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, $\overline{s}$.

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.

1.1 Answers

a) (*i*) Calculating the 5-number summary values is much easier if you first sort the values. For Y:

Y_{sorted}	32	39	41	42	43	50	58	61

and for C:

C_{sorted}	22	28	38	40	44	46	48	52

The minimum and maximum can be read straight off the table; the quartiles and median each require taking the midpoint (or mean) of two values:

	Minimum	Q_1	Median	Q_3	Maximum
Y	32	$(39+41)/2 = 40$	$(42+43)/2 = 42.5$	$(50+58)/2 = 54$	61
C	22	$(28+38)/2 = 33$	$(40+44)/2 = 42$	$(46+48)/2 = 47$	52

(*ii*) Means:

$$\overline{y} = \frac{1}{8}\sum_{i=1}^{8} y_i = \frac{39+43+32+42+50+41+58+61}{8} = \frac{366}{8} = 45.75$$

$$\overline{c} = \frac{1}{8}\sum_{i=1}^{8} c_i = \frac{22+38+28+48+52+40+46+44}{8} = \frac{318}{8} = 39.75$$

(*iii*) Variances and standard deviations:

$$\begin{aligned}
s_y^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 \\
&= \frac{(39-45.75)^2 + (43-45.75)^2 + (32-45.75)^2 + \dots}{7} = \frac{679.5}{7} \\
&= 97.07 \\
s_y &= \sqrt{s_y^2} = 9.852 \\
s_c^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(c_i - \overline{c})^2 \\
&= \frac{(22-39.75)^2 + (38-39.75)^2 + (28-39.75)^2 + \dots}{7} = \frac{731.5}{7} \\
&= 104.5 \\
s_c &= \sqrt{s_c^2} = 10.223
\end{aligned}$$

b) From the formula (given in "covariance" supplemental notes) we can use the given covariance (s_{yc}) and the standard deviations calculated above to get:

$$r_{yc} = \frac{s_{yc}}{s_y s_c} = \frac{63.929}{9.852 \times 10.223} = 0.6347$$

Since this is a positive value, it tells us that consumption and income move together: that is, individuals with higher income tend to have higher consumption (and vice-versa).

c) We want to use rule 4 from the supplemental “Summation” notes:

$$\begin{aligned}
\overline{w} &= \frac{1}{n}\sum_{i=1}^{n}(2 + y_i) \\
&= \frac{1}{n}\left(2n + \sum_{i=1}^{n} y_i\right) \\
&= \frac{2n}{n} + \frac{1}{n}\sum_{i=1}^{n} y_i \\
&= 2 + \overline{y} \\
&= 47.75
\end{aligned}$$

i.e. adding a constant to every observation adds that same constant to the mean.

d) $\overline{s} = \overline{y} - \overline{c} = 45.75 - 39.75 = 6$

This is just the rules for means in action. To see why this works, start from the definition of the mean of S:

$$\begin{aligned}
\overline{s} &= \frac{1}{n}\sum_{i=1}^{n} s_i \\
&= \frac{1}{n}\sum_{i=1}^{n}(y_i - c_i) \\
&= \frac{1}{n}(\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} c_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n} c_i \\
&= \overline{y} - \overline{c}
\end{aligned}$$

e) Of the five-number summary, only the maximum changes (to 61000): the minimum, quartiles, and median are unaffected.

The mean will change significantly. To calculate the new value we can either add up all the values again, or take a shortcut by realizing that we're changing the $\frac{61}{n}$ term in the summation to $\frac{61000}{n}$, so we can simply use: $\overline{y}_{new} = \overline{y}_{old} - \frac{61}{8} + \frac{61000}{8}$. Thus the mean increases from 45.75 to 7663.125.

Notice that in cases such as this one the median (which isn't affected by outliers at all) provides a much better picture of the data, especially if we're interested in the behaviour of a typical household.

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:

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

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".

2.1 Answer

Let's start with first variance formula, and expand as the hint suggests:

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

Now separate the three terms in the summation into three separate summations ("Rule 3"):

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} 2x_i\overline{x} + \sum_{i=1}^{n}\overline{x}^2\right)$$

Now consider the second and third terms separately.

