Final Examination

Econ 103, Statistics for Economists

December 11th, 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: ____________________

Student ID #: ____________________

Signature: ____________________

Question:	1	2	3	4	5	6	Total
Points:	30	25	40	15	40	50	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) If $(2, 6)$ is a 95% CI for μ, we do not reject $H_0\colon \mu = 1$ vs. $H_A\colon \mu \neq 1$ with $\alpha = 0.05$.

Solution: FALSE: we would reject $\mu = 1$ since 1 lies outside the CI.

$\boxed{3}$ (b) A Type I error is rejecting a false null hypothesis.

Solution: FALSE: it is rejecting a *true* null hypothesis.

$\boxed{3}$ (c) The smaller the p-value the stronger the evidence against H_0.

Solution: TRUE

$\boxed{3}$ (d) The power of a hypothesis test equals the probability of making a Type II error.

Solution: FALSE: it is one *minus* the probability of making a type II error.

$\boxed{3}$ (e) If A and B are mutually exclusive events then $P(A \cup B) = P(A) + P(B)$.

Solution: TRUE

$\boxed{3}$ (f) If A and B events such that A implies B then $P(A) \leq P(B)$.

Solution: TRUE

$\boxed{3}$ (g) The concept of *efficiency* involves comparing the MSE of biased estimators.

Solution: FALSE: it involves comparing the *variances* of *unbiased* estimators.

$\boxed{3}$ (h) If X is a continuous RV with pdf $f(x)$ then $f(0)$ gives $P(X = 0)$.

Solution: FALSE: if X is continuous $P(X = x) = 0$ for any x.

$\boxed{3}$ (i) If X and Y are two RVs, then $E[XY] = Cov(X, Y) + E[X]E[Y]$.

Solution: TRUE

$\boxed{3}$ (j) If X and Y are discrete RVs then $p_Y(y) = \sum_{\text{all } y} p_{XY}(x, y)$.

Solution: FALSE: the sum should be taken over all x.

2. For each of the following, provide R code to generate the specified result.

4 (a) Suppose I have a data table called `gradebook` with a column called `midterm1`. Write down R code to display only those rows of `gradebook` for which the entry in `midterm1` is at least 80.

Solution:

```
    gradebook[midterm1 >= 80]
```

6 (b) Write code to plot the CDF of a $\chi^2(5)$ RV from 0 to 10 over a grid of 1001 values.

Solution:

```
x <- seq(from = 0, to = 10, by = 0.01)
y <- pchisq(x, df = 5)
plot(x, y, type = ‘l’)
```

5 (c) Write code to create a vector containing 20 simulated rolls of a fair, six-sided die.

Solution: Various possibilities, the simplest of which is

```
sample(1:6, size = 20, replace = TRUE)
```

10 (d) Write an R function called `my.rt` to make one random draw from the $t(\nu)$ distribution. Your function should take one input argument, the degrees of freedom `nu`, and return the random draw. In your answer you may use any R functions you like *except* for `rt`.

Solution:

```
my.rt <- function(nu){
  normal.draw <- rnorm(1)
  chisq.draw <- rchisq(1, nu)
  return(normal.draw / sqrt(chisq.draw / nu))
}
```

3. This question concerns a game played by rolling a fair, six-sided die with sides numbered 1–6. To play the game you roll the die once. Let x denote the number on the side that shows face-up. If x is even you win x dollars but if x is odd you win $2x$ dollars.

$\boxed{4}$ (a) Suppose you were to play this game an extremely large number of times. On average, how much would you win per play?

Solution: The expected winnings in this game are

$$\frac{1}{6}\left[(2+4+6)+2\times(1+3+5)\right] = 30/6 = 5$$

so you will win, on average, 5 dollars per play in a long sequence of plays.

$\boxed{6}$ (b) Your winnings in one play of this game can be viewed as the realization of a discrete random variable Z. Calculate $Var(Z)$.

