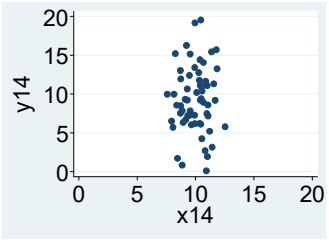
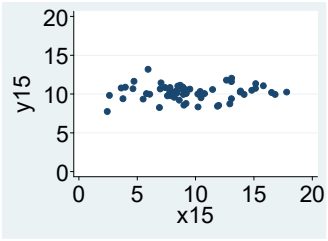
Homework 4: ECO220Y

Required Exercises: Chapter 6: 5, 13, 17, 20, 28, 31, 35

Required Problems:

(1) Answer and EXPLAIN your answer. How do the graphs below *differ*?





(A) the variance of y14 is much larger than the variance of y15
(B) the scatter plot for x15 and y15 shows no relationship whereas x14 and y14 are related
(C) the scatter plot for x14 and y14 shows a steeper relationship than the one between x15 and y15
(D) the scatter plot for x14 and y14 shows a strong relationship than the one between x15 and y15
(E) All of the above

(2) Are each of these observational or experimental data? Explain and ***specifically apply*** course concepts to each.

(a) Data on the interest rate and growth rate of GDP over time in Canada

(b) Obesity rate in mice fed low-carb diet versus regular diet

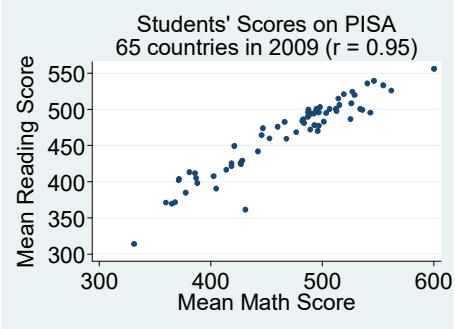
(c) Prices and quantities sold of bath tissue for randomly selected retail outlets

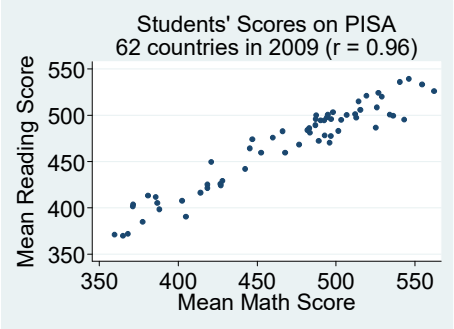
(3) See Exercise 40 (in Chapter 6) in the textbook for background. Consider 2009 data (see StatLink button on page 15 of http://www.oecd.org/pisa/pisaproducts/48852548.pdf.) Here is the variance-covariance matrix for these data.

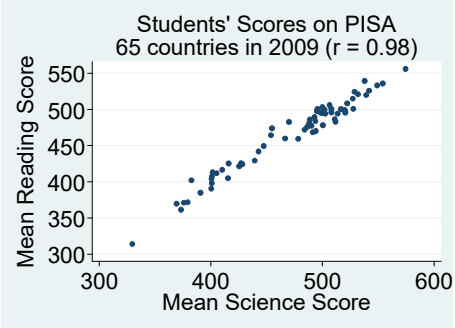
```
. correlate reading math science, covariance;
(obs=65)
             |  reading     math  science
-------------+---------------------------
     reading |   2664.7
        math |  2926.14  3576.25
     science |  2841.29   3255.2  3143.86
```

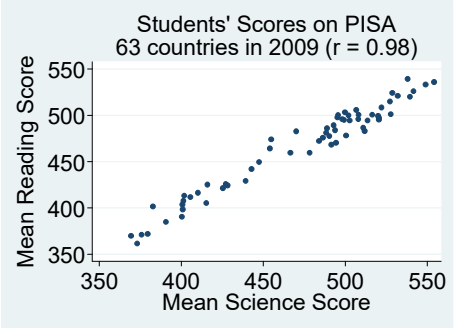
(a) Create a correlation matrix. (Note: You need only the information in the matrix above.)

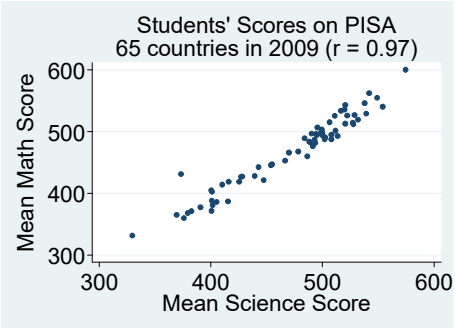
(b) See the following six scatter plots (which continue onto the next page). Are reading and math scores strongly *correlated* or *associated*? Are there any significant concerns about outliers?

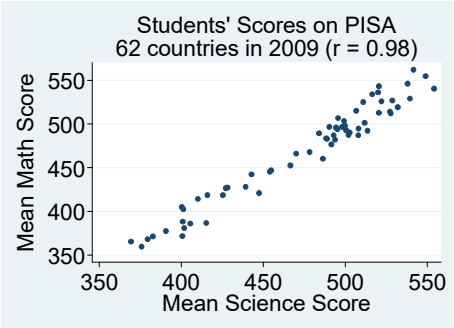












(c) Which kind of data are these? How does that affect the interpretation of the correlations?