Since adding up a constant n times is just n times the constant ("Rule 2"), the third summation becomes:

$$\sum_{i=1}^{n}\overline{x}^2 = n\overline{x}^2$$

For the second term, remember that we can take constants out in front of the summation ("Rule 1"), so the second summation becomes:

$$-\sum_{i=1}^{n} 2x_i\overline{x} = -2\overline{x}\sum_{i=1}^{n} x_i$$

and also applying the second hint given in the question, this becomes:

$$= -2n\overline{x}^2$$

So, putting these into the whole expression again, we get:

$$\begin{aligned} s_x^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2n\overline{x}^2 + n\overline{x}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right) \end{aligned}$$

which is the second formula.

The procedure for covariance is almost exactly the same:

$$\begin{aligned}
s_{xy} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)\\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_iy_i-\overline{x}y_i-x_i\overline{y}+\overline{x}\,\overline{y}\right)\\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_iy_i-\overline{x}\sum_{i=1}^{n}y_i-\overline{y}\sum_{i=1}^{n}x_i+n\overline{x}\,\overline{y}\right)\\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_iy_i-\overline{x}(n\overline{y})-\overline{y}(n\overline{x})+n\overline{x}\,\overline{y}\right)\\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_iy_i-2n\overline{x}\,\overline{y}+n\overline{x}\,\overline{y}\right)\\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_iy_i-n\overline{x}\,\overline{y}\right)
\end{aligned}$$

Another thing to note (following the given *Suggestion*, but not actually required for the assignment) is that the formula for variance is really just a special case of the formula for covariance, where $X = Y$. That is, $\text{cov}(X, X)$, the covariance of X with itself, is exactly the same thing as the variance of X. So you can (following the *Suggestion*) alternatively prove that the covariance equations are the same first, then just plug in $x_i = y_i$ and $\overline{x} = \overline{y}$ to go from the last covariance step directly to the last variance step.

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.

3.1 Answers

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

- (*i*) 0 (this is a continuous distribution with 0 probability of drawing any exact value).
- (*ii*) $P(X \geq 2) = 0$. Though the distribution could generate a value of exactly 1, because it is a continuous distribution the probabilty of getting any exact value is 0.
- (*iii*) $P(X < 1.5) = 0.5$.
- (*iv*) $P(0.5 < X < 2) = 1$. **All** the draws will produce a value in this range.
- (*v*) $P(X < 1.26 \text{ or } X > 1.9) = 0.36$. 26% of draws will be less than 1.26, 10% will be above 1.9, and the remaining 64% will be between 0.26 and 0.9.

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

- (*i*) 0 (see part (a)).
- (*ii*) $P(X \geq 2) = \frac{3-2}{5} = 0.2$.
- (*iii*) $P(X < 1.5) = \frac{1.5-(-2)}{5} = 0.7$.
- (*iv*) $P(0.5 < X < 2) = P(X < 2) - P(X < 0.5) = 0.8 - 0.5 = 0.3$.
- (*v*) $P(X < 1.26 \text{ or } X > 1.9) = 1 - P(1.26 < X < 1.9) = 1 - (P(X < 1.9) - P(X < 1.26)) = 1 - (3.9/5 - 3.26/5) = 1 - 0.128 = 0.872$

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

- (*i*) 0 (see part (a)).
- (*ii*) $P(X \geq 2) = P(Z \geq \frac{2-\mu}{/}\sigma) = P(Z \geq \frac{2-1}{1}) = 1 - P(Z < 1) = 1 - .8413 = 0.1587$. We first convert to a Z-score, then use the table of standard normal probabilities.
- (*iii*) $P(X < 1.5) = P(Z < 0.5) = 0.6915$.
- (*iv*) $P(0.5 < X < 2) = P(Z < 1) - P(Z < -0.5) = .8413 - .3085 = .5328$.
- (*v*) $P(X < 1.26 \text{ or } X > 1.9) = P(X < 1.26) + P(X > 1.9) = P(Z < 0.26) + 1 - P(Z < 0.9) = .6026 + 1 - .8159 = 0.7867$

d) $\mathcal{N}(2,2)$

- (*i*) 0 (see part (a)).