Solution: By the Shortcut Formula $Var(Z) = E[Z^2] - E[Z]^2$. We have

$$\begin{aligned} E[Z^2] &= \sum_{\text{all } z} z^2 p(z) = \frac{1}{6}\left[(2^2+4^2+6^2)+(2\times 1)^2+(2\times 3)^2+(2\times 5)^2\right] \\ &= \frac{1}{6}\left[(4+16+36)+\left(2^2+6^2+10^2\right)\right] = \frac{1}{6}\left[56+(4+36+100)\right] \\ &= 196/6 \end{aligned}$$

Therefore $Var(Z) = 196/6 - 25 = 23/3 \approx 7.67$

$\boxed{10}$ (c) Suppose you play this game 69 times consecutively. Based on the approximation provided by the CLT, roughly what is the probability that your average winnings will be less than \$4.34 per play?

Solution: Since individual throws of a fair die are independent and identically distributed we have $Z_1, \ldots Z_{69} \sim$ iid with mean 5 and variance 23/3 by parts (a) and (b). The question asks us to approximate the value of $P(\bar{Z} < 4.34)$ where $\bar{Z} = (Z_1 + \cdots + Z_{69})/69$. By the CLT $\bar{Z}$ is approximately normally distributed with mean 5 and standard deviation $\sqrt{(23/3)/69} = 1/3$. We have

$$P(\bar{Z} < 4.34) = P\left(\frac{\bar{Z}-5}{1/3} < -1.98\right) = \texttt{pnorm}(-1.98) \approx 0.025$$

For rest of this question we change the rules of the game as follows: after rolling the die and observing x you are given the option to roll a *second time*. If you choose not to roll a second time, your winnings are calculated as before: x dollars if x is even and $2x$ dollars if x is odd. If you *do* choose to roll again your winnings are calculated in the same way but based on whatever number comes up on your *second roll*. For example if your initial roll is 3 and you re-roll and get a 1 then you win 2 dollars.

5 (d) If you want to maximize your average winnings over a large number of plays of this *modified* game, then you should choose to roll a second time if and only if your first roll is a 1, 2, or 4. Briefly explain why.

Solution: Using the same logic as in part (a) your expected payoff from the second roll is 5 dollars. Thus, you should re-roll if x is such that you would win *less* than 5 dollars by sticking with your first roll. Accordingly the strategy that maximizes your expected payoff is to stick with your first roll if it is a 3, 5, or 6 and to re-roll if it is a 1, 2, or 4.

15 (e) Nina played this game once following the strategy given in the preceding part. She won 10 dollars. Given this information, what is the probability that she chose to roll the die a second time?

Solution: Let A be the event that Nina chose to roll the die a second time and B be the event that she won 10 dollars. We are asked to calculate $P(A|B)$. By Bayes' Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

and by the Law of Total Probability

$$P(B) \quad = \quad P(B|A)P(A) + P(B|A^C)P(A^C)$$

Using the strategy from the preceding part, Nina will re-roll if she gets a 1, 2, or 4 her initial roll. Thus, since the die is fair, $P(A) = 1/2$ so we have

$$P(A|B) = \frac{P(B|A)}{P(B|A) + P(B|A^C)}$$

Now, $P(B|A^C)$ is the probability that Nina won 10 dollars given that she did *not* roll twice. Since she is following the strategy from part (d), if Nina does not roll twice then she must have rolled a 3, 5, or 6 initially. Thus $P(B|A^C) = 1/3$. Now, recall that if Nina rolls twice then her winnings depend on the second roll *only*. This means that $P(B|A)$ equals the probability of winning 10 dollars in the *original* version of the game from parts (a)–(c) above, namely $1/6$. Therefore

$P(A|B) = (1/6)/(1/3 + 1/6) = 1/3$. If Nina wins 10 dollars it is more likely than not that she only rolled once.

4. Clayton wants to know the fraction of Penn students who come from Guam, a small western Pacific island, so he polls a random sample of 96: none come from Guam.

