

The DACM Handbook for ECO220Y1Y

*For easy navigation, **download** the pdf file.

Version: July 30, 2025

2025/2026: September 2025 - April 2026

Written by Jennifer Murdock^{a,1}

©2018 Jennifer Murdock. All rights reserved.

1 Goals of the Data Analysis Course Module (DACM)

Get ready to dive into real research and data! You’ve heard of learning by doing. It inspires DACM. There is reading too. By retracing the steps of accomplished researchers all the way from data collection to published results (sometimes appearing as glossy figures in the popular press) you will *do* a lot of statistics and econometrics and deal with a lot of real data. This handbook gives you the curriculum and interactive TA sessions (and optional companion videos) guide you. DACM is an *immersive experience*: there is a flood of data and research and chances to apply a host of ECO220Y course concepts. Here are your learning objectives:

1. Apply the methods you learn in ECO220Y “Introduction to Data Analysis and Applied Econometrics” to answer a variety of research questions using real data
2. Become data literate: understand how data are organized, documented, and readied for analysis and identify reputable sources
3. Become familiar with some major databases and journals that publish replication data
4. Replicate tables and figures, presented in published research, using the original data
5. Build confidence in critically reading and assessing empirical research
6. Understand why researchers sometimes complicate analyses (e.g. transforming or adjusting variables, conditioning on variables) by trying it multiple ways, including the simplest approach
7. Analyze subsets of data (e.g. only homes in a particular climate zone)
8. Fluently employ software – **Excel** – to summarize data and draw statistical inferences
9. Learn to code – using a **Stata** do-file – for basic data manipulation and statistical analysis
10. Develop excellent habits with respect to precision, documentation, and error checking
11. Effectively employ software to apply course formulas to large data sets and also to provide more precise answers than can be obtained from statistical tables
12. Correctly, precisely, and concisely *interpret* (in writing) tables, figures, output, and results

^aProfessor, Teaching Stream, Department of Economics, University of Toronto

¹Thomas Russell, now an Assistant Professor of Economics at Carleton University, provided invaluable insights and suggestions both for the original 2017/18 handbook and substantively revised versions for Summer 2018 and 2018/19.

Contents

1	Goals of the Data Analysis Course Module (DACM)	1
2	DACM Overview	3
2.1	DACM Assessments	3
2.2	Working Through Modules	3
2.3	Do I need to know this stuff for tests and exams?	3
3	Get Started with Excel & Find the Data	4
3.1	Updates to Excel and PCs versus Macs	4
3.2	Limitations of Excel: Watch out for missing values!	4
4	Stata	5
5	What is an <i>interpretation</i>?	5
A	Module A: Describing Data	9
A.1	Module A.1: Types of Data & Analyzing Categorical Data	9
A.2	Module A.2: Histograms & Descriptive Statistics	17
A.0.0	Practice questions for Module A	29
B	Module B: Describing Relationships & Asiaphoria	37
B.1	Module B.1: Association, Correlation, Regression & Composition Effects	37
B.2	Module B.2: PWT & Asiaphoria (Part 1 of 2)	46
B.3	Module B.3: PWT & Asiaphoria (Part 2 of 2)	51
B.0.0	Practice questions for Module B	58
C	Module C: Sampling Distributions & Inference (CI est. & HT)	71
C.1	Module C.1: Sampling Distributions and Simulations	71
C.2	Module C.2: Proportions & Confidence Intervals	81
C.3	Module C.3: Comparing Two Groups & Hypothesis Testing	88
C.0.0	Practice questions for Module C	95
D	Module D: Inference about μ & $(\mu_2 - \mu_1)$ & Using Dummies	107
D.1	Module D.1: Inference about a Mean & Regression Refresher	107
D.2	Module D.2: Inference about a Difference in Means	112
D.3	Module D.3: Review & Dummy Variables in Regression Analysis	120
D.0.0	Practice questions for Module D	132
E	Module E: Multiple Regression	141
E.1	Module E.1: The Big Idea of Multiple Regression & Applied Research	141
E.2	Module E.2: More on Multiple Regression, Including Inference	148
E.3	Module E.3: Interaction Terms & Quadratic Terms	160
E.0.0	Practice questions for Module E	168
F	References	178

2 DACM Overview

The Data Analysis Course Module (DACM) uses cases and data to reinforce and illustrate core ECO220Y curriculum. You engage with data and learn by doing. Section 1 on page 1 lists your learning goals and sets expectations. Page 2 lists the five modules (Modules A through E). For all cases and data, Section F gives full citations. Section 3, starting on page 4, explains how to get Excel (Office 365 ProPlus) and the data files.

2.1 DACM Assessments

The syllabus explains how course grades are computed and the role of DACM. Quercus contains all DACM materials, provides details about DACM work, and lists TA supports.

2.2 Working Through Modules

You work through 14 modules spaced over the course: A.1, A.2, B.1, B.2, B.3, C.1, C.2, C.3, D.1, D.2, D.3, E.1, E.2, and E.3. Each of these 14 modules has roughly 45 minutes of optional companion videos. This handbook gives you the complete lesson plan for each companion video. ***Reading and working through the DACM Handbook is required, but the companion videos are optional.*** For easy search and navigation, ***download this handbook.*** You work along on your own laptop/desktop computer. Utilize TA help when you get stuck.

Each module has required background readings, including figures and tables from published research and refreshers of key concepts covered in our textbook. Carefully study these first.

2.3 Do I need to know this stuff for tests and exams?

Yes. DACM is an integral part of ECO220Y, not an appendage. By working through the DACM curriculum you can expect to deepen your understanding of the course material and improve your performance on other graded assessments, including the cumulative final exam.

To help make the relevance of DACM as obvious as possible, other graded assessments may include questions that use DACM case studies. Further, to make sure you are well prepared for those and any questions on fresh case studies pulled from journal articles and other sources and finish ECO220Y with a working understanding of empirical research, we include:

- **Interpretation tips:** These help you properly interpret the results of data analyses.
- **Test/exam examples:** These are old test/exam questions about a DACM case study. They are sometimes *not* appropriate to work on immediately because they test skills you have not yet learned: the same case can relate to both the first and second half of the course.

Test/exam questions often include both calculations and interpretation. While being able to analyze data and do calculations with software (especially when by-hand calculations would be unreasonable) is an important goal of DACM, it is nearly useless if the *meaning* of the results is unknown or hazy. The interpretation tips and the test/exam examples help you tie together the computations and the interpretations. If you are wondering what is required in an *interpretation*, see Section 5 on page 5.

3 Get Started with Excel & Find the Data

This handbook supports *Microsoft Excel* and the *Analysis ToolPak add-in*. As a U of T student you have access to Office 365, which includes Microsoft Excel, for free: see the page: [Office 365 ProPlus](#). Notice the separate links depending on your device (e.g. windows, mac). You must *first uninstall older versions of Office*.

Update Excel after installing it. If you use a PC, automatic updates should be the default. However, if you use a Mac, you need to turn on automatic updates: [Check for Office for Mac updates automatically](#). After installing Microsoft Excel, open Excel and [Load the Analysis ToolPak in Excel](#).

You download the Excel data files discussed in this handbook from Quercus.

To help you use Excel for the specific tasks in DACM:

- **EXCEL TIPS** pepper this handbook.
- The optional companion videos demonstrate how to use Excel, as you follow along in real time.

3.1 Updates to Excel and PCs versus Macs

This handbook uses on a PC installation of Microsoft Excel 2016. Office 365 updates often, which is generally great, but this can make screen shots slightly out-of-date. Also PC and Mac interfaces have some differences. Where possible, we highlights these.

3.2 Limitations of Excel: Watch out for missing values!

Excel is not powerful statistical software, but it is ubiquitous and most people will use Excel beyond our course (even if not for statistical analyses). Even with just Excel we can replicate cutting-edge research and results. However, there are a *few key limitations of Excel* worth highlighting:

- Excel is used interactively as opposed to writing code and running it. This is dangerous. With code you can fix errors and re-run and you have a record of every step from the raw data to the final result. When working interactively, you must do extra work to document your steps.
- Excel is clunky to use for multiple regression (but it is doable) and it cannot do more advanced statistics/econometrics beyond multiple regression (300-level or higher for undergraduates).
- Excel has serious difficulties in handling missing values. More on that next.

When Excel functions encounter missing values in a variable, they go crazy. Some return an error, others return a zero. Neither is reasonable: if the input to a function is a missing value, the output *should* simply be a missing value. Unfortunately, Excel functions cannot return a truly blank cell. The cases when a zero is returned are particularly dangerous: zero is a real number and hence subsequent functions will treat any zeros as zeros (not missing).

There are workarounds but they are not fun. For example, you can select the entire worksheet, sort by the variable you wish to transform, and only apply the function to the non-missing values (i.e. do

not copy and paste the function to rows where the original variable is missing). If you do this, the new variable will have true missing values (blank cells) in the appropriate spots. More generally, this handbook has many tips to guide you in the safe handling of missing values.

Figure 1 illustrates the pitfalls of missing values in Excel. It shows data with six observations, where the unit of observation is a person. There are three original variables: name, salary, and annual_hrs. The variable salary is a missing value (blank cell) for Xiaodong and Marcus. The variable annual_hrs is a missing value for Tema and Marcus. For a new variable for salary in \$1,000s of dollars (salary_1000), Excel records a zero for Xiaodong and Marcus. The mean of Column D would be wildly incorrect, yet would show no error message. This is the most dangerous situation illustrated. The natural log function returns an error. Column F shows what happens when computing salary per hour: Excel responds to missing values in two different ways even within the same function!