(*ii*) $P(X \geq 2) = P(Z \geq \frac{2-(-1)}{3}) = 1 - P(Z < 1) = 0.1587$. Just like the previous question, we convert to a Z-score, then use the table of standard normal probabilities.

(*iii*) $P(X < 1.5) = P(Z < \frac{1.5-(-1)}{3}) = P(z < 0.83) = .7967$

(*iv*) $P(0.5 < X < 2) = P(Z < \frac{2-(-1)}{3}) - P(Z < \frac{0.5-(-1)}{3}) = P(Z < 1) - P(Z < 0.5) = .8413 - .6915 = 0.1498$.

(*v*) $P(X < 1.26 \text{ or } X > 1.9) = 1 - P(1.26 < X < 1.9) = 1 - (P(Z < \frac{1.9-(-1)}{3}) - P(z < \frac{1.26-(-1)}{3})) = 1 - (P(Z < 0.97) - P(z < 0.75)) = 1 - (.8340 - .7734) = 0.9394$

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.

4.1 Answers

a) The scatterplot should look like the textbook's Figure 2.13 (p. 92), but without the line (although including the line is okay: gretl, for example, automatically throws a regression line on scatterplot graphs—but note that it always uses a straight line, while Figure 2.13 is clearly using a more complicated relationship than a straight line) . The scatterplot could, alternatively, be mirrored such that life expectancy is on the horizontal access and internet time on the vertical access.

b) There is a very obvious positive correlation, particularly among countries with life expectancy above about 60.

c) $r = 0.684037$

d) The friend has confused *correlation* with *causation.* It is much more likely that some other factor (e.g. economic development, technology, etc.) causes *both* values to increase together.

A more subtle reason the friend's interpretation is wrong is that the data set measures only the number of internet users, not the intensity of use of each user. Even if there is causation (which seems unlikely), the data says nothing about the *intensity* of internet use: an individual is counted as an internet user whether she spends 5 minutes or 12 hours per day on the internet.

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.

5.1 Answers

a) (*i*) The sample would be biased because we wouldn't expect that the people coming out of the midterm first are representative: rather, we might expect that they are a mix of the students who studied hardest (and thus answered quickly) and those who didn't study at all (and left much of the exam answers blank).

(*ii*) The technique of putting everyone's name in a hat and selecting at random is a good random sampling technique (assuming each name is the same shape and size, that the hat is carefully and sufficiently mixed, etc.). Bias would still be introduced, however because we might expect only the keener students to be on class on the given day; the students who care less about statistics (or their statistics grade) are more likely to skip Thursday's class to give themselves an earlier start to the studying that they will be doing over Thanksgiving weekend.

(*iii*) The most obvious source of bias is that the people sitting in the front row are probably not representative of the class as a whole: they are typically keener (or braver), and so this most likely does not constitute a representative sample.

Another source of bias is the same as that in the previous example: even if we perfectly randomized from everyone in the room, we are still missing people who don't (regularly) come to class, yet those individuals are still part of the population we are trying to study.

(*iv*) Voluntary surveys are rarely unbiased because different people have different tendencies to respond: those who feel strongly about the issues being studied are more likely to respond, as are those who have more free time, but neither is representative of the group as a whole.

b) The easiest way to get an unbiased sample is to ensure that we are equally likely to choose each member of the population, in this case Economics 250 students. In this case, we could use the class list to select 20 students from the list at random (for example, by using a random number generator, or using the random number table in the back of the textbook (Table B, pages T-4 and T-5) to choose 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?

6.1 Answers

a) The sample space is the same for each die: {1, 2, 3, 4, 5, 6}.

b) For die A, the probability of each outcome equals 1/6. That is, $P(A = 1) = P(A = 2) = \ldots = P(A = 6) = \frac{1}{6}$.

