** TRUE

$\boxed{3}$ (h) The pmf $p(x)$ of a discrete random variable $X$ gives $P(X = x)$.

**Solution:** TRUE

$\boxed{3}$ (i) For any continuous random variable $X$, $P(X \leq 0) = P(X < 0)$.

**Solution:** TRUE

$\boxed{3}$ (j) For any two random variables $X$ and $Y$, $E[XY] = E[X]E[Y]$.

**Solution:** FALSE: this only holds if $Cov(X,Y) = 0$.

2. For each part, write your answer in the space provided. No explanation is needed.

3 (a) What result will I get if I run `pnorm(10, mean = 10, sd = 5)` in R?

**Solution:** 0.5

3 (b) Write an R command to calculate the median of a $\chi^2(2)$ random variable.

**Solution:** `qchisq(0.5, df = 2)`

3 (c) Approximately what result will I get if I run `qnorm(0.16)` in R?

**Solution:** -1

3 (d) Given a dataframe called `grades` with columns `exam1` and `exam2`, write out the full R command to run a regression predicting `exam2` from `exam1`.

**Solution:** `lm(exam2 ~ exam1, data = grades)`

3 (e) Write a single line of R code to display the 4th row of a dataframe called `studentdata`.

**Solution:** `studentdata[4,]`

5 (f) Given a dataframe called `studentdata` with a column called `exam1`, write a single line of R code to display data for all students who scored above 70 on `exam1`.

**Solution:** `subset(studentdata, midterm1 > 70)`

5 (g) Write a single R command to draw three numbers at random from the digits 0–9 *with replacement.*

**Solution:** `sample(0:9, size = 3, replace = TRUE)`

5 (h) Write R code to plot the pdf of a standard normal random variable between -3 and 3 using a grid of $x$-values with a step size of 0.01.

**Solution:**

```
x <- seq(from = -3, to = 3, by = 0.01)
plot(x, dnorm(x), type = ’l’)
```

5 (i) Write an R function called `zscores` that takes a vector `x` as its only input and outputs the z-scores of `x`. You may use any R functions that you like in your answer and may assume that there are no missing values.

**Solution:**

```
zscores <- function(x){
 return((x - mean(x))/sd(x))
}
```

3. Let $Y_1, \ldots, Y_7 \sim$ iid $N(\mu = -3, \sigma^2 = 9)$.

3 (a) Let $X = 1 + Y_1/3$. What kind of random variable is $X$? You do not need to explain your answer, but be sure to specify any and all relevant parameters.

**Solution:** $N(0, 1)$

3 (b) Let $W = (Y_2 + Y_3 + Y_4 + Y_5 + Y_6 + Y_7)/6$. What kind of random variable is $W$? You do not need to explain your answer, but be sure to specify any and all relevant parameters.

**Solution:** $N(\mu = -3, \sigma^2 = 3/2)$

3 (c) What kind of random variable is $X + W$? You do not need to explain your answer, but be sure to specify any and all relevant parameters.

**Solution:** $N(\mu = -3, \sigma^2 = 5/2)$

3 (d) Let $Z = X^2$. What kind of random variable is $Z$? You do not need to explain your answer, but be sure to specify any and all relevant parameters.

**Solution:** $\chi^2(1)$

8 (e) Calculate $E[X^2]$. Briefly explain your reasoning.

> **Solution:** By the shortcut formula $Var(X) = E[X^2] - E[X]^2$. Now, since $X$ is standard normal we know that its variance is one and its mean is zero. Substituting this information, we have $1 = E[X^2]$.

4. On Monday Rodrigo arrives at the office and informs Yiwen that, over the weekend, he has developed extra-sensory perception (ESP). He claims to be able to predict the outcome of a coin-flip with *better* than 50% accuracy. Yiwen is dubious and proposes and experiment in which Rodrigo will be asked to predict the outcomes of 100 flips of a fair coin. She asks Rodrigo to write down his predictions, in order, and seal them in an envelope. To make sure that the experiment is fair, she enlists Rossa to carry out the 100 coin flips while she and Rodrigo both watch. Rossa then opens the envelope and reads the predictions: Rodrigo has successfully predicted 51 of the 100 coin flips.

