QUEEN'S UNIVERSITY
FACULTY OF ARTS AND SCIENCE

APRIL 2014

ECONOMICS 250
Introduction to Statistics

Instructor: Gregor Smith

Instructions:

The exam is three hours in length.

Do all nine (9) questions.

Be sure to show your calculations and intermediate steps.

Put your student number on each answer booklet.

Formulas and tables are printed at the end of this question paper.

You may use a hand calculator. Allowed calculators include those with blue or gold stickers, the Casio 991, the Sharp EL376S, or other non-programmable calculators. No red-sticker calculators or other aids are allowed.

Proctors are unable to respond to queries about the interpretation of exam questions. Do your best to answer the exam questions as they are written.

1. A stock price takes the value \$1 on 8 days, the value \$3 on 4 days, and the value \$5 on 8 days.

(a) Find the sample mean, standard deviation, and coefficient of variation of the stock price.

(b) Find the interquartile range of the stock price.

(c) Suppose that the exchange rate in euros per dollar is 1.679. What are the mean and standard deviation of the price when it is quoted in euros?

2. Suppose the adult population of a Brazil (in millions) is categorized this way:

	Native Born	Immigrant
Employed	96	20
Unemployed	8	4
Not in the Labour Force	56	16

(a) Find the marginal distributions of employment status (the row variable) and immigration status (the column variable).

(b) Find the distribution of employment status separately conditional on each value of the column variable. Is there a relationship between immigration status and employment status?

3. Suppose that adult human weights, labelled w, are uniformly and continuously distributed between 32 and 104, while adult human heights, labelled h, are uniformly and continuously distributed between 48 and 84. The correlation between the two is 0.8.

(a) Find the mean and variance of weight, w, and height, h.

(b) A researcher proposes an indicator of health given by:

$$y = w - h.$$

Find the mean and variance of the indicator y.

4. A specific cancer affects 1% of the population. A medical test for that illness has a false positive rate of 8%; in other words, 8% of those without the illness will be falsely indicated as having the illness, on the basis of the test. For people who do have cancer, though, the test correctly reveals the illness 90% of the time.

(a) What is the unconditional or marginal probability of a positive test result?

(b) Given a positive test result, what is the probability of having the illness?

5. A number n of unrelated people participate in a clinical trial for a new drug. Suppose the proportion of times that drug is effective in the population is labelled p and that it takes the value 0.4.

(a) Suppose a small study involves $n = 10$. Find the probability that the sample proportion of successes, $\hat{p}$, lies in this range: $0.3 \leq \hat{p} \leq 0.5$.

(b) Now suppose a larger study involves $n = 36$. Use the normal approximation to find the probability that the sample proportion of successes lies between 0.3 and 0.5.

(c) In the second, large study, suppose that you found the drug was effective for 4 people. You do not know the true value of p. Find a 90% confidence interval for the population proportion.

(d) Suppose that in the second study you instead found no successes. What would the width of the confidence interval be in that case? Can you suggest a modification to your methods for this case?

6. An economist is studying the unemployment rate for young workers in Spain. She conducts a survey of 1000 workers and tests the null hypothesis that the unemployment rate is 20% against the alternative that it is greater than 20%.

(a) If the critical value she uses is 22% (*i.e.* she rejects the null if the unemployment rate is above this value) then what is the significance level (α) of the test?

(b) If the true value of the unemployment rate in the population is 23% then what is the power of the test?

7. Suppose that exchange-rate changes, labelled x, can take on the values -2, -1, 0, 1, and 2, each with probability 0.2.

(a) Find the population mean and standard deviation of the random variable x.

(b) An investigator collects 30 observations on these changes and finds an average value of 0.2 and a sample standard deviation of 1.5. Find a 90% confidence interval for the population mean.

(c) Using the sample statistics from part (b), the same investigator then tests the null hypothesis that the population mean is zero against the alternative that it is greater than 0. Report a range for the P-value from this test.

8. Suppose that sample average income in India is \$3000 and sample average income in Pakistan is \$2950. These averages are found from surveys of 25 households in each country. The sample standard deviation of income is 100 in each country. Suppose that income is normally distributed.

(a) Find a 95% confidence interval for the difference between the two population average incomes.

(b) Test the hypothesis that the population average income is the same in the two countries against the alternative that average income is greater in India, at the 10% significance level.

9. Suppose that a labour economist runs a linear regression of wages, y, as the response variable, on years of experience, x, as the explanatory variable, using a sample of thousands of workers. She finds a fitted, least-squares regression line:

$$\hat{y} = 20 + 0.6 \cdot x$$

The slope coefficient has a t-statistic of 2.5 and the R^2 statistic is 0.8.

(a) Does the evidence support a statistical relationship between years of experience and wages?

(b) Based on this evidence, what wage would you predict for a worker with 10 years of experience?

(c) Comment on whether there is a likely candidate as a lurking variable in this statistical relationship.

Economics 250
Winter Term 2013
Final Exam Answer Guide

1. (a) The mean is 3. The variance (being careful to divide by 19 not 20) is 3.3684 so the standard deviation is 1.835. The coefficient of variation is 61.177%.

(b) The first and third quartiles are 1 and 5 so the IQR is 4.