	A	B	C	D	E	F
1	name	salary	annual_hrs	salary_1000	ln_salary	salary_hr
2	Roger	45,000	1,885	45	10.7144178	23.872679
3	Tema	70,000		70	11.1562505	#DIV/0!
4	Xiaodong		1,995	0	#NUM!	0
5	Cherry	65,000	2,015	65	11.0821425	32.2580645
6	Marcus			0	#NUM!	#DIV/0!
7	Sheela	155,000	2,460	155	11.9511804	63.0081301

	A	B	C	D	E	F
1	name	salary	annual_hrs	salary_1000	ln_salary	salary_hr
2	Roger	45000	1885	=B2/1000	=LN(B2)	=B2/C2
3	Tema	70000		=B3/1000	=LN(B3)	=B3/C3
4	Xiaodong		1995	=B4/1000	=LN(B4)	=B4/C4
5	Cherry	65000	2015	=B5/1000	=LN(B5)	=B5/C5
6	Marcus			=B6/1000	=LN(B6)	=B6/C6
7	Sheela	155000	2460	=B7/1000	=LN(B7)	=B7/C7

Figure 1: Illustration of issues with missing values. Instead of missing values, Excel functions return: 0, #NUM!, or #DIV/0!. The right side shows the functions behind each cell. The FORMULATEXT() function or the shortcut **Ctrl + `** (PC) or **control + `** (mac) shows the functions behind the numbers in Excel.

4 Stata

After mastering Excel, you will have a chance to learn the basics of coding in Stata. You will see Stata *output* throughout this handbook. In the second half of the Winter term you will learn the basics of *coding* in Stata. We will support you via Quercus resources and interactive TA sessions. In some ways, getting started with coding and new software (Stata) is harder, but in other ways it is much easier than working interactively using Excel. Once you're started, you will see the power and performance possible with coding. Even if these skills are not directly useful to you in your upper-level courses, it is great to build your understanding of how analysts work.

5 What is an *interpretation*?

Interpreting a number, graph, or other result requires clearly explaining what it means. Generally, all interpretations must, at a minimum, meet these requirements:

1. Be context specific. It is *not* an interpretation if you could cut-and-paste your words and reuse them for another case study.
2. Make the big picture clear. For numbers, round them off in the interpretation: too much precision can distract people. For figures, focus on the overall picture and not every little blip and bump. Stepping back, what do the results mean?
3. Specify the units of measurement.

4. When needed, change the units of measurement to make it easier for people to grasp.

Let's illustrate 1, 2, 3, and 4. It is *not* an interpretation to say "water use is 5,020.8." Here's an interpretation: "According to the OECD, in 2015 Canada extracted just over 5 billion cubic meters of fresh water to supply the public utility water system." (Section F gives full citations.) The original number, 5,020.8, is measured in millions of cubic meters. It is important to specify that this is only the water extracted for the public utility water system: far, far more fresh water is extracted directly by users. Even better, for the units, let's change to litres of water, which most people will understand better than cubic meters: 1 cubic meter = 1,000 litres. Hence, replace "5 billion cubic meters" with "5 trillion litres." (This would fill 2 million Olympic-sized swimming pools, where each pool holds 2.5 million litres of water.)

5. Offer relevant color commentary. Is the result particularly large/small, weak/strong, marginal (a close call)/definitive, and/or surprising/unsurprising?

Comparing water use in other countries, or in Canada in other years, allows such assessments.

6. For interpreting a **difference or a change**, there are more requirements:

- (a) Be clear on causality. Can we say what caused the difference or the change? If so, explain. If not, offer a descriptive interpretation that clearly describes the observed difference.
- (b) Interpret the sign (positive or negative). If $C = (A - B)$ is positive then A is C units *larger than* B . Alternatively, B is C units *smaller than* A . If $F = (D - E)$ is negative then D is F units *smaller than* E . Alternatively, E is F units *larger than* D . Of course you have to say what the units are (these are just generic examples). The point is: it is *not* an interpretation to say that the difference between A and B is C . That leaves the reader wondering which is bigger. Also, it is confusing to talk about a negative difference. Interpret the sign by explaining which is higher or lower and by how much.
 - i. After we cover *statistical significance* in the second half of the course, you must also address whether the observed difference or change is statistically different from zero.
- (c) When a difference is between two things each measured as a percent, specify whether you are measuring the difference as a *percent difference* OR a *percentage point difference*. If $C = (A - B)$ and A is a percent, B is a percent, and C is positive, then A is C *percentage points* higher than B . In contrast, A is $100(C/B)$ *percent* higher than B .

Let's illustrate 6a and 6b. It is *not* an interpretation to say "the difference in salaries is -12.3352 between males and females."² Here's an interpretation: "According to the Ontario disclosure data for all public sector employees making at least \$100,000, among faculty members with a professorial or lecturer job title at the University of Waterloo, on average females made \$12,300 less than males in 2016. However, while this salary discrepancy is large, it does not control for differences in rank, discipline, years of service or any other variables related with both salary and sex so we *cannot* say that \$12,300 measures the amount of unfair sex-based discrimination." (The reason that 6c is not relevant here is because salary is not measured as a percent.)

Before attempting to interpret any results, carefully read any provided background and the titles, notes, captions, and labels given with a table or figure. While notes given in small font below tables

²For the source of this example, see the Supplement for Question (1) in [April 2018 Test #5](#) (with [solutions](#)).

or figures may *appear* to be mere footnotes that could be skipped, notes are actually *important*. Often they contain crucial information necessary to make sense of the table or figure. Also, do not overlook the title(s), which, at their best, contain a nice headline summary of what you are about to see. In other words, a well-prepared table or figure should make your job of interpreting the results – meeting all of the above requirements – a bit easier (so long as you take the time to read what is provided).

But, do I need to be so specific in my interpretation – meeting all of the points above – if the reader can see the original figure, table, or output for themselves? Yes. Your job is explain in words what the results mean.³ If you are asked for an interpretation, you *cannot* leave it to your reader to figure that out by looking at the original results. That said, you must also keep your interpretations *concise*. You certainly do not want to clutter up the key meaning with a bunch of technical details and marginally-relevant background information. Notice that the two sample interpretations given above (about water use and Waterloo salaries) pack a lot of relevant information into one or two sentences. Even when you are writing for an expert audience, you are not off the hook. Researchers writing academic journal articles have to interpret the key results shown in their figures and tables for the reader. Generally, you are trying to write interpretations that would be clear to a less-than-expert audience. It is helpful to imagine you are writing for other students in our course who are not really sure what to make of the number/figure/table/result and need your *clear and firm* guidance. *Firm* means that even if your readers have some misconceptions, your interpretation makes sure they think about the results correctly. Your interpretation cannot be ambiguous and leave room for people to misunderstand what the results mean.

To illustrate interpreting figures and 6 above (including 6c), consider “Postdoctoral Fellowships and Career Choice in Science,” which appears in *The NBER Digest* (<https://www2.nber.org/digest/jul18/jul18.pdf>).⁴

The person summarizing the (much longer) original research paper starts by clearly identifying the research question “What works in supporting the pipeline of scientific talent development?” They continue with relevant background “The National Institutes of Health (NIH) has been asking that question for decades, and has funded undergraduate and graduate fellowships, research grants, and other programs designed to train and encourage promising young scientific researchers.”

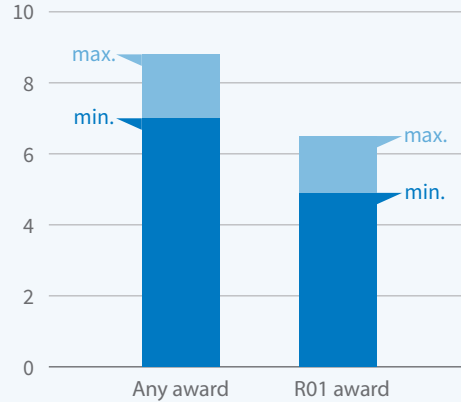
One figure (reproduced on the next page) captures the key results and these are explained (in words). “The study finds that receiving an NRSA fellowship increases the likelihood that a researcher will continue to be involved in NIH-funded research later in his or her career by between 6.3 and 8.2 percentage points. The probability that a researcher will subsequently receive an NIH-funded R01 grant award, an indication of an independent research career, rises by between 4.6 and 6.1 percentage points.”

³Personally, I remember learning this as an undergraduate. I was shocked that I needed to tell people what the tables and figures said. However, explaining in words what the results are and what they mean *is* required.

⁴The *NBER Digest* is a great way for you to keep abreast of cutting-edge research in economics: it gives a short one-page summary, with one clear figure, of a current research working paper. The writing is excellent and accessible for a broad audience (not only PhD economists).

National Institutes of Health Postdoctoral Fellowships and Subsequent Research Funding

Fellowship winners' change in probability of receiving future NIH awards (percentage points)



An R01 award is highly prestigious and indicative of a successful independent research career
Source: Researchers' calculations using data from the National Institutes of Health

Notice that the interpretations imply causality. Heggeness et al. (2018) have done a lot to deal with lurking/unobserved/confounding/omitted variables (Section 6.5 of the textbook) in these observational data, which would include ability (skill, perseverance, and talent). Young scientists with high ability are likely to *both* win a fellowship *and* have a successful scientific career (regardless of any fellowship). The researchers are *not* simply comparing those who receive an NRSA fellowship with those who do not. For example, a simple comparison shows the probability of an R01 award is 0.088 for those who do not receive a NRSA fellowship whereas it is 0.192 for those who do. Descriptively we can say that the probability of winning a R01 award is 10.4 percentage points higher for NRSA fellowship winners ($10.4 = 19.2 - 8.8$). Notice that 10.4 exceeds the maximum causal estimate of 6.1: the number 10.4 is a *biased* estimate of the causal effect. On a more basic level, remember to write *percentage points* because the probability is 118 *percent* higher ($118 = 100 * (0.192 - 0.088)/0.088$): 118 and 10.4 are very different numbers.

You may be hoping for some more examples of proper interpretations, which would earn full marks on a test/exam. Rather than crowd everything here, watch out for the interpretation tips and specific examples of test/exam questions that pepper this handbook. These help you learn how to implement the six general requirements (starting on page 5) in a variety of specific circumstances.