$\boxed{15}$ (a) Rodrigo claims that the results of the experiment prove that he has ESP but Yiwen isn't convinced. She decides to use what we've learned about tests for proportions in Econ 103 to test the null hypothesis that Rodrigo is just guessing the outcomes of the coin flips at random against the two-sided alternative with $\alpha = 0.05$. To ensure that her test is as accurate as possible, Yiwen *fully imposes the null* when specifying her test statistic. What test statistic does Yiwen use? What is its sampling distribution under the null? What is the numeric value of her test statistic? What is the critical value for her test? What is the outcome of the test?

> **Solution:** The test statistic is $(\widehat{p} - 0.5)/\sqrt{(0.5)^2/100}$. Substituting the fact that Rodrigo got 51/100 correct, the value of the test statistic is 0.2. Under the null hypothesis, this is a realization from a distribution that is approximately $N(0, 1)$, by the CLT. The approximate critical value for a two-sided test with $\alpha = 0.05$ based on the normal distribution is 2. Since 0.2 is less than 2, we fail to reject the null hypothesis that Rodrigo is simply guessing at random.

$\boxed{5}$ (b) Rodrigo objects to Yiwen's procedure claiming that she should have tested against the *one-sided* alternative. Re-do the preceding part using the one-sided alternative and briefly explain which, if any, of the following items will change: the test statistic, the sampling distribution of the test statistic under the null, the numeric value of the test statistic, the critical value, and the outcome of the test.

> **Solution:** The only thing that changes is the critical value: it is now *smaller* than two. To find its precise value, we need to use R: `qnorm(0.95)`. However, we know that the critical value must be greater than one since `pnorm(1)`$\approx 0.84$ is one of the values we have memorized from the Empirical Rule. Thus we would

still fail to reject the null.

15 (c) Let $p$ denote Rodrigo's accuracy in predicting coin flips and assume, for the sake of argument, that he really does have ESP so that $p > 1/2$. (For example, if he correctly predicts the outcome of a coin flip 60% of the time, then $p = 0.6$.) Calculate the approximate power of a one-sided test with $\alpha = 0.16$ based on Yiwen's experiment as a function of $p$. Note that your answer should be an R command *that depends on* $p$, not a specific numeric value.

**Solution:** The first step is to work out the rejection rule in this case. The one-sided critical value for $\alpha = 0.16$ is one so we reject if $(\widehat{p} - 0.5)/(\sqrt{0.5^2/100}) > 1$. Rearranging, this is equivalent to rejecting if $\widehat{p} > 0.55$. The power of the test is the probability that this event occurs as a function of $p$. Whatever is the true value of $p$, we have $(\widehat{p} - p)/(\sqrt{p(1-p)/100}) \approx N(0,1)$ by the CLT. In other words: $\widehat{p} \approx N\left(\mu = p, \sigma^2 = p(1-p)/100\right)$. Thus,

$$\begin{aligned}\text{Power}(p) &= P(\widehat{p} > 0.55) = 1 - P(\widehat{p} \leq 0.55)\\ &= 1 - \texttt{pnorm(0.55, mean = p, sd = sqrt(p*(1-p)/100))}\end{aligned}$$

5. This question concerns a dataframe called `earnings` containing data on the height in inches (`height`), sex (`female` = 1 denotes female), and annual earnings in US dollars (`earn`) of a random sample of 1379 individuals. Here are the first few rows:

```
   earn height female
1 50000     74      0
2 60000     66      1
3 30000     64      1
4 50000     63      1
5 51000     63      1
6  9000     64      1
```

and here are some summary statistics:

| | earn | height | female |
|---|---|---|---|
| Mean | 20015 | 67 | 0.62 |
| Median | 16400 | 66 | 1 |
| S.D. | 19764 | 4 | 0.48 |

To answer this question you will need to consult the regression results that appear on the last page of this exam. (You may want to tear out the page of regression results to avoid having to flip back-and-forth.)

