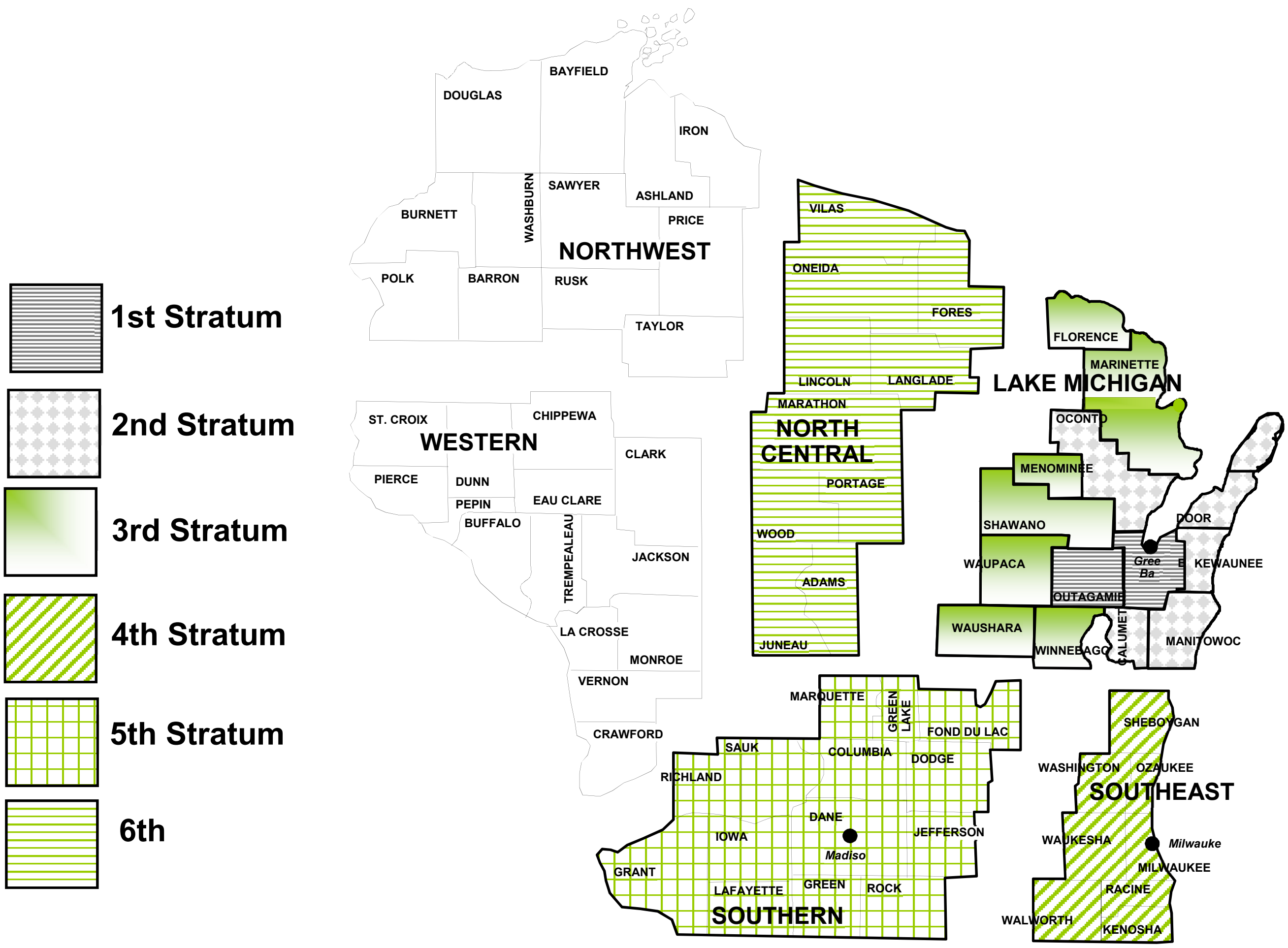
ECO220Y: Homework 1

Required Exercises: Chapter 2: 39, 41; Chapter 3: 1, 3, 5, 9, 15, 22; Chapter 4: 1, 7, 9, 35, 49 (**Note:** The first sets of exercises at the end of each chapter correspond to specific sections. These are best done *as you complete each section* (i.e. before you finish the reading).)