A Module A: Describing Data

A.1 Module A.1: Types of Data & Analyzing Categorical Data

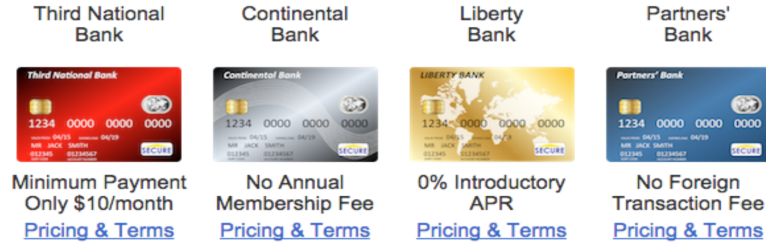
Concepts: Recognize variable types – quantitative (interval) and categorical (nominal) – and data structures – cross-sectional, time series, and panel. Use tabulations and cross tabulations.

Case studies: We replicate parts of an academic journal article “Millennial-Style Learning: Search Intensity, Decision Making, and Information Sharing,” abbreviated Carlin et al. (2017).

Required readings: Chapters 2 and 4. Next is an introduction to Carlin et al. (2017). After that is a review of three common data structures: cross-sectional, times series, and panel.

- In an online experiment, 1,603 respondents from [Amazon Mechanical Turk](#) watch a short video:

Now suppose that you need to apply for a new credit card. You’ve received the four card offers below from four different issuers. Each one has a description from that card’s issuer. Which one would you choose?



- One of the four is the **dominant card**: it is clearly the best choice of the four offered.

Details of the Four Credit Card Offers

(Terms that are worse than the dominant card are highlighted in red)

	Activation fee	APR changes	APR level	Credit limit
Dominant card	\$60	Fixed	13.99%	\$700
High activation fee card	\$110	Fixed	13.99%	\$700
APR can change card	\$60	Can change	13.99%	\$700
High APR & variable limit card	\$60	Fixed	14.99%	Variable

- Two things vary randomly across the 1,603 respondents: the video and taglines.
 1. The **baseline video** is a humorous cartoon about what to watch out for when choosing a credit card. The **implemental video** adds a recap to the baseline video.
 2. The graphic above shows **superfluous taglines**. For example, “No Annual Membership Fee” is superfluous (misleading) because *none of the four cards have an annual membership fee*. Other respondents have the same choice but with **no taglines**.

Table A.1: Summary of Experimental Design: Number of Respondents Receiving Each Treatment

	No Taglines	Superfluous Taglines	Total
Baseline Video	407	394	801
Implemental Video	397	405	802
Total	804	799	1,603

- Each of the 1,603 respondents are *randomly assigned* to one of the **four cells** in Table A.1.

- Three common data structures are cross-sectional, times series, and panel. Whereas variables are about the *columns* in data, data structures are about the *rows*.
 - In **cross-sectional data** each row records the values of the variables for a cross-sectional unit. The unit of observation could be people, countries, employers, etc. For example, for 10 buildings record: building id, square feet, electricity use in 2018, and location. The cross-sectional unit is a building. These data have 10 observations (rows) and 4 variables.
 - In **time series data** each row records the values of the variables for a time period. The unit of observation could be a year, quarter, etc. For example, for Canada in 2018 for each month record: month name, inflation, unemployment, and interest rate. The unit of observation is a month. These data have 12 observations (rows) and 4 variables.
 - In **panel data** there is both cross-sectional and time series variation: it is two dimensional. The unit of observation is both a cross-sectional unit and a time period. For example, for Canada and the U.S. in 2018 for each country in each month record: country name, month name, inflation, unemployment, and interest rate. The unit of observation is a country-month. These data have 24 observations (rows) and 5 variables.
- A common mistake is to confuse the type of data with the number of variables. Any of the three types of data could have many or few variables. If you ask 100 people a survey with many questions you get cross-sectional data with many variables (because it's a long survey). Those data are cross-sectional because the unit of observation is a person and there are 100 observations (rows). If every day in March you record a digital diary about yourself with your weight, hours of sleep, mood, etc. you get time series data with many variables. Those data are time series because the unit of observation is a day and there are 31 observations (rows).

Datasets: For Carlin et al. (2017), [cred_card.xlsx](#), where “cred_card” abbreviates credit card choice. For the survey of various data sources, [assor_ctor_goog_oecd.xlsx](#), where “assor_ctor_goog_oecd” abbreviates assorted data from a variety of sources (City of Toronto, Google Finance, and the OECD).

Interactive module materials for Module A.1:

1. For Carlin et al. (2017), use [cred_card.xlsx](#). **Browse** the worksheets “cred_card” and “readme.”
 - (a) Which kind of data are these? **Verify** that these are cross-sectional data with 1,603 observations and 68 variables. The unit of observation is a person (respondent).
EXCEL TIPS: For large data sets, scrolling is inefficient. Jump to the last cell in a column with a non-missing value with the shortcut **Ctrl** + **↓**. Similarly, use **Ctrl** + **↑**, **Ctrl** + **→**, and **Ctrl** + **←**. Each can be combined with **Shift** to select a range for copying or pasting. For example, **Ctrl** + **Shift** + **↓** selects all cells from the current through to the first missing value. To keep going after a missing value, continue holding down **Ctrl** + **Shift** and press **↓** again. For mac users, use **command** (or **control**) instead of **Ctrl**.
 - (b) Which kind of variables are in these data? **Verify** there is one identifier variable (resp_id) and that many variables are nominal (categorical).

Note: The variables choicetime, choiceclicks, numcards, age, and timedom through view-foreign are clearly interval. Some are in a gray area: starttime, endtime, and the ten

variables for opinions on a 1 to 7 Likert scale (e.g. lik_share_fam). We often treat Likert scale variables as interval. (For example, U of T reports mean course evaluations.)

- (c) What fraction of the 1,603 respondents chose the dominant card? **Verify** that you obtain 0.4885. What fraction of the 1,603 respondents already have a credit card? **Verify** that you obtain 0.7442. In answering, use the fact that these two variables are indicator (aka dummy) variables: one if yes and zero if no.

EXCEL TIPS: Copy the variables chosedom and havecard to a new worksheet. To ensure you get all observations, select the entire column. Use the AVERAGE function to compute the mean of each variable (noting that the mean of a 0/1 variable is the fraction of 1's). For example: =AVERAGE(A2:A1604). Note: When naming worksheets, avoid special characters (such as !, \$, %, quotes, or spaces) and make sure the first character is a letter (a - Z) and *not* a number (those can cause cryptic error messages later on).

- (d) **Replicate** Table A.1 on page 9 (summarizing the experimental design). This is a **cross tabulation**: it tells the frequency of each possible pair of values for two variables.

EXCEL TIPS: Select the entire columns of both variables (video and tagline) and click the PivotTable button under the Insert tab (put output in a new worksheet). In the PivotTable Fields area, drag video to the ROWS area and drag tagline to the COLUMNS area. Drag another copy of either tagline *or* video to the Σ VALUES area and select count. (With a pc, clicking on a field under PivotTable Fields yields a drop-down menu where you can un-check “(blank)” to clean up the table when there are no blanks. If you select only the rows with data, not the entire columns, you can avoid this step.)

Count of video	Superfluous taglines	(blank)	Grand Total
Baseline	407	394	801
Implemental	397	405	802
(blank)			
Grand Total	804	799	1603

Interpretation tips: **Q1.** What does the number 394 mean? Of the 1,603 participants in the study, 394 participants watched the baseline video *and* saw superfluous taglines. **Q2.** What does the number 804 mean? Of the 1,603 participants in the study, 804 participants saw no taglines, regardless of which video they saw.

- (e) Figure 6 shows key results. **Replicate** it. Also, **find** the exact height of each grey bar.

EXCEL TIPS: Copy the variables video, tagline and chosedom to a new worksheet (named “Replicate Fig 6”). Select those variables and insert a PivotChart (and associated PivotTable) in that same worksheet. (Macs by default give both the chart and table.) Drag tagline and video to the AXIS area (dragging tagline to be first and video to be second). Drag chosedom to the Σ VALUES area and using the drop down menu, select Value Field Settings, and choose Average. (For mac users, either right click or click the small “i” to get to Value Field Settings.) (The mean of a 0/1 variable is the *share* of 1's.)

Figure 6. Choice Proportion of the Dominant Card in Each of the Four Experimental Treatments

