Final Examination

Econ 103, Statistics for Economists

December 16th, 2014

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	20	20	20	50	70	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. Consider the following simple dataset with nine observations of two variables:

x	y
1	1
2	1
3	1
1	2
2	2
3	2
1	3
2	3
3	3

(a) (4 points) Calculate $\bar{x}$ and $\bar{y}$.

Solution: The sample mean is 2 for both x and y since $3 \times (1+2+3)/9 = 2$.

(b) (4 points) Calculate s_x^2 and s_y^2.

Solution: The calculation is the same for both:

$$3 \times \left[(1-2)^2 + (2-2)^2 + (3-2)^2\right]/8 = 3/4$$

(c) (6 points) Calculate s_{xy}.

Solution:

$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
-1	-1	1
0	-1	0
1	-1	-1
-1	0	0
0	0	0
1	0	0
-1	1	-1
0	1	0
1	1	1

Summing the third column and dividing by $n-1$ gives the covariance. Since the sum is zero, so is the covariance.

(d) (6 points) Calculate the slope and intercept of a linear regression model that uses this dataset to predict y from x.

Name: ____________________ Student ID #: ____________________

Solution: The regression slope is s_{xy}/s_x^2. Since the covariance is zero, so is the regression slope. Since the regression line goes through the means of the data, $\bar{y} = a + b\bar{x}$ but since $b = 0$, we have $a = \bar{y}$.

2. (20 points) Let $Y \sim \text{Bernoulli}(1/3)$ and define X *conditional* on Y as follows: if $Y = 0$ then $X \sim \text{Bernoulli}(3/4)$, otherwise $X \sim \text{Bernoulli}(4/5)$. Write the joint pmf of X and Y in a 2×2 table. Put the X-values in the *rows* and the Y-values in the *columns*.

Solution:

$$\begin{aligned}
P(X=0,Y=0) &= P(X=0|Y=0)P(Y=0) = 1/4 \times 2/3 = 1/6\\
P(X=0,Y=1) &= P(X=0|Y=1)P(Y=1) = 1/5 \times 1/3 = 1/15\\
P(X=1,Y=0) &= P(X=1|Y=0)P(Y=0) = 3/4 \times 2/3 = 1/2\\
P(X=1,Y=1) &= P(X=1|Y=1)P(Y=1) = 4/5 \times 1/3 = 4/15
\end{aligned}$$

So we have:

		Y	
		0	1
X	0	1/6	1/15
	1	1/2	4/15

3. Suppose I take a meter stick and break it into two pieces. The exact point at which I break it, S, is random and follows a Uniform distribution. Thus, the length of the first piece is simply S while the length of the second piece is $1 - S$.

(a) (10 points) Let A be the *area* of a rectangle with sides S and $1 - S$. Calculate the expected value of A.

Solution: Here S is a uniform random variable and we are asked to calculated the expected value of $A = S(1 - S) = S - S^2$. Since the pdf of S is one,

$$E[A] = \int_0^1 (S - S^2)\, dx = \left.\frac{S^2}{2} - \frac{S^3}{3}\right|_0^1 = 1/2 - 1/3 = 1/6$$

which is just under 0.17 squared meters.

(b) (10 points) The R command `runif(n)` draws `n` independent, Uniform$(0, 1)$ random variables. Using this command, write R code to verify your solution to the preceding part via Monte Carlo simulation using 1000 draws.

Solution: Many correct answers. Here's the simplest:

```
S <- runif(1000)
A <- S * (1 - S)
mean(A)
```

4. To pay off his gambling debts to Rodrigo, Rossa has taken a part-time job as a plumber and needs to measure the length of two pipes. When he uses his measuring tape, Rossa makes normally distributed errors with variance σ^2 and mean zero: if an object's true length is ℓ, his measurement is $L \sim N(\ell, \sigma^2)$. Suppose that each measurement error is independent of the others and let ℓ_A, ℓ_B denote the true lengths of pipes A and B.