Note: Make sure you have completed the *Quiz and Prerequisite Review for ECO220Y1* (on Quercus).

Required Problems:

(1) This map of the state of Wisconsin (divided by counties) shows the sampling plan for the 1998 Wisconsin Fishing and Outdoor Recreation survey. The anglers (people who fish) in the 1st area had the highest probability of being included in the sample, anglers in the 2nd area had the second highest, …, and anglers in the 6th area had the lowest probability of being included in the sample. Anglers outside these areas had no chance of inclusion. Respondents were contacted by telephone (using a random digit dial (RDD) sampling method).



(a) Is this random sampling, stratified random sampling, cluster, other? Would sampling weights be needed?

(b) Using this sample, can we make inferences about all Wisconsin anglers?

(c) Based on this sampling plan, which group of anglers did the researchers wish to make the most precise inferences about (have the least sampling noise for)?

(d) Discuss some likely non-sampling errors for this specific case.

(2) To consider the relationship between the number of children a woman has and number her mother had you randomly sample 500 women past childbearing age and ask: Q1. How many children did you have? Q2. How many children did your mother have? Answer the following questions.

(a) How many variables are in these data? What kind of data are each variable (interval or nominal)? How many observations? What is the sample size? Are these data cross-sectional, time series or panel?

(b) The answers are recorded in variables named kids_mom and kids_daughter. Does the cross-tabulation show a relationship? If so, is it a positive association or a negative association? Explain.

```
           |                 kids_daughter
  kids_mom |         0          1          2          3 |     Total
-----------+--------------------------------------------+----------
         1 |        73         22          5          1 |       101
         2 |       110         58         30         11 |       209
         3 |         0         97         21          2 |       120
         4 |         0         30         15          4 |        49
         5 |         0          9          6          1 |        16
         6 |         0          2          1          1 |         4
         7 |         0          0          1          0 |         1
-----------+--------------------------------------------+----------
     Total |       183        218         79         20 |       500
```

(c) Comment on the appropriateness of using a cross-tabulation to summarizes these data. If instead of number of children we had data on the age of the mother when she had her first child and the age of daughter when she had her first child: would a cross-tabulation be an appropriate way to summarize such data?

(3) Consider “Who Earns the Minimum Wage?” by Angella MacEwen, May 25th, 2016, on “The Progressive Economics Forum” (http://www.progressive-economics.ca/2016/05/25/who-earns-minimum-wage/, retrieved June 2, 2016):

UPDATE: All numbers exclude self-employed workers. The Labour Force Survey doesn’t provide wage data for self-employed workers, and self-employed workers aren’t subject to minimum wage laws. “Proportion of workers” is more accurately “Proportion of employees”. The number of employees per province can be found in CANSIM Table 282-0012.

How many people even earn minimum wage? Well if you take everyone earning less than the primary minimum wage, there were 1,253,000 workers earning minimum wage or less in 2015 (many provinces have exceptions or lower wages for students or alcohol servers).

	Minimum wage (2015)	Employees earning minimum wage or less (000’s)	Proportion of workers
BC	$10.45	98.2	5.2%
AB	$11.20	100.1	5.2%
SK	$10.50	22.2	4.7%
MB	$11.00	50.4	9.1%
ON	$11.25	675.5	11.6%
QC	$10.55	232.6	6.6%
NB	$10.30	18.5	6.0%
NS	$10.60	30.9	7.9%
PEI	$10.75	6.6	10.6%
NFLD	$10.50	18.0	8.4%

Source: Labour Force Survey microdata 2015, Government of Canada Minimum wage database

But if we’re talking about increasing the minimum wage, workers who earn just above that get a raise too. So how many workers benefit directly from a $15 minimum wage? Well, in 2015 about 25% of all workers in Canada made $15 / hr or less. That’s more than 4 million workers. This varies significantly by province, from 18% of all employees in Alberta, to 38% of all employees in PEI.

	Employees earning $15/hr or less (000's)	Proportion of workers
BC	482.5	25.5%
AB	350.5	18.2%
SK	106.6	22.7%
MB	162.8	29.6%
ON	1,670.1	28.6%
QC	971.4	27.5%
NB	111.5	36.0%
NS	130.6	33.5%
PEI	23.9	38.4%
NFLD	69.6	32.6%

Source: Labour Force Survey microdata, 2015

Finally, women are disproportionately represented in these low wage jobs – fully one-third of women earn less than $15 / hr, compared to only 22% of men.