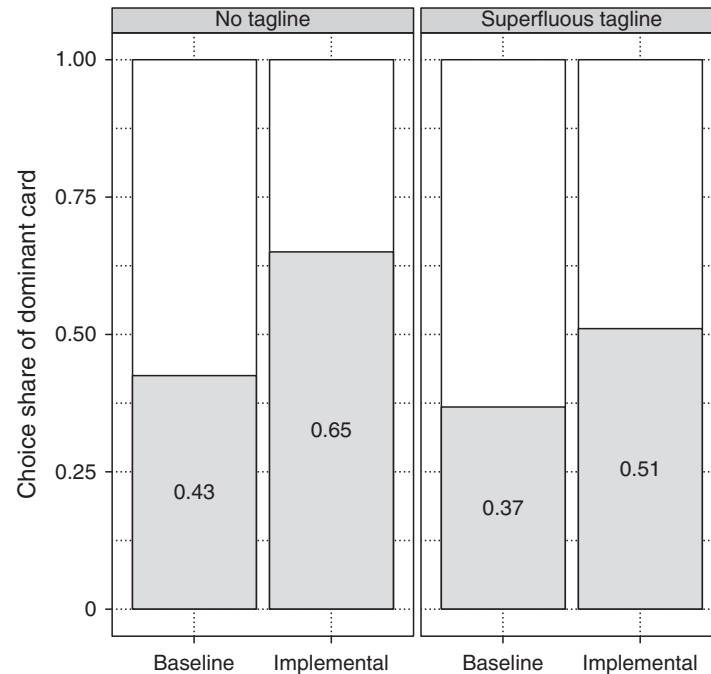
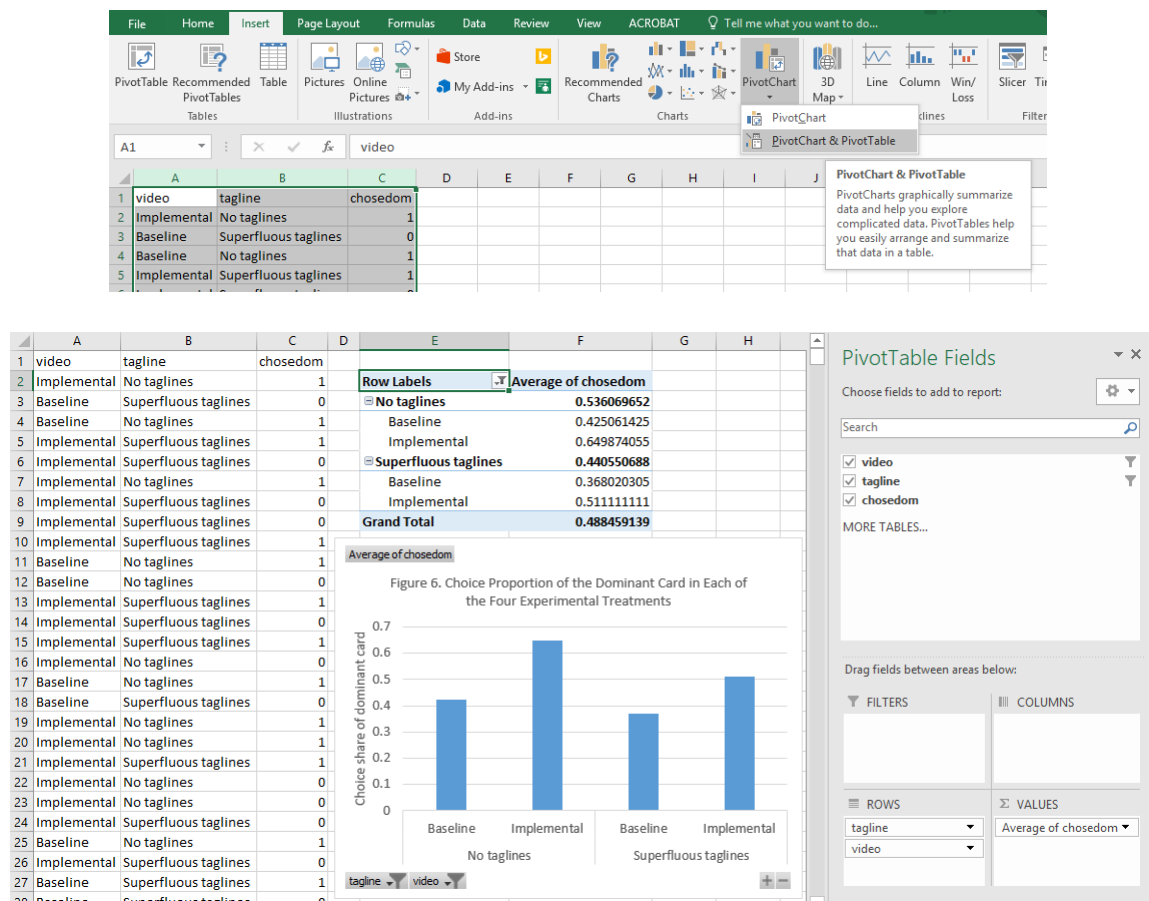
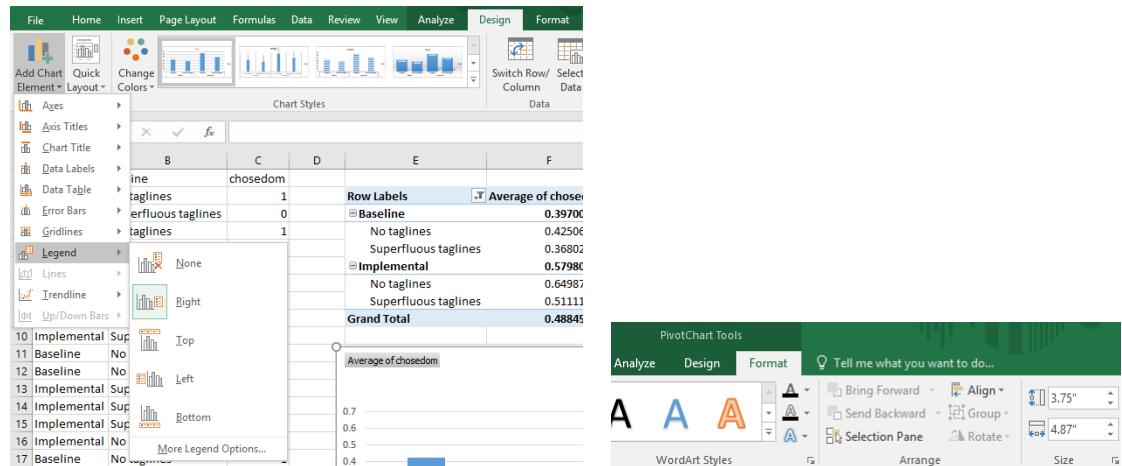


Figure 6: Carlin et al. (2017), p. 13.



To label your figure for best readability, use the Add Chart Element button: remove the

Legend and add Axis Titles. Also, un-check (blank) like in part 1d. Select the chart and click the Format tab to resize. Alternatively, right click (with mac, hold the Control key while clicking) on the chart and select Format Chart Area... from the drop-down menu.



Interpretation tips: Q1. What do the numbers 0.425061425 and 0.649874055 in the PivotTable mean? Among those participants who saw no taglines and watched the baseline video, 42.5% picked the best credit card among the four offered. Among those participants who saw no taglines and watched the longer (implemental) video with a helpful recap, 65.0% picked the best credit card among the four offered. This suggests the longer video helped people (not subjected to misleading advertising) make better choices: they are 22.5 percentage points more likely to make a good financial decision. A 22.5 percentage point increase is huge: a 53 percent increase in those selection of the best card ($= 100 * \frac{65.0 - 42.5}{42.5}$). Further, because these data are experimental – participants are randomly assigned a video to watch – we say the increase is *caused* by the longer video.

- (f) **Review** the first column of results in Table 6 on page 14. **Replicate** the first four rows of the “Chosen (%)” column. Also, **report** the exact percentages.

EXCEL TIPS: Select the variable chosen_terms and insert a PivotTable in a new worksheet. Drag chosen_terms to the ROWS area and drag another copy of chosen_terms to the Σ VALUES area. One way to obtain percentages is to copy-and-paste Column B to Column C and add a Column D to compute the percentages. Usefully, there are two ways to see the formulas in Excel: use shortcut **Ctrl + `** or use the function FORMULATEXT, which returns the formula in a cell. For example, cell E2 below contains =FORMULATEXT(D2). Notice the use of the \$ to anchor to a cell when copying-and-pasting. Here the same total in cell C6 (1,603) is relevant for each percentage. If cell E2 had =100*C2/C6 then copying it to cell E3 would yield =100*C3/C7, which would return an error message #DIV/0! (because C7 is blank). Instead, the \$ says to *literally* copy C6.

	A	B	C	D	E
1	Row Labels	Count of chosen_terms	Count of chosen_terms	Percent	
2	Dominant card	783	783	48.845914	=100*C2/\$C\$6
3	High activation fee card	161	161	10.043668	=100*C3/\$C\$6
4	APR can change card	397	397	24.766064	=100*C4/\$C\$6
5	High APR & variable limit card	262	262	16.344354	=100*C5/\$C\$6
6	Grand Total	1603	1603	100.000000	=100*C6/\$C\$6

Note: Columns A and B in the Excel output above provide a **tabulation** of the variable

recording which of the four credit cards each respondent selected. Hence, if you are asked to tabulate a variable, this is what is meant. In addition to reporting the number of times each value of a variable occurs, a tabulation usually also includes the percent of observations taking each value (i.e. also Column D in the Excel output above).

EXCEL TIPS (IMPORTANT!): Note the percent column is obtained by multiplying the fraction by 100. Use this approach – *especially* when creating or managing a variable in a dataset – and *do NOT* use the option to Format Cells as a Percentage in Excel. This is crucial. Otherwise, the true units of measurement will be hidden from you and this will mess up your calculations and interpretations of any statistics that are not unit-free.

EXCEL TIPS (OPTIONAL): Click on your PivotTable in part 1f. In the drop-down menu for chosen_terms in the Σ VALUES area, select Value Fields Settings, click the Show Values As tab, and select % of Column Total. By default, it formats the proportions (values from 0 to 1) as percents (values from 0 to 100). BE CAREFUL using these values in any subsequent calculations: this formatting hides the real number in the cell.

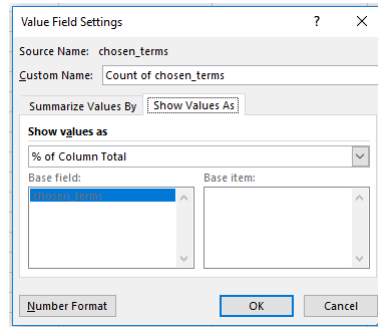


Table 6. Summary Statistics of Choice and Attention (Quartiles of Time Spent Viewing Pricing and Terms and Number of Views of Pricing and Terms)

	Chosen (%)	Time (s) (25th %ile)	Time (s) (50th %ile)	Time (s) (75th %ile)	Views (25th %ile)	Views (50th %ile)	Views (75th %ile)
Dominant card	48.9	1.00	17.00	33.65	1	2	5
High fee card	10.0	1.40	11.90	22.10	1	2	3
Unfixed APR card	24.8	2.20	14.50	28.55	1	2	4
High APR card	16.3	1.50	12.20	23.20	1	2	3.5
First card	26.2	2.70	18.90	26.63	1	2	4
Second card	25.0	2.40	14.30	18.58	1	2	5
Third card	24.8	0.95	11.10	15.94	1	2	4
Fourth card	24.0	0.00	12.00	16.72	0	1	3
0% intro APR	27.3	0.00	10.70	23.30	0	1	3
Minimum payment	17.9	0.00	9.40	22.50	0	1	3
No membership fee	42.4	0.00	12.00	27.40	0	1	3.5
No foreign transaction fee	12.4	0.00	7.40	22.20	0	1	3

Notes. The top four rows show results based on the structure of pricing and terms. The second set of four rows show results based on card position (from left to right). The third set of four rows show results based on the superfluous tagline, among those in the superfluous tagline condition. %ile, percentile.