   (a) (4 points) Rossa decides to start with pipe A. Following the adage "measure twice, cut once," his instinct is to make *two* measurements of the pipe and use the *average* to estimate ℓ_A. Calculate the bias and variance of this estimator.

   Solution: This is just the sample mean of two independent $N(\ell, \sigma^2)$ random variables so it is an unbiased estimator of ℓ with variance $\sigma^2/2$.

   (b) (4 points) Rossa notices that pipe A is clearly longer than pipe B and comes up with an idea: rather than measuring each pipe twice, he'll lay the pipes end to end and measure the sum and difference of their lengths. Let D be Rossa's measurement of the *difference* of lengths, and S be his measurement of the *sum* of lengths. Assume that Rossa lines up the pipes perfectly: when he measures any length (of a single pipe, a sum or a difference) his measurement equals the true length plus a $N(0, \sigma^2)$ error, as above. What is the distribution of D? What is the distribution of S?

   Solution: The true difference of lengths is $\ell_A - \ell_B$ so D is a $N(\ell_A - \ell_B, \sigma^2)$ RV. The true sum of lengths is $\ell_A + \ell_B$ so S is a $N(\ell_A + \ell_B, \sigma^2)$ RV.

   (c) (6 points) Rossa decides to estimate ℓ_A using $(S + D)/2$ and ℓ_B using $(S - D)/2$. Are these estimators unbiased? If so, prove it. If not, calculate the bias of each.

   Solution: By the Linearity of Expectation both are unbiased:

   $$E[(S + D)/2] = (E[S] + E[D])/2 = [(\ell_A + \ell_B) + (\ell_A - \ell_B)]/2 = \ell_A$$

   $$E[(S - D)/2] = (E[S] - E[D])/2 = [(\ell_A + \ell_B) - (\ell_A - \ell_B)]/2 = \ell_B$$

   (d) (6 points) Calculate the variance of the two estimators from the preceding part.

Solution: Using the fact that S and D are independent:

$$Var[(S+D)/2] = [Var(S) + Var(D)]/4 = \sigma^2/2$$

$$Var[(S-D)/2] = [Var(S) + Var(D)]/4 = \sigma^2/2$$

So the variance is *the same* as if Rossa had measured each pipe twice!

5. Petra has a dataframe called `reaction` containing measurements of the reaction times of 19 students given in seconds. Although 19 observations is a relatively small sample size, you may assume for the purposes of this question that the approximation based on the Central Limit Theorem applies. Each row of `reaction` corresponds to an individual: the value in the column `dom` gives that individual's reaction time using her *dominant* hand while the value in the column `nondom` gives her reaction time using her *non-dominant* hand. For example, I am left-handed so my value for `dom` would be my reaction time with my *left hand*. Here are the first six rows of the dataframe and some summary statistics:

```
    dom nondom
1 0.159  0.188
2 0.176  0.194
3 0.180  0.171
4 0.130  0.195
5 0.180  0.199
6 0.121  0.179
```

	dom	nondom
Sample Mean	0.180	0.202
Sample S.D.	0.045	0.048
Correlation	0.83	

(a) (5 points) Give the units of each of the summary statistics from the above table.

Solution: The sample means are measured in seconds as are the sample standard deviations. The sample correlation is unitless.

(b) (5 points) All of the measurements in `reaction` are smaller than a second so Petra runs the R command `reaction <- 1000 * reaction` to convert the dataset to *milliseconds*. Give the updated values for each of the above summary statistics.

Solution: Correlation is unchanged, and everything else is multiplied by 1000:

	`dom`	`nondom`
Sample Mean	180	202
Sample S.D.	45	48
Correlation	0.83	

(c) (5 points) Petra wants to use the data contained in `reaction` to determine whether people's reaction times differ when they use their dominant versus non-dominant hand. Is this a problem based on two independent samples or matched pairs? Explain briefly.

Solution: This is a matched pairs problem. We have *two* measurements of each individual: one in which she uses her dominant hand and another when she uses her non-dominant hand. Thus the two columns *cannot* be independent.

(d) (15 points) Write R code that computes a 90% confidence interval for the difference of population mean reaction times: *non-dominant* minus *dominant*.