$\boxed{5}$ (a) Apply the "textbook" procedure for constructing an approximate 95% CI for a population proportion based on the CLT to the data given in the problem statement.

Solution: Since $\widehat{p} = 0$, the textbook interval $\widehat{p} \pm \texttt{qnorm}(1 - \alpha/2)\sqrt{\widehat{p}(1-\widehat{p})/n}$ simply gives $(0, 0)$.

$\boxed{5}$ (b) Repeat the preceding part using the *refined* interval.

Solution: To construct the refined 95% interval we add four "fake" observations to the dataset: two zeros and two ones. This gives us a new dataset with 100 observations, two of which are ones so that $\widetilde{p} = 0.02$ and $\widetilde{n} = 100$ and the approximate 95% interval is

$$0.02 \pm 2\sqrt{0.02 \times 0.98/100}$$

which gives $(-0.008, 0.048)$.

$\boxed{5}$ (c) Compare and contrast the two intervals you constructed above. Which makes more sense and why? Briefly explain your answer.

Solution: The textbook interval is nonsensical in this example: it implies that we have complete certainty that absolutely no one at Penn comes from Guam. This is a ridiculous conclusion to draw from the data we have. This is an example where the approximation from the CLT breaks down because the sample size is too small relative to the population proportion we are trying to estimate. (We would expect a priori that very few students at Penn come from Guam.) The refined interval corrects this defect and provides a much more reasonable answer although, naturally, we shouldn't take the LCL literally since a population proportion cannot be negative!

5. Professor Quack has developed a diet plan where you are allowed to eat anything you want as long you wash down every meal with a spoonful of pickle juice. He claims that if you follow this diet you will lose, on average, 3kg over 4 weeks. Matt decides to carry

out an experiment to test this claim. He recruits a random sample of 25 subjects and puts all of them on the "pickle juice diet." Let X_i be person i's weight (in kg) before beginning the diet and Y_i be person i's weight (in kg) after four weeks on the diet. The summary statistics from Matt's experiment are as follows:

	X	Y
Sample Mean	83	82
Sample S.D.	6	10
Correlation	0.3	

Throughout this question please work with the approximation provided by the CLT. For simplicity you may treat this approximation as though it were exact.

3 (a) Is this an independent samples or matched pairs problem? Explain in one sentence.

Solution: Matched pairs: we're interested in how much weight people lose over the course of the diet which involves a before-and-after comparison for each subject.

4 (b) Let L_i denote the (positive) amount of weight that subject i lost over the course Matt's experiment: $L_i = X_i - Y_i$. Calculate $\bar{L}$, the sample mean of the L_i.

Solution: $\bar{L} = \bar{X} - \bar{Y} = 83 - 82 = 1$

6 (c) Continuing from the preceding part calculate S_L^2, the sample variance of the L_i.

Solution: $S_L^2 = S_X^2 + S_Y^2 - 2r_{XY}S_XS_Y = 6^2 + 10^2 - 2 \times 0.3 \times 6 \times 10 = 100$

Suppose Matt decides to test the null hypothesis that population mean weight loss for people on the "pickle juice" diet is zero against the one-sided alternative of positive weight loss at the 2.5% significance level.

$\boxed{3}$ (d) What is the critical value for Matt's test?

Solution: `qnorm(0.975)`≈ 2

$\boxed{4}$ (e) What is the value of Matt's test statistic?

Solution:

$$\frac{\bar{L}}{S_L/\sqrt{n}} = \frac{1}{10/\sqrt{25}} = 0.5$$

$\boxed{2}$ (f) Does Matt reject the null hypothesis at his specified significance level?

Solution: No: his test statistic is smaller than his critical value.

$\boxed{3}$ (g) Write down the R command Matt would use to calculate the p-value for his test.