Figure of Table 6: Carlin et al. (2017), p. 10.

Note: Table 6 has some typos. For the rows of results from “First card” to “Fourth card” they accidentally report the mean time spent instead of the 75th percentile: the correct values are 35.2 (not 26.63), 25.9 (not 18.58), 22.4 (not 15.94), and 22.1 (not 16.72), respectively.

Interpretation tips: **Q1.** In the first column of results in Table 6, what does 10.0 mean? Among the 1,603 participants, 10 percent selected the particular inferior credit card that had a higher activation fee (\$110 versus \$60) among the choice of four cards (one unambiguously the best, and three inferior cards). **Q2.** In the first column of results in Table 6, what does 42.4 mean? (Read the *Notes* below the table.) For the 799 participants shown superfluous taglines, which are misleading ads highlighting a feature that *all* of the cards have (which falsely implies the card with the ad is special), 42.4 percent selected the card that advertised “No membership fee.” Of the four possible misleading ads, “No membership fee” caused the biggest distraction: it increased the fraction of people selecting the card the most. If the misleading ads had no influence on peoples’ choices, we would expect the last four results to be 25, 25, 25, and 25 (rather than 27.3, 17.9, 42.4, and 12.4).

Test/exam examples: Carlin et al. (2017) appears often. You are ready for the Fall 2024, Summer 2019, October 2018, and October 2017 questions.

- Question (1), [October 2024 Test #1](#) (with [solutions](#))
- Question (1), [Summer 2019 Test #1](#) (with [solutions](#))
- Question (1), [October 2018 Test #1](#) (with [solutions](#))
- Questions (2) and (6), [October 2017 Test #1](#) (with [solutions](#))
- Question (1), [April 2018 Final Exam](#) (with [solutions](#))
- Questions (21) - (27), [October 2015 Test #1](#) (with [solutions](#)).

2. **Open** the file [assor_ctor_goog_oecd.xlsx](#), which contains data from a variety of sources (City of Toronto, Google Finance, and the OECD).

(a) **Browse** the worksheet “City of Toronto (Wellbeing).”

- i. Which kind of data are these? **Verify** that these data are cross-sectional with 140 observations and 16 variables. The unit of observation is a neighborhood in the City of Toronto. This is not a random sample: it includes the population of all neighborhoods in the City of Toronto.
- ii. Which kind of variables are in these data? **Verify** that there is one identifier variable (Neighbourhood) and that all of the other variables are interval (quantitative).

(b) **Browse** the worksheet “Google Finance (Apple Stock).”

- i. Which kind of data are these? **Verify** that these data are time series with 3,999 observations and 2 variables. The unit of observation is a day. This is not a random sample: it includes the population of all stock prices during days of trading in that period (July 1, 2005 - May 31, 2017).
- ii. Which kind of variables are in these data? **Verify** that there is one identifier variable (Date) and that the other variable is an interval (quantitative) variable.

Note: These data would still be time series data even if additional daily variables about Apple, such as a daily measure of Apple’s press coverage, were also included.)

(c) **Browse** the worksheet “OECD (Ren, Ene., CO2, GDP, Oil).”

- i. Which kind of data are these? *Verify* that these data are panel (longitudinal) data with 390 observations and 6 variables. The unit of observation is a particular country in a particular year. This is not a random sample: it includes the population of all 26 OECD member nations with available data during that 15-year period (2000 - 2014). Note that $26 \times 15 = 390$.
- ii. Which kind of variables are in these data? *Verify* that there are two identifier variables (COUNTRY and YEAR) and that the other four variables are interval (quantitative) variables.

Note: The three examples above show all kinds of data: cross-sectional, time series, and panel. However, none show a nominal (categorical) variable (which is not also an identifier variable). Recall that you have seen such examples in part 1b using [cred_card.xlsx](#). For example, `highest.ed` is a straight-up example of a nominal (categorical) variable. Sex (also categorical) is recorded as a dummy (0/1 variable) named `male`. Also, all of the experimental variables are nominal (e.g. `video`, `tagline`).

Test/exam examples: For data structures, see:

- Question (1), [May 2023 Test #1](#) (with [solutions](#))
- Question (4), [October 2021 Test #1](#) (with [solutions](#))

A.2 Module A.2: Histograms & Descriptive Statistics

Concepts: Using histograms and descriptive statistics (e.g. mean, median, s.d.) to summarize quantitative (interval) variables.

Case studies: We use data from an academic journal article “A New Era of Pollution Progress in Urban China?” abbreviated Zheng and Kahn (2017). These data, compiled from China’s *Statistic Yearbooks* and *City Statistical Yearbooks*, include variables measuring weather, geography, air pollution (PM10), city-level GDP, and other socioeconomic variables. For currency conversions, a variable from FRED (China / U.S. Foreign Exchange Rate) has been merged on. Also, the World Health Organization (WHO) compiles data on air pollution for cities worldwide (not only China) and includes two common measure of air pollution (PM2.5 and PM10).

Required readings: Chapter 5. Background reading on measuring air pollution:

- To measure air pollution, Zheng and Kahn (2017) use PM10: the concentration of particulate matter (i.e. small particles such as dust) with a diameter of 10 micrometers (μm) or less (μ is short for micro and is *not* related to the population mean μ). Another measure of air pollution is PM2.5 for fine particulate matter with a diameter of 2.5 micrometers or less. The units of measurement of PM10 or PM2.5 are micrograms per cubic meter air ($\mu g/m^3$).

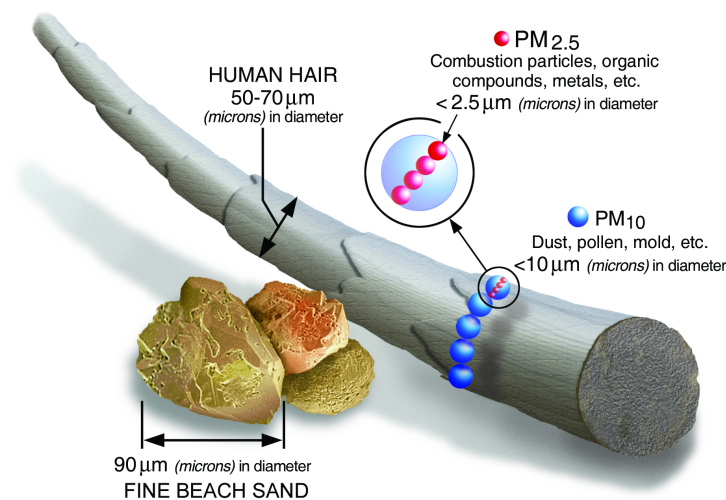
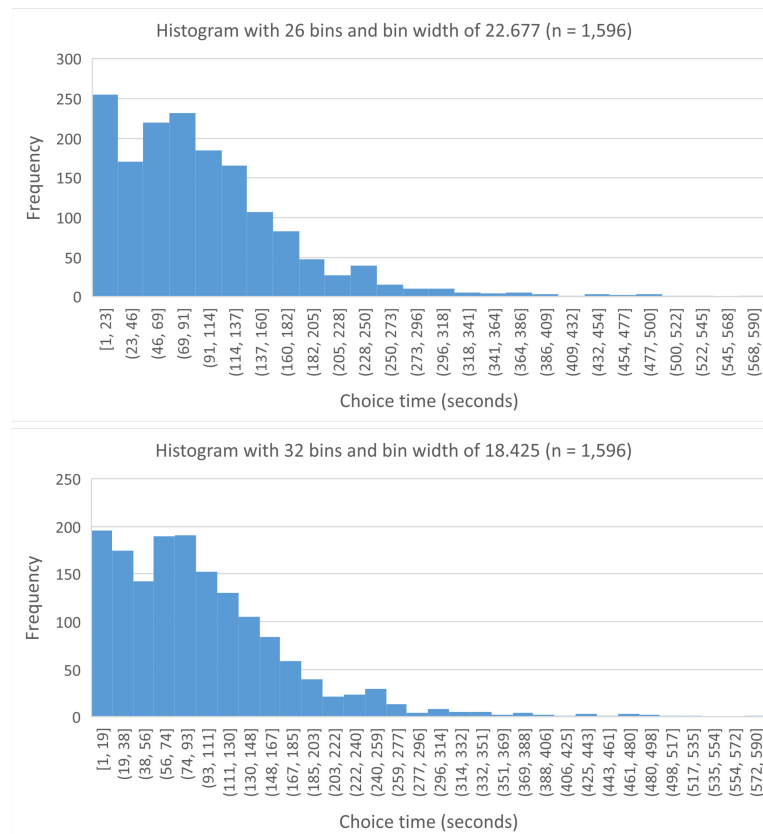


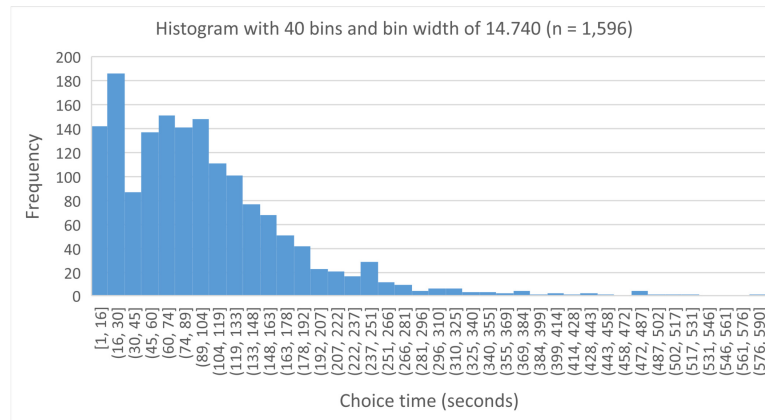
Figure 1: U.S. EPA (online), “Particulate Matter (PM) Basics”, retrieved July 17, 2017.