Solution: Many possibilities. Here's one:

```
react.diff <- reaction$nondom - reaction$dom
SE <- sd(react.diff) / sqrt(length(react.diff))
ME <- qnorm(0.95) * SE
mean(react.diff) + c(-ME, ME)
```

(e) (15 points) Now suppose that, instead of calculating a confidence interval, Petra wanted to test the null hypothesis that reaction times are *the same* regardless of whether one uses one's dominant or non-dominant hand against the two-sided alternative. Calculate the value of the appropriate test statistic.

Solution: It doesn't matter whether we do the calculation in seconds or milliseconds: the test statistic will take the same value. For simplicity, I'll use milliseconds. The numerator of the test statistic is $\bar{D} = 202 - 180 = 22$ milliseconds. To calculate the denominator, we first need the sample variance of the differences D_i, which we calculate as follows:

$$s_D^2 = s_X^2 + s_Y^2 - 2s_X s_Y r_{XY} = 45^2 + 48^2 - 2 \times 45 \times 48 \times 0.83 = 743.4$$

Thus, we have

$$SE(\bar{D}) = \sqrt{743.4/19} \approx 6.3$$

so the test statistic is $22/6.3 \approx 3.5$.

(f) (5 points) Approximately what is the p-value for Petra's test from the preceding part? What should she conclude?

> **Solution:** The test statistic is about 3.5 which is *very large*: the p-value is definitely smaller than 0.01 so Petra has found strong evidence that people's reaction times are *slower* when they use their non-dominant hand.

6. This question concerns a dataframe called `birthdata` containing observational data on 1000 mothers and their first-born children: `birthweight` is a given child's birth weight in grams, `weeksgest` is the number of weeks between that child's conception and his or her birth (i.e. weeks of gestation), and `smoker` is a dummy variable that takes on the value one if that child's mother smoked during pregnancy. Here are the first few rows:

```
  birthweight weeksgest smoker
1        4252        38      1
2        4229        42      0
3        4338        41      0
4        3850        39      0
5        3430        41      0
6        3260        39      0
```

To answer this question, refer to the regression results on final page of the exam.

(a) (6 points) What is the sample mean birth weight for children whose mother smoked during pregnancy? How does this compare to the sample mean birth weight for children whose mothers did *not* smoke during pregnancy?

> **Solution:** To answer this part, we use the results of Regression #1. The sample mean birth weight for children whose mothers smoked during pregnancy is about $3472 - 293 = 3179$ grams compared to about 3472 grams for children whose mothers did not smoke during pregnancy.

(b) (6 points) Construct an approximate 95% confidence interval for the population mean difference of birth weights between children whose mothers smoked during pregnancy and those whose mothers did not.

> **Solution:** The standard error for the difference of means is approximately 51 grams, so -293 ± 102 or equivalently $(-395, -191)$ is an approximate 95% CI.

(c) (6 points) Suppose you wanted to carry out a two-sided test of the null hypothesis that the children of smokers and non-smokers weigh the same, on average, at birth.

What is the value of your test statistic? Write out the full R command needed to calculate the p-value for this test. Approximately what would be your result?

Solution: Again, using the results of Regression #1, the test statistic is approximately $|-293/51| \approx 5.7$. The R command to calculate the two-sided p-value is `2 * (1 - pnorm(5.7))` which is essentially zero: a standard normal RV *practically never* takes on a value more than 3 std. devs. from its mean.

(d) (6 points) Interpret your results from the preceding two parts. Do they provide evidence of a causal relationship between smoking and birth weight?

Solution: We have found very strong evidence that children born to mothers who smoked during pregnancy weight less at birth. The difference appears to be considerable: on the order of several hundred grams. We need to be cautious about saying more than this, however: we have not proven that smoking *causes* lower birth weight since this data comes from an observational study. It could be that mothers who smoke are also unhealthy in other ways that are more important in influencing birth weight than smoking behavior.

(e) (5 points) What is the sample correlation between `birthweight` and `weeksgest`?