(d) Consider the 2012 PISA data (see StatLink button on page 19 of http://www.oecd.org/pisa/keyfindings/pisa-2012-results-volume-I.pdf). Compare and contrast these results with those from 2009 discussed earlier.

```
. correlate reading math science, covariance;
(obs=65)

             |  reading     math  science
-------------+---------------------------
     reading |  2216.65
        math |  2505.97  3075.51
     science |  2338.37  2736.36  2576.39
```

(4) Read SW11 (see Readings page on Quercus). Assess your understanding of this reading with this quiz.

(4.1) What does the phrase "empirical evidence" mean? It refers to evidence based on ____.

- **(A)** many experts' intuitions
- **(B)** observations recorded as data
- **(C)** the outcome of deductive reasoning
- **(D)** the results of simulations derived from theory
- **(E)** rigorous mathematical modeling of a phenomenon

► **Questions (4.2) – (4.4):** Consider the example about applying fertilizer to some randomly selected plots of farmland. Suppose there are 50 plots in the treatment group and 50 plots in the control group.

(4.2) This is an example of what kind of data?

- **(A)** time series data
- **(B)** experimental data
- **(C)** observational data
- **(D)** natural experiment data
- **(E)** longitutudinal (panel) data

(4.3) What happens to the 50 plots in the control group?

(A) all get fertilizer
(B) none get fertilizer
(C) some selected plots get fertilizer
(D) some randomly selected plots get fertilizer
(E) all are subject to careful control such that they each receive exactly the same water, sunlight, weeding, seeds, wind, slope, etc.

(4.4) What is the distinguishing feature of a randomized controlled experiment in the farming example?

(A) that the 100 plots are randomly divided into the two groups
(B) that the 50 plots of land in each group are perfectly identical in every respect
(C) that the plots in the treatment group are carefully matched to plots in the control group
(D) that the randomization process has been controlled to ensure that the plots are comparable
(E) that the researchers have verified that all other variables are held constant across these plots

► **Questions (4.5) – (4.9):** Which could correctly fill in the blank. Data on the unemployment rate, inflation rate, and growth rate, in each province for each of the last twelve months would be an example of ____ data?
(A) correct **(B)** incorrect

(4.5) observational
(4.6) experimental
(4.7) cross-sectional
(4.8) time series
(4.9) longitudinal (panel)

(4.10) By using observational data that shows that in schools with smaller class sizes the learning outcomes are typically better than in other schools with larger classes, why is it difficult to answer "Does reducing class size improve elementary school education?" It is difficult because ____.

(A) there are no data available that quantitatively measure outcomes for learning
(B) there is a lot of variability in outcomes across students: each student is different
(C) class sizes vary little across schools making it hard to separate the signal from the noise
(D) factors like neighborhood wealth vary across schools and affect class sizes and outcomes
(E) all of the above

► **Questions (4.11) – (4.14):** You wonder if the format of a questionnaire affects how students respond to questions about their undergraduate experience. You select a random sample of 30 students and randomly divide them into two groups. One group answers the questionnaire online while the other does so using pen and paper. Which are valid criticisms of your study design? **(A)** valid; **(B)** not valid

(4.11) While the results will not by systematically wrong, there will be a fair bit of sampling noise and this will limit your ability to answer your research question.

(4.12) You have made no attempt to ensure that the two groups are *otherwise identical* and this means that your data should not be used to answer your research question.

(4.13) You have failed to ensure that other factors are *held constant* across the two groups and this will lead to an overestimate of the causal effect of the questionnaire format.

(4.14) You should have conducted a randomized controlled experiment rather than relying on observational data.

(5) Get comfortable using the terms “endogenous,” “exogenous,” and “endogeneity bias.” Go back through the examples in Lecture 4 – Canadian inflation and interest rates, chocolate consumption and Nobel Laureate production, and drug dosage and hours of sleep – and apply these terms appropriately. Answer with several sentences for each case.

(6) Consider the cross-tabulation below of two dummy variables from a survey of 774 respondents in a study like Carlin et al. (2017) from Module A in DACM. The variable male takes a value of 1 if the respondent is male and 0 otherwise. The variable chosedom takes a value of 1 if the respondent chose the dominant credit card and 0 otherwise. What is the coefficient of correlation between these two variables?

chosedom	male 0	1	Total
0	178	215	393
1	183	198	381
Total	361	413	774

(7) See the data and scatter diagram below from the Council of Ontario Universities. (Data retrieved on September 22, 2017 from http://cou.on.ca/numbers/multi-year-data/enrolment/.) It records the number of full time equivalent (FTE) students enrolled in undergraduate (UG) programs across all of Ontario’s universities annually since 2000.

