Economics 250 — Assignment 1 **Due 13 October 2016, by end of class**

Complete this exercise and submit it in class on the due date. You should do the assignment on your own and hand in your own work. Remember to include the departmental cover page on your assignment.

You are welcome to use statistical software for any question (except where "by hand" is explicitly specified).

Short answer questions require no more than a few sentences. Longer answers are no more likely to receive better grades than short, concise answers.

1 Calculating summary statistics [25 points]

Consider the following data set for citizens of Fictionland, measuring income (Y) and consumption (C) in thousands of dollars per year.

Y	39	43	32	42	50	41	58	61
C	22	38	28	48	52	40	46	44

a) For each of the variables Y and C calculate **by hand, showing your work**:

(*i*) the "five-number summary"

(*ii*) the mean

(*iii*) the variance and standard deviation

b) The covariance of the data sets Y and C given above is 63.929. Using **only** this covariance value and values you reported in part *a*), calculate the correlation between Y and C. In general terms (i.e. without using statistics terminology) what does the sign of this correlation tell you about the relationship between consumption and income?

c) Suppose that the values of Y are missing a government-provided income supplement of 2: that is, you are really interested in W, which is just like Y but with 2 added to each observation (i.e. $41, 45, 34, \ldots$). Use the rules of summation to find the mean of W **without** needing to add together all the new w_i values.

d) Suppose that the variable you are really interested in is savings: $S = Y - C$, that is, the data set $s_1 = y_1 - c_1$, $s_2 = y_2 - c_2$, and so on. Using only values you reported in part *a*), but without calculating each s_i value, calculate the mean value of savings, $\bar{s}$.

e) You discover that you made an error when inputting your data, and that the income value "61" in the data set should actually be "61000" (i.e. your sample includes a multimillionaire). Would any of the five-number summary values change? Would the mean change? If so, calculate the new values.

2 Alternative formulas [10 points]

We learned in class, from the textbook, and can see from our formula sheet that the formulas for sample variance and covariance are:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$
$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)$$

Your friend, who took STAT 263 instead of ECON 250, disagrees, and says that the correct formulas are actually the following:

$$s_x^2 = \frac{1}{n-1}\left(\left(\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^2\right)$$
$$s_{xy} = \frac{1}{n-1}\left(\left(\sum_{i=1}^{n} x_i y_i\right) - n\overline{x}\,\overline{y}\right)$$

Settle the disagreement by using our rules for summations[1] to show that the two formulas are equivalent.

Hint: start from the first formulas by expanding the squared or multiplied terms inside the summation.
Hint 2: You may find it useful to note that $\overline{x} = \frac{1}{n}\sum x_i$ *can also be rearranged as* $\sum x_i = n\overline{x}$.

Suggestion: If you are feeling particularly adventurous, answer the question for the covariance formula first, and then show that the variance formula is really just a special version of the covariance formula.

[1] From the handout given in class, which is also available from the "Outline" page of the course website (`https://imaginary.ca/econ250/outline`) under "Additional Materials" → "summation".

3 Distributions [20 points]

Consider the following distributions. (Note that, following our textbook's notation, the second parameter of the normal distributions is the standard deviation (σ), not the variance (σ^2), as is sometimes used instead when denoting normal distributions.)

a) $\mathcal{U}(1, 2)$

b) $\mathcal{U}(-2, 3)$

c) $\mathcal{N}(1, 1)$

d) $\mathcal{N}(-1, 3)$

For each of the above distributions, calculate the probability that a random draw from the distribution would give you a value:

i) exactly equal to 1.2.

ii) greater than or equal to 2.

iii) less than 1.5.

iv) between 0.5 and 2.

v) less than 1.26 or greater than 1.9.

4 Scatter plot and correlation [15 points]

The data set "`internetandlife`" that comes with our textbook and is also posted on the course website (under "Data sets" on the "Resources" page) contains observations on the percentage of internet users and life expectancy for 181 countries.

a) Produce a scatterplot of the data. (*Tip: the data set contains two countries with missing data: Channel Islands, and Serbia. Depending on the software you use, it may be necessary to manually delete these two rows.*)

b) Based on the graph, does there appear to be a correlation between life expectancy and internet use? Is the correlation positive or negative?

c) Calculate the correlation coefficient.

d) You show your answers to parts *a*) – *c*) to a friend who cares a great deal about his life expectancy (but is not taking Economics 250). Based on your answers, your friend rushes home and spends the rest of the day on Facebook in the hope of living longer. Assuming that the data is from a trustworthy source, give a reason why your friend's interpretation of the data is potentially incorrect.

5 Sampling [10 points]

With the professor's permission and assistance, you decide to design a study involving 20 of next year's Economics 250 students to better understand how Economics 250 students behave.

a) Consider the following sampling approaches. For each technique, give a reason why your sample would be biased. Since the population is Economics 250 students who understand the importance of a good data set, you can safely assume that any students selected will participate (that is, assume that no one refuses to take part in the study).

  (*i*) You choose the first 20 students to leave the classroom during the first midterm.

  (*ii*) You visit the classroom on the Thursday before Thanksgiving, ask everyone present to put their name in a hat, and draw 20 names at random.

  (*iii*) You come to class and, noticing that the front two rows of seats have exactly 20 students, choose those students for your sample.

  (*iv*) You use the class list to send an e-mail to the entire class asking for volunteers to participate in your study. *Assume that you get exactly 20 responses.*

b) Briefly (i.e. in one or two sentences) describe how you could construct an unbiased sample of 20 Economics 250 students.

6 Probabilities [20 points]

Suppose that you have a pair of six-sided dice with the usual values 1–6 on the 6 sides. Die A is a fair die: each number has the same probability of being rolled. Die B is weighted: it rolls a 1 just 6% of the time; each of 2–5 have the same probability of being rolled; and a 6 is rolled just as often as every other number *combined*.

a) For each die, write down the sample space for a single roll of the die.

b) For each die, find the probability of each possible outcome from a single roll of the die.

c) Suppose that you roll both dice together and sum the two values. Calculate:

  (*i*) the probability that the sum equals 2

  (*ii*) the probability that the sum equals 10

  (*iii*) the probability that the sum equals 11 or 12

  (*iv*) the probability that the sum *does not* equal 11 or 12

  (*v*) the probability that the two dice show the same value

d) Instead of rolling both dice, you roll the weighted die twice. What is the probability that the two rolls have the same face value?