- To give some sense of these, according to WHO (2016), the annual mean of PM10 in 2012/13 in: Beijing, China is $108 \mu g/m^3$; Toronto, Canada is $14 \mu g/m^3$; Los Angeles, U.S. is $20 \mu g/m^3$; Rome, Italy is $28 \mu g/m^3$; Tokyo, Japan is $28 \mu g/m^3$; Delhi, India is $229 \mu g/m^3$.
- Drawing a histogram is both art and science. While your bins *must* all be the same width, *must* not overlap, and *must* cover the entire range of values (or explain any special treatment of outliers), you *do* have some choices. The goal is to *visually summarize* the distribution. Various software packages make different decisions. Excel 2016 even makes different choices for its two histogram tools. Next are some formulas giving *suggestions* about the bins. You may choose something a bit lower or higher than any of these formulas, making a *subjective* judgment about

the best summary of a particular distribution. All suggest more bins (narrower bins) for bigger sample sizes (i.e. a larger number of observations n). The number of bins is always an integer. Once you choose the number of bins, you have no choice about the width (and vice versa).

- Number of bins = \sqrt{n} . E.g., with 400 observations of a variable, choose ≈ 20 bins.
 - Number of bins = $\frac{10\ln(n)}{\ln(10)}$. E.g., with 4,000 observations of a variable, choose ≈ 36 bins.
 - Or, a combination. E.g., Stata uses = $\text{MIN} \left\{ \sqrt{n}, \frac{10\ln(n)}{\ln(10)} \right\}$. (In practice, this means that you use \sqrt{n} whenever the number of observations is less than 862.)
 - Number of bins = $1 + 3.3\log_{10}(n)$. E.g., with 34 observations, choose ≈ 6 bins.
 - Width of each bin = $\frac{3.5*sd}{\sqrt[3]{n}}$, where sd is the standard deviation. E.g., with 850 observations of a variable with a s.d. of 11.48, choose a bin width of ≈ 4.2 . **Note:** This is for Normal (Bell shaped) distributions. It usually suggests fewer bins and is not ideal for other shapes.
- For interpreting histograms, see Sections 5.1 - 5.5 (textbook). For the shape, use common terms (below). These overall descriptions should *not* depend on the exact choices for the bins.
 - Symmetric or skewed? Symmetric, positively (right) skewed, or negatively (left) skewed
 - Modality? Unimodal, bimodal, or multimodal (three or more major peaks)
 - Famous shape? Uniform (full story in Section 9.9) or Bell (Normal) (Section 9.10)
 - To illustrate, recall [cred.card.xlsx](#) from Module A.1 and the variable choicetime. There is 1 missing value and 6 extreme values over 600 seconds. For the 1,596 (= 1,603 – 1 – 6) remaining observations, the min value is 0.8 and the max value is 590.4. These histograms use suggestions from three formulas above: bin width $\frac{3.5*77.4}{\sqrt[3]{1,596}} \approx 23$, $\frac{10\ln(1,596)}{\ln(10)} \approx 32$ bins, or $\sqrt{1,596} \approx 40$ bins.





Interpretation tips: What do the above three histograms show? Despite differing numbers of bins, all three histograms give the same overall visual summary of how much time respondents spend choosing a credit card. Specifically, they show positive skew: the vast majority of respondents spend a modest amount of time selecting a card – less than 3 minutes – but a small number take quite long (some take more than 9 minutes). They show evidence of bimodality: a group making the choice very quickly (a fraction of a minute) with another group taking roughly a couple of minutes.

Optional: The *Journal of Economic Perspectives* targets a general audience, including undergraduates. For Zheng and Kahn (2017), we cover multiple regression (Table 1) in the last third of our course.

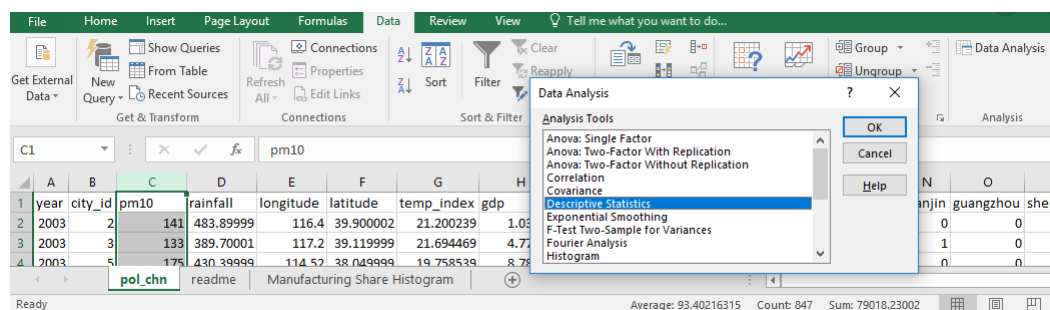
Datasets: For Zheng and Kahn (2017), [pol_chn.xlsx](#), where “pol_chn” abbreviates pollution in urban China; For WHO (2016), [who-aap-database-may2016.xlsx](#) (filename used by the WHO).

Interactive module materials for Module A.2:

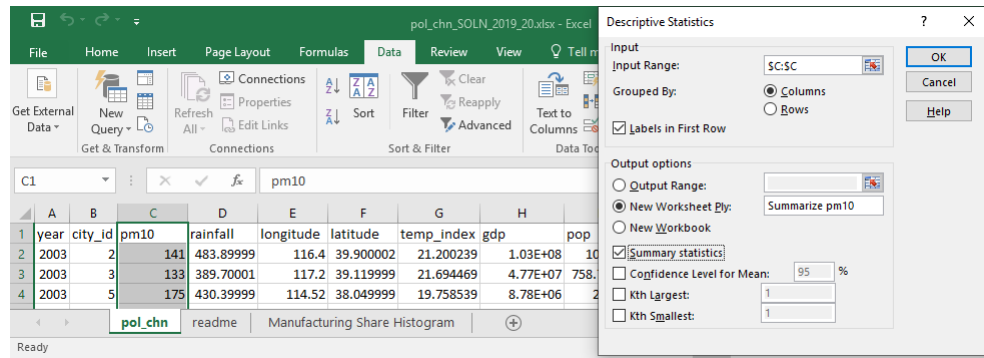
1. Consider Zheng and Kahn (2017) and **open** the file [pol_chn.xlsx](#).

- (a) **Browse** the data and data description in worksheets “pol_chn” and “readme.”
- (b) **Compute** the usual suite of descriptive statistics for the variable pm10.

EXCEL TIPS: Click Data Analysis on the Data tab. (If it does not appear, install the Data Analysis ToolPak add-in: Section 3 on page 4.) Select Descriptive Statistics.



Set the Input Range as the entire column (\$C:\$C). Check the boxes for Labels in First Row and Summary statistics. Output to a new worksheet named “Summarize pm10.” Always include the variable name and any option to include labels: this is the bare minimum to document your work. Similarly, name your worksheets clearly. These good habits allow you to retrace your steps and remember the relevant specifics of your analyses.



EXCEL TIPS: For more than the one largest and smallest value, click the boxes Kth Largest and Kth Smallest. For example, enter 2 for the second largest and smallest values.

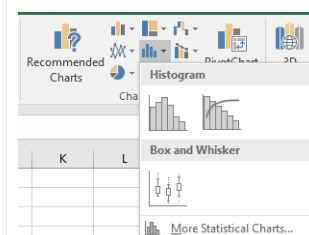
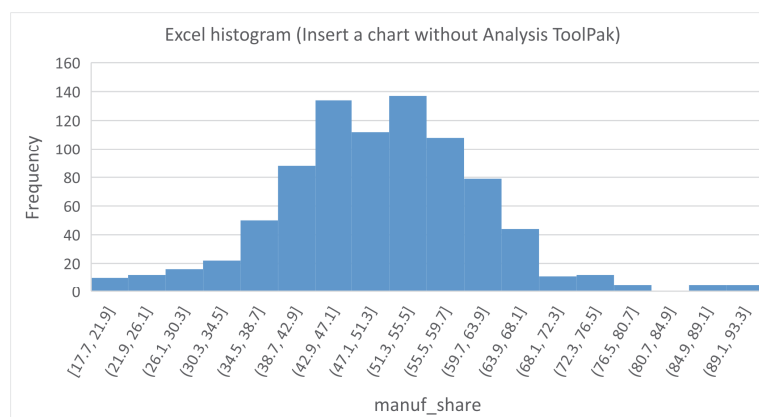
- (c) **Verify** the Excel output below (left) matches yours. **Compare** it with the Stata output (right). While both report common summary statistics, there are differences. Excel gives the standard error (covered later in our course), mode, and total sum (which divided by n is the mean). Stata gives percentiles (50th percentile = median) and the four smallest and largest observations (not just the smallest and largest one). Both measure kurtosis and skewness, which you can ignore. Stata helpfully includes the variable description.