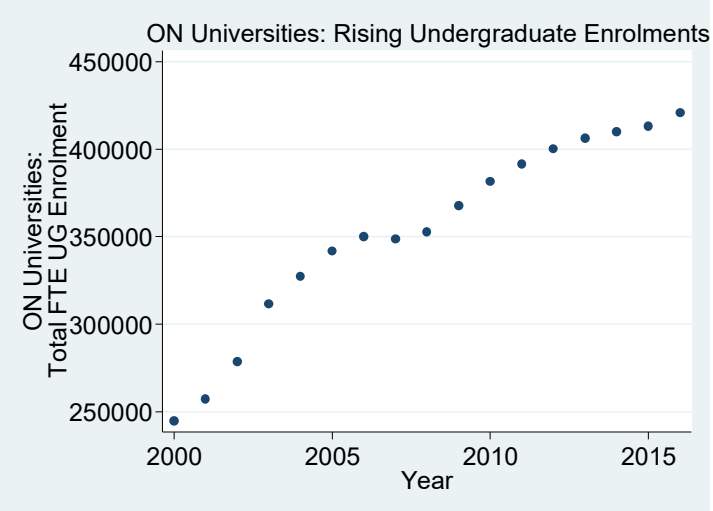
(a) The coefficient of correlation between the variables year and UG_tot_FTEs is 0.9683. In light of the given background information, what does that value of the correlation mean?

(b) How would the coefficient of correlation change if enrolments were measured in 1,000s of FTE students?

(c) How would the coefficient of correlation change if year were recorded as 0, 1, 2, …, 16 instead of 2000, 2001, …, 2016?

(d) How would the coefficient of correlation change if the level of enrolment in the year 2000 were 420,000 instead of 244,945? Also, how would that affect its ability to summarize the strength of the relationship?

year	UG_tot_FTEs
2000	244945
2001	257488
2002	278765
2003	311660
2004	327371
2005	341882
2006	350030
2007	348611
2008	352945
2009	367901
2010	381583
2011	391502
2012	400272
2013	406407
2014	410086
2015	413206
2016	420687



(8) Researchers often compute the covariance and correlation between pairs of dummy variables. This simple context is a great opportunity to build conceptual understanding. *Without actually computing the coefficient of correlation*, use the cross tabulations below to assess whether the correlation will be: exactly zero, very close to zero, positive, or negative. In cases where it is not zero nor close to zero (no relationship), share an assessment of the strength of the relationship.

(a) First situation to assess:

x | y	0	1	Total
0	949	233	1,182
1	2,268	550	2,818
Total	3,217	783	4,000

(b) Second situation to assess:

x | y	0	1	Total
0	28	386	414
1	562	24	586
Total	590	410	1,000

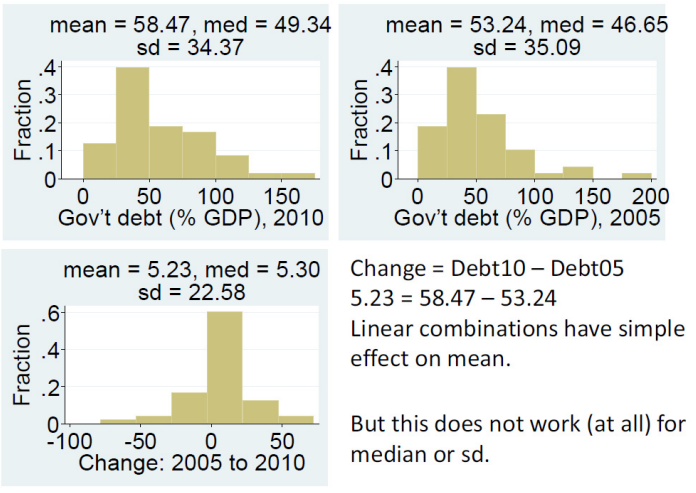
(c) Third situation to assess:

x | y	0	1	Total
0	382	659	1,041
1	152	807	959
Total	534	1,466	2,000

(d) Fourth situation to assess:

x | y	0	1	Total
0	1,638	702	2,340
1	462	198	660
Total	2,100	900	3,000

(9) Recall Slide 25 of Lecture 3 (reproduced to the right), which looked at change in debt as a percent of GDP from 2005 to 2010 for a 48 countries using World Bank data. Also, recall Section 5.6 of the textbook "Adding Measures of Centre and Spread." Expanding on it, for the mean, we can also subtract (not just add). Notice the mean difference is 5.23 and that this is the same as the difference in the means: (58.47 - 53.24). If two variables are uncorrelated then the variance of the sum of the variables is the sum of the variances (Section 5.6). It is also true that if two variables are uncorrelated then the variance of the *difference* is the *sum* of the variances. (Remember that you cannot add standard deviations no matter what the correlation because $\sqrt{x+y} \neq \sqrt{x} + \sqrt{y}$.) Given all of that, why is $(34.37^2 + 35.09^2) = 2413$ not even close to being equal to $(22.58^2) = 510$?



Change = Debt10 – Debt05
5.23 = 58.47 – 53.24
Linear combinations have simple effect on mean.

But this does not work (at all) for median or sd.

World Bank data again, Central gov't debt, n = 48 countries 25