For die B, we are directly given $P(B = 1) = 0.06$. We also know $P(B = 2) = P(B = 3) = P(B = 4) = P(B = 5) = x$; where x is the value to be calculated. As for $P(B = 6)$, we know it equals the probability of all other rolls combined, in other words, $P(B = 6) = P(B = 1)+P(B = 2)+\ldots+P(B = 5) = 0.06+4x$. We also know that the probability of getting any roll equals 1, i.e. $P(B = 1)+P(B = 2)+\ldots+P(B = 6) = 1$. Putting in everything we know thus lets us find x:

$$\begin{aligned}1 &= 0.06 + x + x + x + x + (0.06 + 4x)\\ 1 &= 0.12 + 8x\\ x &= \frac{0.88}{8} = 0.11\end{aligned}$$

So, for die B, we have probabilities:

$$\begin{aligned}P(B = 1) &= 0.06\\ P(B = 2) &= P(B = 3) = P(B = 4) = P(B = 5) = 0.11\\ P(B = 6) &= 0.5\end{aligned}$$

You could also skip a couple steps in finding x by realizing that $P(B = 6)$ must equal 0.5 (since it has the same probability of everything else combined), but you would get the same answer.

c) Note: in the following, let C denote the sum of the two dice face values.

(*i*) $P(C = 2) = P(A = 1 \text{ and } B = 1)$: the only way to roll a 2 is for each die to show a 1. Since A and B are independent, this equals $P(A = 1)P(B = 1) = (\frac{1}{6})(0.06) = 0.01$.

(*ii*) There are 3 ways we can roll a 10: roll a 4 on A and a 6 on B; roll 5 on each; or roll 6 on A and 4 on B. Since all of these rolls are disjoint (they cannot occur at the same time) we can simply add them together:

$$\begin{aligned}P(C = 10) &= P(A = 4)P(B = 6) + P(A = 5)P(B = 5) + P(A = 6)P(B = 4)\\ &= (\frac{1}{6})(0.5) + (\frac{1}{6})(0.11) + (\frac{1}{6})(0.11)\\ &= \frac{0.72}{6} = 0.12\end{aligned}$$

(*iii*) There are two ways to roll an 11: roll a 5 on A and a 6 on B, or roll a 6 on A and a 5 on B; and one way to roll a 12: roll a 6 on each die. Again, since these are all disjoint, we can sum them together:

$$\begin{aligned}P(C = 11 \text{ or } C = 12) &= P(A = 5)P(B = 6) + P(A = 6)P(B = 5) + P(A = 6)P(B = 6)\\ &= (\frac{1}{6})(0.5) + (\frac{1}{6})(0.11) + (\frac{1}{6})(0.5)\\ &= \frac{1.11}{6} = 0.185\end{aligned}$$

(*iv*) Not equalling 11 or 12 is just the complement of equalling 11 or 12:

$$P(\text{not } (C = 11 \text{ or } C = 12)) = 1 - P(C = 11 \text{ or } C = 12) = 1 - 0.185 = 0.815$$

(*v*) The easier way to answer this question is to reason as follows: for *any* roll of B, there is exactly a 1-in-6 probability of A coming up to the same value, so the probability is 1/6.

An alternative approach is to calculate all the possibilities:

$$\begin{aligned}P(\mathit{same}) &= P(A = 1 \text{ and } B = 1) + P(A = 2 \text{ and } B = 2) + \ldots + P(A = 6 \text{ and } B = 6)\\ &= \frac{1}{6}(0.06) + \frac{1}{6}(0.11) + \frac{1}{6}(0.11) + \frac{1}{6}(0.11) + \frac{1}{6}(0.11) + \frac{1}{6}(0.5)\\ &= \frac{0.06 + 4(0.11) + 0.5}{6} = \frac{1}{6}\end{aligned}$$

d) Unlike in the answer to (*v*), above, the same reasoning approach won't work here: given some first roll of B, the probability that the second roll is the same is *not* equal for any first roll value: if the first roll was a 6, for example, the probability of getting the same roll a second time is *not* the same as it would be if the roll was a 1.

We thus use a calculation (here B_1 refers to the first roll of B and B_2 refers to the second roll):

$$\begin{aligned}P(\mathit{same}) &= P(B_1 = 1 \text{ and } B_2 = 1) + P(B_1 = 2 \text{ and } B_2 = 2) + \ldots + P(B_1 = 6 \text{ and } B_2 = 6)\\ &= (0.06)(0.06) + 4(0.11)(0.11) + (0.5)(0.5)\\ &= 0.302\end{aligned}$$