Solution: To answer this, we use the results of Regression #2. The R-squared is 0.2 so the correlation is $\sqrt{0.2} \approx 0.45$.

(f) (6 points) Suppose we wanted to use `weeksgest` *alone* to predict `birthweight`. For two newborns who differ by one week in gestation time, by how much would we predict that their birth weights differ?

Solution: Again using the results of Regression #2, we see that the regression slope is about 113. For each additional week of gestation, we would predict a birth weight that is 113 grams higher.

(g) (5 points) What are the units of the slope in Regression #2?

Solution: Grams per week (of gestation).

(h) (6 points) What is the meaning of the intercept in Regression #2?.

Solution: Taken literally, the intercept tells us that we would predict a birth weight of about -1 kilogram for a child born after zero weeks of gestation. Clearly this is not a meaningful quantity!

(i) (6 points) If you were given the task of predicting birthweight as accurately as possible *either* using `smoker` *or* using `weeksgest` but not both, which would you use? How much more accurate is your preferred model? Explain briefly.

Solution: We should use `weeksgest` rather than `smoker`. The regression using only `smoker` predicts to an accuracy of about 540 grams on average while the regression using only `weeksgest` predicts to an accuracy of about 491 grams on average. The better model is, on average, about 49 grams more accurate.

(j) (6 points) Suppose you wanted to predict `birthweight` using *both* `smoker` and `weeksgest`. Two of the four regressions are relevant for this task, although they differ in the *way* in which they use the information from the two variables. Which models are they, and how do they differ in the relationship they fit between `birthweight` and `weeksgest` depending on the value of `smoker`? In your answer, discuss only the regression *models*, not the *results* of fitting these models to `birthdata`.

Solution: The relevant regressions are #3 and #4. Regression #3 fits a linear relationship between `weeksgest` and `birthweight` in which the intercept is allowed to vary depending on whether the mother is a smoker. Regression #4 expands upon Regression #3 by allowing for a different intercept *and* slope for the relationship between `weeksgest` and `birthweight` depending on whether the mother is a smoker.

(k) (12 points) For each of the models you listed in your answer to the preceding part, use the appropriate regression results to write out the *rule* we would use to predict `birthweight` from `weeksgest` for a child whose mother smoked during pregnancy. Repeat for a child whose mother did *not* smoke during pregnancy.

Solution: Under Regression #3 we predict a birthweight of about $-940 + 112 \times$ `weeksgest` grams for the child of a mother who did *not* smoke. For the child of a mother who *did* smoke, we predict $-1219 + 112 \times$ `weeksgest` grams. Under Regression #4, we predict a birthweight of around $-1069 + 115 \times$ `weeksgest` grams for the children whose mothers did *not* smoke compared to $-607 + 96 \times$ `weeksgest` grams for those children whose mothers *did* smoke.

Regression #1

```
lm(formula = birthweight ~ smoker, data = birthdata)
            coef.est coef.se
(Intercept) 3472.48    18.30
smoker      -292.91    50.96
---
n = 1000, k = 2
residual sd = 540.20, R-Squared = 0.03
```

Regression #2

```
lm(formula = birthweight ~ weeksgest, data = birthdata)
            coef.est coef.se
(Intercept) -1009.00   281.10
weeksgest     112.82     7.13
---
n = 1000, k = 2
residual sd = 490.87, R-Squared = 0.20
```

Regression #3

```
lm(formula = birthweight ~ smoker + weeksgest, data = birthdata)
            coef.est coef.se
(Intercept) -940.49   276.31
smoker      -278.90    45.49
weeksgest    111.99     7.00
---
n = 1000, k = 3
residual sd = 482.11, R-Squared = 0.23
```

Regression #4

```
lm(formula = birthweight ~ smoker + weeksgest + smoker:weeksgest,
    data = birthdata)
                 coef.est coef.se
(Intercept)      -1069.20   303.79
smoker             461.93   728.26
weeksgest          115.26     7.70
smoker:weeksgest   -18.85    18.49
---
n = 1000, k = 4
residual sd = 482.10, R-Squared = 0.23
```