Solution: `1 - pnorm(0.5)`

Name: ____________________ Student ID #: ____________________

Alyson reads about Matt's results in the prestigious *West Philadelphia Journal of Dietary Science* and decides to replicate his experiment using a larger sample of subjects from the same population. To obtain approval from the Institutional Review Board at Penn, she must carry out a power calculation. In her study Alyson plans to use the same statistical test as Matt, with the same significance level, null, and alternative hypothesis. Analyzing power requires knowledge of the population standard deviation of the L_i. Since Alyson doesn't know this quantity she approximates it using the *sample* standard deviation from Matt's experiment.

15 (h) How large a sample size should Alyson recruit to ensure that the power of her test will be at least 0.84 if Professor Quack's claim is correct, i.e. if the diet causes an average weight loss of 3kg?

Solution: Substituting the value of S_L from Matt's experiment in place of the unknown population value, the test statistic is $\sqrt{n}\bar{L}/10$. Since $\sqrt{n}(\bar{L}-\mu_L)/10 \sim N(0,1)$ it follows that $\sqrt{n}\bar{L}/10 \sim N(\mu_L\sqrt{n}/10, 1)$. If Professor Quack's claim is correct, $\mu_L = 3$. Since Alyson will reject the null whenever the test statistic is greater than 2, we need to find the smallest value of n for which a $N(3\sqrt{n}/10, 1)$ RV will take on a value greater than 2 with probability of at least 0.84. Changing n only shifts the mean. Since the variance equals one, we need to set the mean equal to 3 so that 0.16 of the probability lies below 2 and the rest lies above. Thus we solve $3\sqrt{n}/10 = 3$ for n which gives us $n = 100$. Alyson needs a much larger sample size than Matt used to have the desired power against the alternative that corresponds to Professor Quack's claim.

6. An R data table called `houses` contains the sale price and characteristics of a random sample of 128 houses sold in Kansas City in a single year. The first few rows of the data table are as follows:

```
> head(houses)
  neighborhood offers sqft brick bedrooms bathrooms  price
1            B      3 1990    No        2         2 105600
2            A      3 1900    No        3         3 102500
3            A      3 1860    No        2         2  91100
4            A      2 1780    No        3         2 114600
5            C      3 2150   Yes        4         3 160600
6            C      2 2110    No        3         2 142600
```

In this question we will only work with the columns `sqft`, `brick` and `price`: `brick` is a categorical variable that indicates whether or not the house in question is made of brick, `sqft` gives the size of the house in square feet, and `price` is the sale price of the house in US dollars. The final two pages of this exam contain regression results and plots that relate to this question. You may want to tear them out for easy reference when answering the following. You may assume throughout this question that there are no missing values.

Parts (a) and (b) refer to the *first* of the two plots on the final page of this exam.

5 (a) Give R code to create the plot, including axis labels and title.

Solution:

```
boxplot(price ~ brick, data = houses, xlab = ‘‘Brick House?’’,
                       ylab = ‘‘House Price ($)’’,
                       main = ‘‘House Prices in Kansas City’’)
```

6 (b) Explain what this plot shows using bullet points with no more than three bullets.

Solution: This is a boxplot. It compares the minimum, 25th percentile, median, 75th percentile and maximum of the sale prices of a sample of brick houses in Kansas city to those of non-brick houses. Brick houses cost more.

Suppose I wanted to test the null hypothesis that the average price for brick and non-brick houses in Kansas City are the same against the two-sided alternative.

2 (c) Which set of regression results should I consult?

Solution: Regression #2

3 (d) On average, how much more does a brick house cost in Kansas City?

Solution: About 26,000 more.

3 (e) Approximately what is the p-value of my test?

Solution: The test statistic is roughly $25.8/4.5 \approx 5.7$ so the p-value is less than 0.001.

3 (f) Is there convincing evidence that brick houses cost more? Explain in one sentence.

Solution: Yes: we would resoundingly reject the null hypothesis here even if we chose a tiny value for α.