5 (a) Is there evidence of skewness in `height` or `earnings`? Explain briefly.

> **Solution:** A rule of thumb for detecting skewness is to compare the sample median to the sample mean. For both of these variables, the median is below the mean, suggesting that both are somewhat right-skewed.

5 (b) About how many of the individuals in this dataset are female? Explain briefly.

> **Solution:** The sample mean of `female` is 0.62. Since this is a dummy variable taking the value one if a given individual is female, this means that means that 62% of the 1379 individuals in the dataset are female, so roughly 855.

10 (c) Who's taller on average in this dataset: males or females? About how much taller? Does the difference in sample means in this dataset provide compelling evidence of a difference of mean heights in the population from which these individuals were sampled? Support your answer by constructing an approximate 95% confidence interval for the difference of population mean heights (female minus male) and interpreting the results.

> **Solution:** To answer this question we use the results from the regression that has `formula = height ~ female`. The coefficient `female` gives the difference of mean heights: females minus males. We see that, on average, females are 5.54 inches shorter than males. The standard error associated with this estimate is 0.15 leading to a margin of error of 0.3 for an approximate 95% confidence interval: $5.54 \pm 0.3$ or equivalently $(5.24, 5.84)$. We have found strong evidence that females are substantially shorter than men in the population: this isn't merely an artifact of sampling error.

15 (d) Who earns more on average in this dataset: males or females? About how much more? Does the difference in sample means in this dataset provide compelling evidence of a difference of mean earnings in the population from which these individuals were sampled? Support your answer by testing the null hypothesis that males and females earn the same amount against the two-sided alternative with $\alpha = 0.05$ and interpreting the results. Be sure to specify the value of the test statistic, the decision rule, critical value, and outcome of the test. Approximately what is the associated p-value?

Name: ____________________ Student ID #: ____________________

**Solution:** To answer this question we use the results from the regression that has `formula = earn ~ female`. The coefficient `female` gives the difference of mean earnings: females minus males. We see that, on average, females earn \$14307 less per year. The standard error associated with this estimate is about 1029 so our test statistic for the null hypothesis that the population mean earnings are the same for males and females is $14307/1029 \approx 14$. The critical value for a two-sided test with $\alpha = 0.05$ is 2. Accordingly, our decision rule is to reject the null *either* for values of the test statistic *greater* then 2 or *less* than -2 since we are carrying out a two-sided test. Since 14 is larger than 2, we reject the null. We know that the probability of a standard normal RV taking on a value outside of $(-3, 3)$ is less than 0.01. Given that our test statistic is more than *four times* larger than 3, the p-value in this case is for all intents and purposes zero. We have found extremely strong evidence that females earn less, on average, than males in the population: this isn't merely an artifact of sampling error.

$\boxed{5}$ (e) What is the sample correlation between `height` and `earn`? Whare the units of this summary statistic?

**Solution:** It's unitless and its value is $\sqrt{0.09} = 0.3$.

$\boxed{5}$ (f) What is the sample covariance between `height` and `earn`? What are the units of this summary statistic?

**Solution:** Its units are inches × dollars and its value is $0.3 \times 4 \times 19764 \approx 23716$.

$\boxed{5}$ (g) What is the value of the estimated intercept for the regression that uses `height` to predict `earn`? What are its units? Briefly explain the meaning of this estimate.

**Solution:** The intercept is -84078.32 and the units are dollars. This is the amount that we would predict a person whose height is zero inches would earn in a year. Clearly this is a completely meaningless quantity!

$\boxed{5}$ (h) What is the value of the estimated slope for the regression that uses `height` to predict `earn`? What are its units? Briefly explain the meaning of this estimate.

**Solution:** The estimated slope is about \$1563 dollars per inch. This means that, for two people who differ by one inch in height, we would predict that the taller person earns about \$1563 dollars more per year.

5 (i) Construct an approximate 95% confidence interval for the regression slope from the previous part and interpret it. In particular, do you find evidence that there is a positive relationship between height and income in the population?