	Employees earning less than $15 (000's)		Low wage employees as % of all employees	
	Men	Women	Men	Women
15-24	776.5	892.0	65%	74%
25-54	681.1	1,112.9	13%	22%
55+	243.1	373.9	18%	29%
Total	1,700.7	2,378.8	22%	32%

Source: Labour Force Survey microdata, 2015

(a) Consider the last table. We can construct a cross-tabulation that breaks down all employees earning less than $15 per hour by age and sex. Recalling that the numbers originally reported were in 1,000's, consider the shell below for the cross-tabulation results (in 1's, not 1,000's) and fill in the blanks.

```
           |          sex
       age |    Female        Male |     Total
-----------+----------------------+----------
     15-24 | _________            |
     25-54 |                      | _________
       55+ |                      |
-----------+----------------------+----------
     Total |                      | _________
```

(b) Reconsider Part **(a)**. What if instead of starting with all employees earning less than $15 per hour, we started with all employees. Recalling that the numbers originally reported were in 1,000's, consider the shell below for the cross-tabulation results (in 1's, not 1000's) and fill in the blanks.

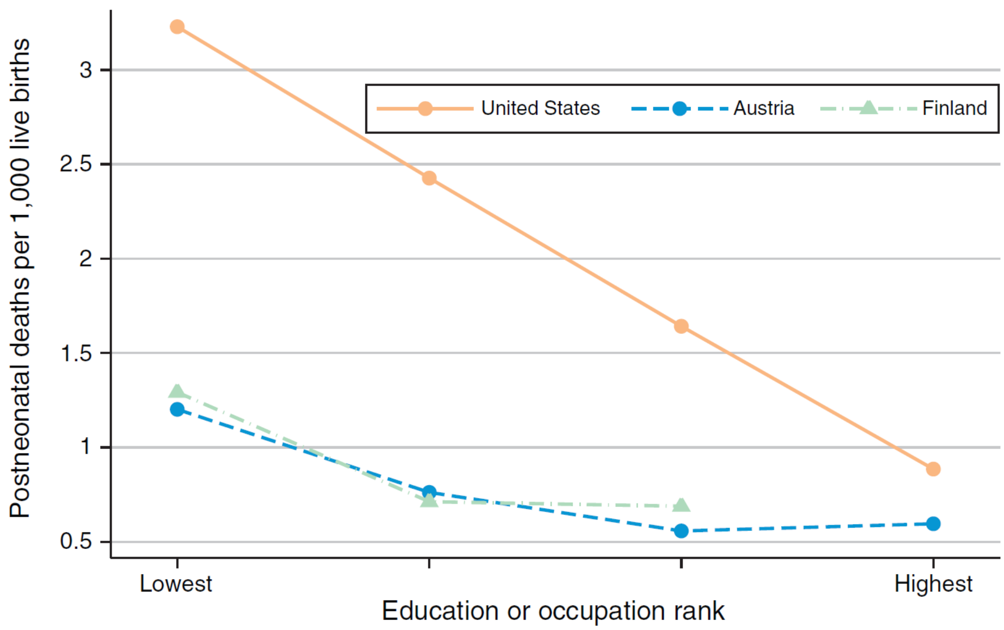
```
           |          sex
       age |    Female        Male |     Total
-----------+----------------------+----------
     15-24 |                      |
     25-54 |                      |
       55+ |                      |
-----------+----------------------+----------
     Total | _________  _________ | _________
```

(4) One important goal of ECO220Y1Y is to prepare you to critically read and understand the findings in current research, including research published in peer-reviewed academic journals. Consider a 2016 academic article "Why Is Infant Mortality Higher in the United States than in Europe?" by Alice Chen, Emily Oster, and Heidi Williams in the journal

American Economic Journal: Economic Policy (DOI: 10.1257/pol.20140224). Academic research in economics often appears in the popular press. For example, the Chen, Oster and Williams (2016) paper is discussed in a June 6, 2016 article in the New York Times "The U.S. Is Failing in Infant Mortality, Starting at One Month Old" (http://www.nytimes.com/2016/06/07/upshot/the-us-is-failing-in-infant-mortality-starting-at-one-month-old.html). Here is the abstract of the academic paper:

> **ABSTRACT** The United States has higher infant mortality than peer countries. In this paper, we combine microdata from the United States with similar data from four European countries to investigate this US infant mortality disadvantage. The US disadvantage persists after adjusting for potential differential reporting of births near the threshold of viability. While the importance of birth weight varies across comparison countries, relative to all comparison countries the United States has similar neonatal (<1 month) mortality but higher postneonatal (1–12 months) mortality. We document similar patterns across census divisions within the United States. The postneonatal mortality disadvantage is driven by poor birth outcomes among lower socioeconomic status individuals.

Consider Figure 6A (below) entitled "Gradient in Postneonatal Death Rates by Socioeconomic Status and Location: Cross-Country." As a general rule for interpreting figures, make sure to carefully read the axes *and* the notes below the table.



Notes: This figure shows the gradient in postneonatal death rates by socioeconomic status and location. The sample for all countries cover 2000 through 2005. The sample is limited to singleton births at ≥ 22 weeks of gestation and ≥ 500 grams with no missing covariates. Because the death rates are postneonatal, the sample also excludes infants who died before one month of age.

(a) First (before comparing countries), explain what the four dots for the United States mean. Make sure to indicate the units of measurement and be specific.

(b) Compare and contrast the results for the three countries. (Note: Compare and contrast means to discuss any important similarities and differences.) You should make at least three distinct points. Make sure to be context specific and to *interpret* the results in plain English. (In other words, your *interpretation* should not be talking about dots and lines, but rather the meaning and patterns in plain English: what messages should a person take away from this presentation of statistics?)

(5) People entering our course often struggle with conditional statements. However, being fluent in conditional statements is an important course requirement that comes up repeatedly. Fortunately, the textbook addresses this topic right away in the early chapters: pages 62 through 68 in this week's readings. Let's practice the kind of questions that often appear on tests and exams: questions that ask you to apply course concepts to real research. Figure 2 (below) is from page 938 of an article titled "Transparency, Reproducibility, and the Credibility of Economics Research" in a 2018 issue of the *Journal of Economic Literature* (Christensen and Miguel (2018); DOI: 10.1257/jel.20171350). Aside from providing you with important practice of current course concepts, this article is particularly relevant to ECO220Y1Y because the Data Analysis Course Module (DACM) relies on important academic journals in economics, like the *American Economic Review* (AER), making replication data available so that we can use those data in DACM.

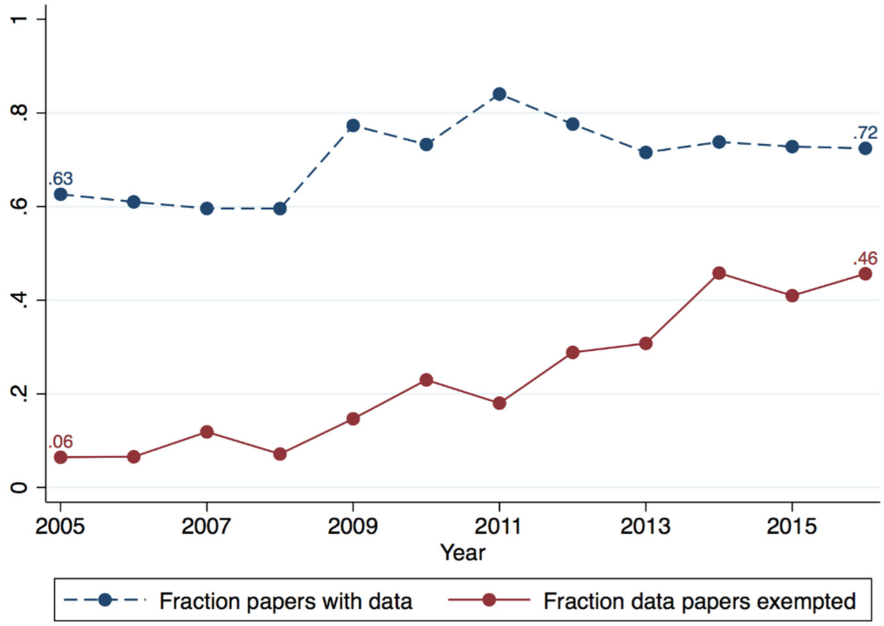


Figure 2. *AER* Papers with Data Exempt from the Data-Sharing Requirement

Note: Figure shows annual data on the fraction of *American Economic Review* papers that use data, and the fraction of those data-using papers that were exempted from the data-sharing policy.