Suppose I wanted to use square-footage *alone* to predict house prices in Kansas City based on a simple linear regression model.

2 (g) Which set of regression results should I consult?

Solution: Regression #4

8 (h) The second plot on the final page of this exam plots the data and regression line. Give the R code to produce this plot, including all axis labels and the title. You do *not* need to give the code to run the regression: you can use the coefficient values from the regression output I provide when plotting the regression line.

Solution:

```
plot(price ~ sqft, data = houses, xlab = ‘‘Square Feet’’,
                   ylab = ‘‘House Price ($)’’,
                   main = ‘‘House Prices in Kansas City’’)
abline(a = -10091.13, b = 70.23)
```

3 (i) What is the sample correlation between house prices and square-footage?

Solution: $\sqrt{0.31} \approx 0.56$

$\boxed{3}$ (j) Based on the regression results, how much more would we predict that a house would cost if it were 100 square feet larger?

Solution: The regression slope is about 70 dollars per square foot, so we would predict that a house that is 100 square feet larger would cost 7000 dollars more.

$\boxed{3}$ (k) Construct an approximate 95% confidence interval for the regression slope, including the appropriate units.

Solution: Approximately 70 ± 19 or $(51, 89)$ dollars per square foot.

Now suppose I wanted to use both `brick` and `sqft` to predict house prices. There are two ways I could do this: by allowing *only* a different intercept for brick houses or by allowing *both* a different intercept and and a different slope.

$\boxed{3}$ (l) Suppose I only allow a different intercept, *not* a different slope. Based on the appropriate set of regression results, how much larger would a non-brick house have to be for us to predict it to have the same sale price as a brick house?

Solution: From Regression #1 a brick house commands a premium of around 23,450 dollars while the predicted increase in price per additional square foot is about 66 dollars. Dividing, the non-brick house would have to be about 355 square feet larger to command the same price premium.

$\boxed{6}$ (m) Do the regression results provide convincing evidence that brick houses command a higher premium *per square foot* than non-brick houses? Explain briefly in bullet points using no more than two bullets.

Solution: From Regression #3, the slope for brick houses is estimated to be about 25 dollars/sqft greater than for non-brick houses. The approximate 95% CI, however, is about 25 ± 37 dollars per square foot which comfortably includes zero: the evidence is suggestive but not particularly convincing.

Regression #1

```
lm(formula = price ~ brick + sqft, data = houses)
            coef.est coef.se
(Intercept) -9444.29 16577.13
brickYes    23445.10  3709.81
sqft           66.06     8.27
---
n = 128, k = 3
residual sd = 19644.14, R-Squared = 0.47
```

Regression #2

```
lm(formula = price ~ brick, data = houses)
            coef.est  coef.se
(Intercept) 121958.14   2593.50
brickYes     25810.91   4527.59
---
n = 128, k = 2
residual sd = 24051.17, R-Squared = 0.21
```