> **Solution:** The margin of error is $2 \times 133.45 \approx 267$ yielding a confidence interval of $1563 \pm 267$ or $(1296, 1830)$. This is very convincing evidence that earnings are positively related to height in the population. Moreover, the effect is very large: the smallest value in this confidence interval would imply a slope of just under \$1300 dollars per additional inch of height!

5 (j) Which more accurately predicts `earnings`: `height` or `female`? Explain briefly.

> **Solution:** The residual standard deviation for the regression using `height` to predict earnings is about 18854, while that of the regression using `female` is a bit lower: 18513. So it appears that `female` is a slightly more accurate as a predictor of `earn` but neither model is particularly good: in each case we're only predicting to an accuracy of around \$19,000.

15 (k) Kevin argues that the results discussed above provide evidence that women are discriminated against in the labor market. He says "the only reason that there is a positive relationship between height and earnings is that women are systematically and unfairly paid less than men and women also happen to be shorter, on average." Amanda does not believe that women are discriminated against and argues that the situation is exactly the *reverse* of what Kevin claims. She says: "taller people earn more than shorter people since being tall is valuable in many careers. Since women are shorter than men, this fully explains why they earn less, on average." Use what you have learned in Econ 103 along with the *full* set of regression results on the final page of this exam to discuss Kevin and Amanda's interpretation of the regressions results and whether you agree with either, both or neither of them. You will be graded on the clarity of your answer and the extent to which it incorporates the tools and concepts we studied in the course. Write your answer using bullet points with *no more than five bullets*: all other things equal, a concise answer will be treated more favorably.

> **Solution:** There are various possible answers. At a minimum students should recognize that they need to consider the final set of regression results, the one with `formula = earn ~ height + female`. By allowing a different intercept in the relationship between earnings and height for males and females, this regression allows us to see whether the difference in earnings between men and women is completely explained by the fact that women are shorter than men.

When both height and female are used to predict earn we there is still a strongly positive relationship between height and income (with an approximate 95% confidence interval of $551 \pm 370$). This is a much smaller estimate that we had without female in the regression, which is explained by the fact that women are shorter on average and also earn less. However, as we see from the estimate for the coefficient female this regression predicts that, when comparing a man and woman of the same height, we would predict that the women earns about $11,255 less per year. This is a very large difference and we have strong evidence that isn't merely sampling variability: the associated 95% confidence interval is $11254 \pm 2898$ or equivalently $(8368, 14164)$. Even after taking into account differences in height, women earn substantially less than men. In this sense, Amanda is incorrect: the difference in earnings between men and women cannot be fully accounted for by differences in height, even if we believed that being taller is intrinsically valuable in the labor market. At the same time, Kevin is incorrect: even after adjusting for differences in earnings between men and women, which could be evidence of discrimination, there is *still* a positive relationship between height and earnings. So what can we say about discrimination? This is an observational dataset: the sex and height were not randomly assigned to individuals. This makes it hard to be sure of what's going on since there are many possible confounding variables. There are various possible arguments one could make here.

```
lm(formula = earn ~ female)
            coef.est  coef.se
(Intercept)  28926.92    811.86
female      -14307.02   1028.65
---
n = 1379, k = 2
residual sd = 18513.23, R-Squared = 0.12
```

```
lm(formula = height ~ female)
            coef.est coef.se
(Intercept) 70.05     0.12
female      -5.54     0.15
---
n = 1379, k = 2
residual sd = 2.70, R-Squared = 0.50
```

```
lm(formula = earn ~ height)
            coef.est  coef.se
(Intercept) -84078.32   8901.10
height        1563.14    133.45
---
n = 1379, k = 2
residual sd = 18853.92, R-Squared = 0.09
```

```
lm(formula = earn ~ height + female)
            coef.est  coef.se
(Intercept)  -9636.63  12953.75
height         550.54    184.57
female      -11254.57   1448.89
---
n = 1379, k = 3
residual sd = 18460.37, R-Squared = 0.13
```