Source: Data is taken from the Annual Report of the Editors, which appears annually in the *Papers and Proceedings* issue of the *AER*. Figure available in public domain: http://dx.doi.org/10.7910/DVN/FUO7FC.

You may be wondering what an exception means. Sometimes papers that use data are published without also publishing the data needed to verify the findings in the paper. The journal's rules say those replication data should, in general, be published. However, editors of journals can waive those rules: in other words, give a paper using data and exception from the data-sharing rule. That is, publish a paper using data without publishing the data needed to verify the results are correct. The most common reason for an exemption is that the data are proprietary: e.g. Amazon allows a researcher to use their private data files on the condition that those data be kept confidential.

(a) If in 2018 the AER (*American Economic Review*) published 110 papers, where 77 are papers that use data and 47 are exempted from the data-sharing policy, then what are the values that would appear in the graph above (if the figure continued past 2016 to 2018)?

(b) If in 2016 the AER (*American Economic Review*) published 100 papers, how many of those papers were exempted from the data-sharing policy?

(6) Recall Simpson's Paradox and composition effects. Bickel et al. (1975) "Sex Bias in Graduate Admissions: Data from Berkeley; *Measuring bias is harder than is usually assumed, and the evidence is sometimes contrary to expectation*" document an example from admissions at the University of California, Berkeley (DOI: 10.1126/science.187.4175.398). Of the 8,442 male applicants, 44.28% (3,738) were admitted. In contrast, of the 4,321 female applicants, only 34.53% (1,492) were admitted. However, studying admissions at the departmental level, females were generally *more* successful. That sounds like a paradox. (The explanation is that females tended to disproportionately apply to programs like English, which have low admission rates for everyone, whereas men tended to disproportionately apply to programs like the physical sciences, which have higher admission rates for everyone.) The following parts give (hypothetical) examples and divide applicants into two groups: A and B. These groups could divide applicants by sex or by other categories such as whether or not an applicant benefits from a legacy preference or whether or not an applicant requires financial aid. For each, identify whether or not it is an example of Simson's Paradox and explain why or why not.

(a) Consider the admissions data below. First, fill in the missing values in the 12 empty cells. Note that the notation "$\hat{P}_A$" means the sample proportion of applicants from Group A admitted. In plain English, it is simply the admissions rate. For this first part, two values have already been filled in for you (in *italics*). Referencing the completed table, is this an example of Simpson's Paradox? Explain why or why not.

	Group A			**Group B**		
	Admitted	Applied	$\hat{P}_A$	Admitted	Applied	$\hat{P}_B$
Department 1	30	100	*0.300*	660	2200	*0.300*
Department 2	40	200		300	1500	
Department 3	60	150		1200	3000	
Department 4	35	50		700	1000	
Overall						

(b) Consider the admissions data below. First, fill in the missing values in the 14 empty cells. Referencing the completed table, is this an example of Simpson's Paradox? Explain why or why not.

	Group A			**Group B**		
	Admitted	Applied	$\hat{P}_A$	Admitted	Applied	$\hat{P}_B$
Department 1	40	100		300	1000	
Department 2	60	200		400	2000	
Department 3	75	150		600	1500	
Department 4	40	50		350	500	
Overall						

(c) Consider the admissions data below. First, fill in the missing values in the 14 empty cells. Referencing the completed table, is this an example of Simpson's Paradox? Explain why or why not.

	Group A			**Group B**		
	Admitted	Applied	$\hat{P}_A$	Admitted	Applied	$\hat{P}_B$
Department 1	350	1000		224	560	
Department 2	1050	3000		96	240	
Department 3	700	2000		312	780	
Department 4	1400	4000		120	300	
Overall						

(d) Reviewing your answers to the previous parts, what are the TWO things that need to exist in this *specific case* to create a Simpson's Paradox situation?