<u>pm10</u>					
Mean	93.40216315	. summarize pm10, detail			
Standard Error	1.109286217	Particulate matter conc. (micrograms/cubic meter air; diameter<=10 micrometers)			
Median	91.1805				
Mode	97				
Standard Deviation	32.26478671				
Sample Variance	1041.016462				
Kurtosis	47.32347917				
Skewness	3.796381164				
Range	521				
Minimum	29				
Maximum	550				
Sum	79018.23002				
Count	846				

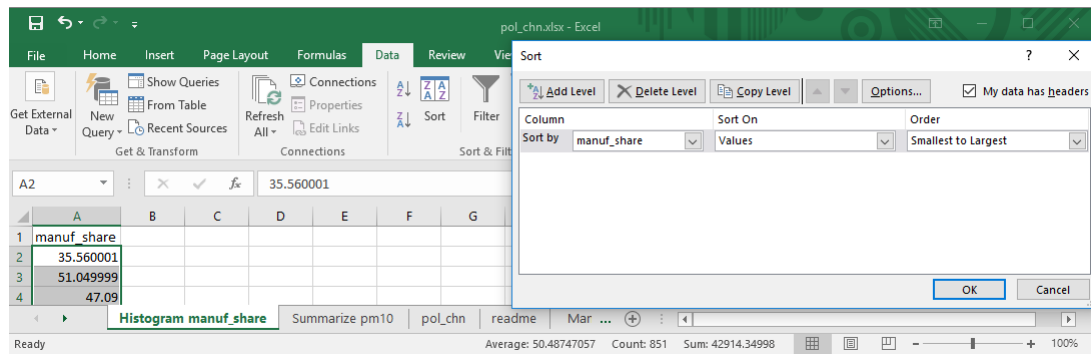
Percentiles		Smallest			
1%	37	29		Obs	846
5%	50	30		Sum of Wgt.	846
10%	59.773	30			
25%	74	33			
50%	91.1805			Mean	93.40216
				Std. Dev.	32.26479
75%	109	192		Variance	1041.016
90%	128	198		Skewness	3.789647
95%	142	239		Kurtosis	50.03717
99%	176	550			

Interpretation tips: What does 93.40216 mean? Air pollution, as measured by the concentration of particulate matter with a diameter of 10 micrometers or less (PM10), across 85 Chinese cities from 2003 through 2012, is $93 \mu\text{g}/\text{m}^3$ on average for the 846 non-missing observations. This is notable air pollution compared to Toronto in 2012/13 with only $14 \mu\text{g}/\text{m}^3$. (See part 1(l)ii on page 26 for how to get the number of cities and years.)

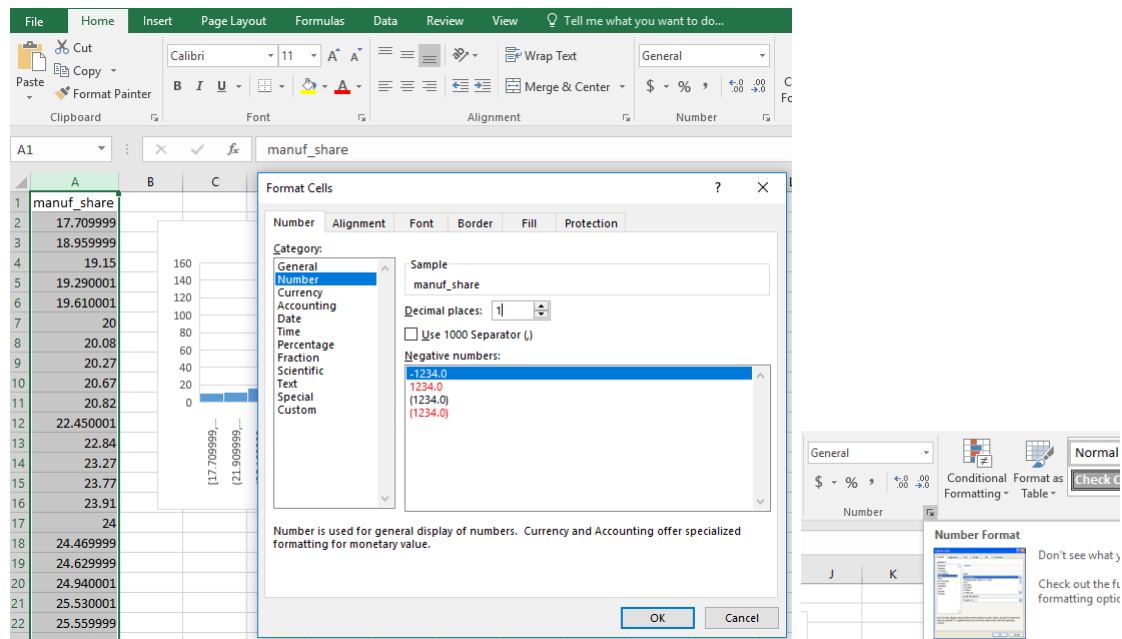
- (d) **Replicate** the histogram below (not worrying about adding the titles).



EXCEL TIPS: In the 2016 version, Excel added histograms to the standard set of charts (without add-ins like the Analysis ToolPak). Create a new worksheet titled “Histogram manu.share.” Select the entire column for the manu.share variable and copy it to the new worksheet. Sort it in ascending order. This helps address any missing values for a variable. (In this case, it has no effect because there are no missing values of the manu.share variable. It is a good habit because missing values are common.) **Be careful with sorts: select ALL variables and ALL rows or you will destroy your data.** It is safer in a worksheet with only one variable (but you still must select all observations).

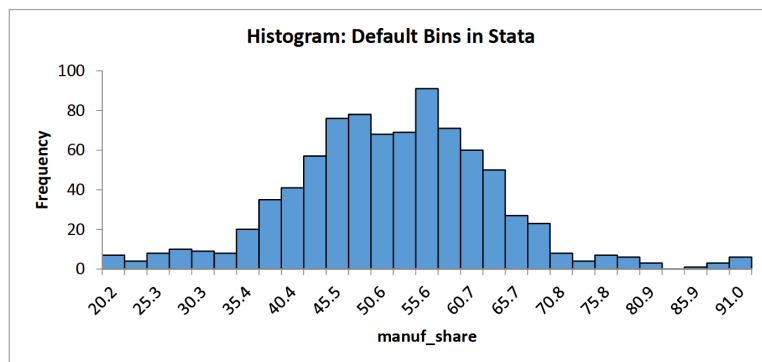


Select all non-missing values, including the variable name, using the keyboard shortcuts on page 10. Under the Insert tab, click the button with the miniature histogram shown on page 20. To reduce the number of decimal places on the x-axis bins (improving readability), select the entire column for manu.share. Under the Home tab, pull up more options for Number (see below right). Alternatively, you can right click on the selected column and select Format Cells from the drop-down menu. Select Number, and select 1 decimal place. You can do this after making the histogram (it automatically updates).



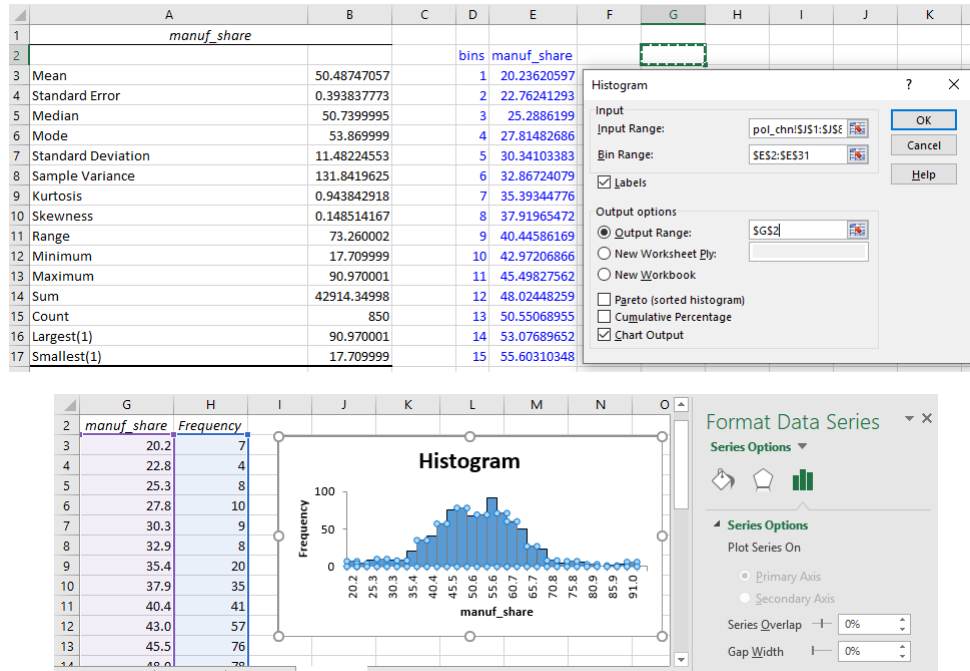
Note: The histogram chart in Excel 2016 chooses width of each bin = $\frac{3.5 \cdot sd}{\sqrt{n}}$ by default. The bin width on page 20 is 4.2: e.g., $51.3 - 47.1 = 4.2$. The sample size is $n = 850$ and the standard deviation of manufacturing share is 11.48. Plugging in, we obtain $4.2 = \frac{3.5 \cdot 11.48}{\sqrt{850}}$.

- (e) Now we explore how our bin choices affect a histogram. **Browse** the worksheet “Manufacturing Share Histogram.” It shows descriptive statistics and a histogram for `manuf_share`.

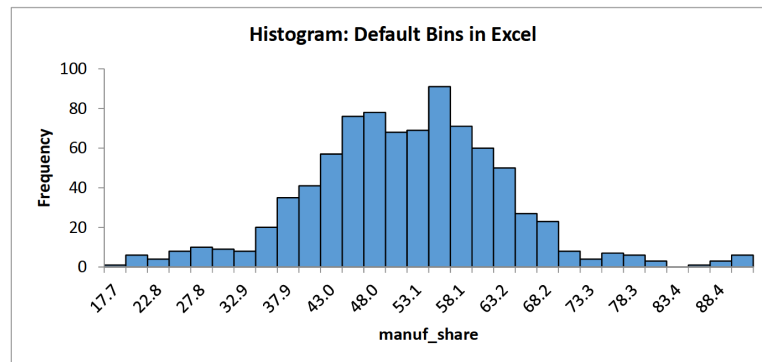


EXCEL TIPS (IMPORTANT!): We now switch the Data Analysis ToolPak histogram tool. With a pc (not a mac) you *can* use the histogram chart as in part 1d. However, the mac version does *not* allow you adjust the histogram (e.g. you cannot change the number of bins). The ToolPak version has flexibility for everyone. However, the histogram chart in part 1d *is* still useful for everyone for a quick and basic histogram.