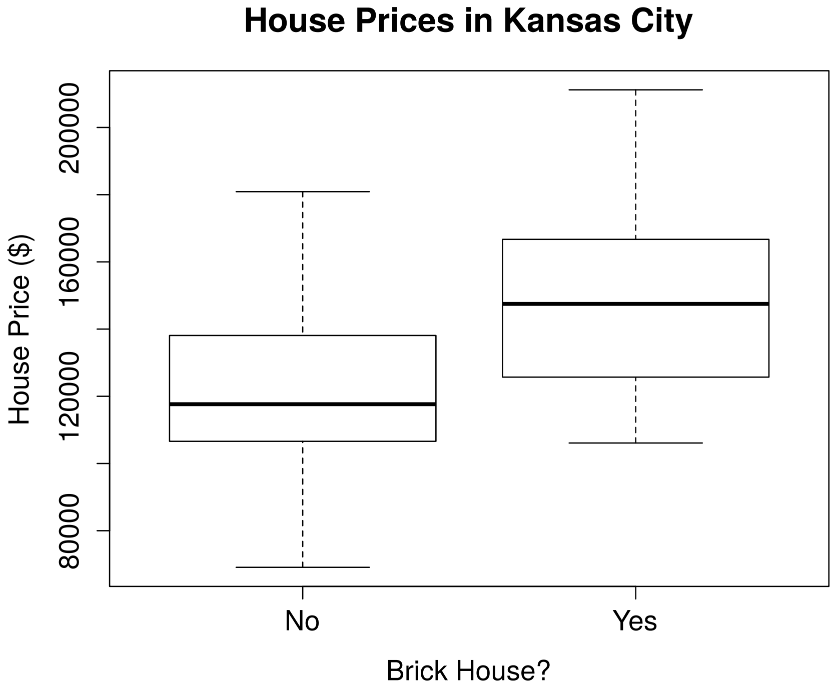
Regression #3

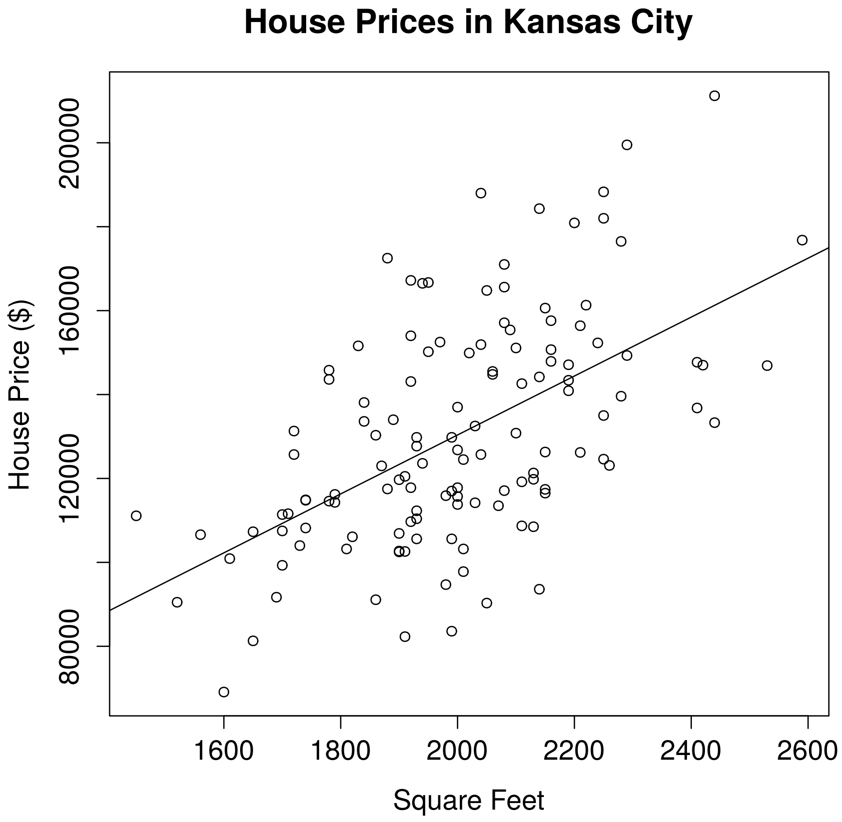
```
lm(formula = price ~ brick + sqft + brick:sqft, data = houses)
              coef.est  coef.se
(Intercept)     4448.23  19396.56
brickYes      -27193.38  37234.31
sqft              59.07      9.69
brickYes:sqft     25.13     18.39
---
n = 128, k = 4
residual sd = 19576.29, R-Squared = 0.48
```

Regression #4

```
lm(formula = price ~ sqft, data = houses)
            coef.est  coef.se
(Intercept) -10091.13  18966.10
sqft            70.23      9.43
---
n = 128, k = 2
residual sd = 22475.53, R-Squared = 0.31
```

Name: ______________________ Student ID #: ______________________





Name: ______________________ Student ID #: ______________________