(c) Multiplying by a constant multiplies the mean and standard deviation by the same constant, so those values in euros are 5.037 and 3.081.

2. (a) For employment status: E: 0.58; U: 0.06; N: 0.36 is the marginal distribution. For immigration status: Native: 0.8; Immigrant: 0.2 is the marginal distribution.

(b) For Native: E: 0.6; U: 0.05; N: 0.35. For Immigrants: E:0.5; U: 0.10; N: 0.4. Yes, there is a relationship. For Brazil, immigrants have a lower employment ratio, a higher unemployment rate, and a lower rate of participation in the labour force. (If you like you can show the two variables are not independent using the product rule.)

3. (a) For w the mean is 68 and the variance is 432. For height the mean is 66 and the variance is 108.

(b) The difference has mean 2. The variance is:

$$\sigma_y^2 = 432 + 108 + 2(1)(-1)0.8\sqrt{432}\sqrt{108} = 540 - 345.52 = 194.4.$$

(Notice that the variance of a difference is NOT the difference between the two variances. Also notice that you need the covariance, not just the correlation. remember to refer to the formula sheet if you do not recall these formulas.)

4. (a) The probability of a positive test is $0.009 + 0.0792 = 0.0882$ or 8.82%.

(b) The probability of cancer conditional on a positive test is $0.009/0.0882 = 0.1020$ or 10.2%.

5. (a) From Table C the included probabilities are 0.2150 + 0.2508+ 0.2007 = 0.6665 so the probability of being in the given range is 66.65%. (Remember that with this small sample size you cannot find the probability using the normal approximation.)

(b) Now $\hat{p} \sim N(0.4, 0.0816)$. Standardizing the upper end point gives:

$$z = 1.2255.$$

In Table A that leaves 0.8898 below that value, by averaging the two values in the table. That means there is 11.025% above that value so 22.05% in two tails or 77.95% in the given range.

(c) The confidence interval is:

$$0.111 \pm 1.645(0.0524) = 0.111 \pm 0.0861 = (0.0249, 0.1972).$$

(d) With no success the width of the confidence interval would be 0. So in that case we can use the 'plus 4' or Wilson version: which would give $\tilde{p} = 2/40 = 0.05$ and use that to form a confidence interval.

6. (a) The standard deviation is 0.012649, so standardizing gives:

$$z = \frac{0.22 - 0.20}{0.012649} = 1.58.$$

From Table A that means there is probability 0.9429 below that point or 5.71% above that value, so that is the significance level for this one-tailed test.

(b) Now if $p = 0.23$ then the standard deviation is 0.0133. So the dividing line in the distribution under the alternative implies:

$$z = \frac{0.22 - 0.23}{0.0133} = -0.75.$$

(Notice that the standard deviation in the denominator has changed too.) From Table A that means 0.2266 below that point. So power is 0.7734 or 77.34%.

7. (a) The mean is 0. The variance is 2 so the standard deviation is $\sqrt{2} = 1.414$.

(b) The confidence interval is:

$$0.2 \pm 1.699 \frac{1.5}{\sqrt{30}} = 0.2 \pm 0.465 = (-0.265, 0.665),$$

using the appropriate t-statistic with $df = 29$.

(c) Our test statistic is:

$$t = \frac{0.2 - 0}{0.274} = 0.73.$$

For a one-tailed test with $df = 29$ Table D shows the P-value is between 0.20 and 0.25. (Resist the temptation to say whether you reject the null or not. That is not how you report results when you quote a P-value.)

8. (a) These are independent samples not matched ones, so see the formula sheet for the relevant expressions if you do not remember them. The CI is:

$$50 \pm 2.064\sqrt{400 + 400} = 50 \pm 58.38 = (-8.38, 108.38)$$

(A common error is simply to use $100/\sqrt{25}$ for the standard deviation, but that is obviously not the standard deviation of the difference.)

(b) Our test statistic is:

$$t = \frac{50 - 0}{28.28} = 1.768.$$

With $df = 24$ the critical value for a one-tailed test is 1.318 so we reject the null hypothesis.

9. (a) Yes. The R^2 statistic shows that experience statistically explains 80% of variation in wages. And the t-statistic on the slope is large and thus has a small P-value (so the population value of the slope is very unlikely to be zero).

(b) Using the fitted values, we would predict a wage of 26.

(c) The most likely lurking variable is years of education. That has an effect on wages separate from that of experience. Two workers could have very different experience (or age) yet similar wages if their educations differ. A lurking variable could be included in a multiple regression.

Omitted Question

7. Suppose that in Kingston 7 out of 50 people surveyed report having had the flu this winter while in Ottawa 20 out of 200 people report having had the flu.

(a) Find a 95% confidence interval for the difference between the Kingston and Ottawa flu rates.

(b) Test the null hypothesis that the two flu rates are equal, against the alternative that they are not equal, using a significance level of 5%.

7. (a) The difference is 0.14-0.10 or 0.04. The margin of error is 1.96 (0.053) or 0.104 so the 95% confidence interval is (-0.64, 0.144).

(b) We used the pooled proportion, 0.108, to find the standard deviation, which is 0.0491. So our test statistic is

$$z = \frac{0.04 - 0}{0.0491} = 0.81466.$$

The two-sided critical value is 1.96 so we do not reject the null that the rates are equal.