- (f) Continuing, **read** the bullets below explaining the cells in **blue**:
- Recall from page 18 that Stata uses a different formula to make suggestions about the bins for a histogram. In the worksheet “Manufacturing Share Histogram,” **find** the value 29.15476 produced by the formula: $\text{number of bins} = \text{MIN} \left\{ \sqrt{n}, \frac{10 \ln(n)}{\ln(10)} \right\}$.
EXCEL TIPS: This cell uses: `=MIN(B15^0.5, 10*LN(B15)/LN(10))`, where the cell B15 is the count of observations (n) given with the descriptive statistics. Alternatively, `=MIN(COUNT(pol_chn!J:J)^0.5, 10*LN(COUNT(pol_chn!J:J))/LN(10))`.
 - Next, **review** the rounded suggested number of bins (an integer) and suggested width.
Note: Once you determine the number of bins, you can find the width of each bin by taking the range of your variable (which, recall, is the maximum value minus the minimum value) and dividing it by the number of bins.
EXCEL TIPS: Notice how the worksheet does this in cell B21 with `=B11/B20`.
 - Next, **review** the two columns in **blue** that set up the right endpoint of each bin (bins 1 through 29) using the minimum value in the data and the suggested bin width.
EXCEL TIPS: Notice the use of the \$ to anchor to a cell when copying-and-pasting. For example, `=B$12 + D3*B$21` for the first bin.
- (g) **Replicate** the histogram in part 1e using the bins as defined in Stata.
EXCEL TIPS: Create a copy of the worksheet “Manufacturing Share Histogram.” Delete everything to the right of the two columns in **blue** that set up the right endpoints of each bin (i.e. *keep* the **blue** columns that define the bins). Click the Data Analysis button in the Data tab and select Histogram. Select the Input Range from the original data (`pol_chn!J1:J851`), select the Bin Range from your current worksheet (include the column label), check the Labels box, select `G2` as the Output Range (which makes the histogram appear in your current worksheet), and check the Chart Output box. To fine-tune your histogram, click the chart area to Format Data Series: set Gap Width to zero. Format the numbers in the table produced with the histogram to the first decimal place (see Column G in the screenshot below).

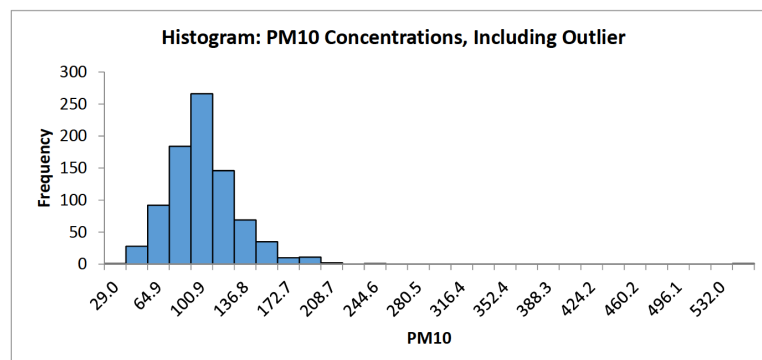


- (h) **Create** a histogram of `manuf_share` using $\sqrt{n} + 1$ bins. (This is the Data Analysis default. Stata uses \sqrt{n} in this case.) **Verify** that your histogram looks similar to this one.



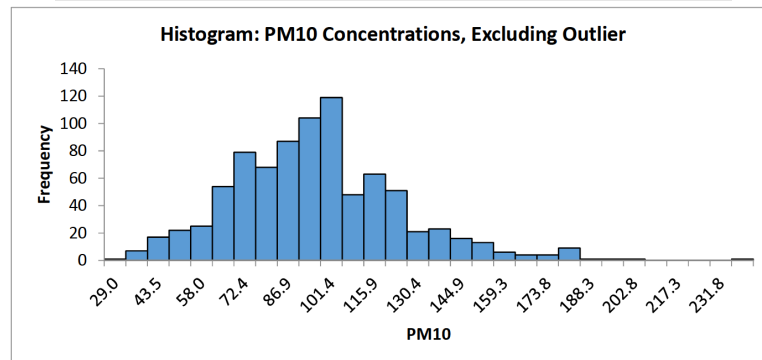
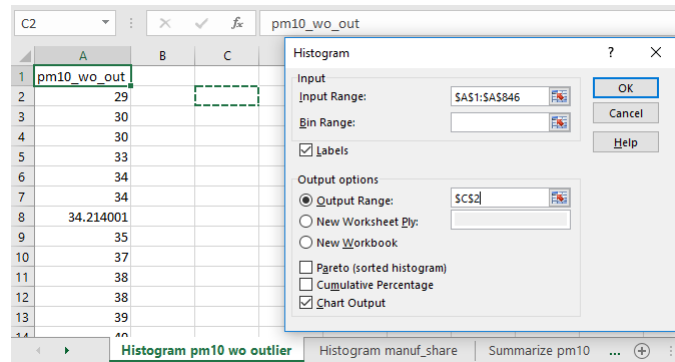
EXCEL TIPS: In Data Analysis select Histogram. Leave Bin Range empty. For Input Range, select only actual observations. Excel sets n to the number of rows selected, ignoring missing values. An entire column (`pol_chn!$J:$J`) means n is the maximum number of rows possible, which yields a histogram with too many (very skinny) bins.

- (i) A histogram of PM10 visually reveals an outlier. (The descriptive statistics in part 1b on page 19 also highlighted this outlier.) **Notice** it in the histogram below: $550 \mu\text{g}/\text{m}^3$.

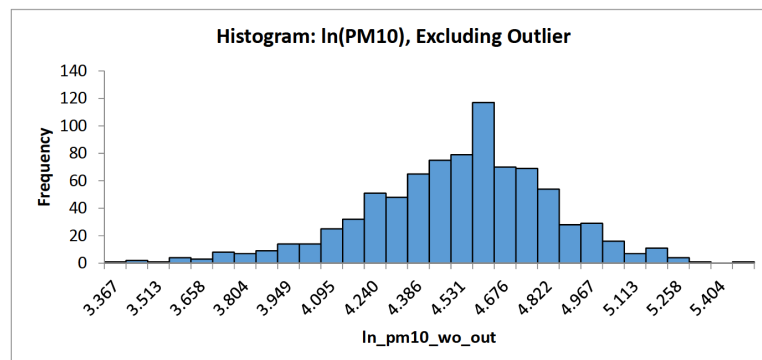


- (j) **Create** a histogram of PM10 excluding the outlier with PM10 of 550 for the city with id 315 (Karamay) in 2003. **Verify** that your histogram looks similar to the one below.

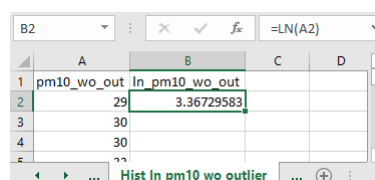
EXCEL TIPS: Copy the entire column for pm10 to a new worksheet named “Histogram pm10 wo outlier.” Sort it in ascending order. Clear the cell containing the last (highest) value. Rename the variable pm10-wo-out. Select the non-missing values (and the variable name) and insert a histogram (with Data Analysis) into that same new worksheet.



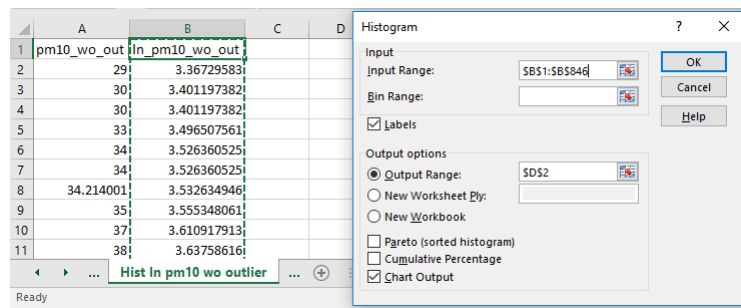
- (k) **Create** a histogram of the natural log of PM10 (excluding the outlier).



EXCEL TIPS: Copy the entire column for pm10_wo_out (from the previous part) to a new worksheet named “Hist ln pm10 wo outlier.” Use the function LN() to create a variable named ln_pm10_wo_out. After setting up the first cell, double click on the little green square in the bottom right (see screenshot below) to autofill the column.



Insert a histogram of `ln_pm10_wo_out` in this same new worksheet.



- (1) Use [pol_chn.xlsx](#) to review the useful PivotTables from Module A.1.
- How many different years and how many different cities are there in these data?
Verify there are 10 unique years (2003 through 2012) and 85 unique cities.
EXCEL TIPS: One way to count unique values is to select a variable and insert a PivotTable. Use the COUNT function, *not* including (blank), to count the row labels (see cell D16 and its formula in E16 below). Repeat these steps for the `city_id` variable.

1	year	city_id		
2	2003	2	Row Labels	
3	2003	3	2003	
4	2003	5	2004	
5	2003	7	2005	
6	2003	17	2006	
7	2003	18	2007	
8	2003	19	2008	
9	2003	20	2009	
10	2003	29	2010	
11	2003	32	2011	
12	2003	39	2012	
13	2003	40	(blank)	
14	2003	41	Grand Total	
15	2003	42		
16	2003	54		10 =COUNT(D3:D12)
17	2003	63		

You can also cross tabulate these two variables. Select `year` and `city_id` and insert a PivotTable and drag `year` to COLUMNS, `city_id` to ROWS, and drag (another copy of) `city_id` to Σ VALUES. To improve readability, from the Design tab, select Show in Tabular Form under Report Layout. The cross tabulation output shows that these data are a *balanced panel*. Each city is observed each year: $n = 850$ is exactly 10 years times 85 cities. In an *unbalanced panel*, some cities would be observed only some years.

1	year	city_id		
2	2003	2		
3	2003	3		
4	2003	5		
5	2003	7		
6	2003	17		
7	2003	18		
8	2003	19		
9	2003	20		
10	2003	29		
11	2003	32		
12	2003	39		
13	2003	40		
14	2003	41		
15	2003	42		
16	2003	54		
17	2003	63		
18	2003	64		

- ii. Across all 85 cities in China, **compute** the average pollution level and manufacturing share for each year.

EXCEL TIPS: Insert a new worksheet. Copy-and-paste the variables year, pm10, and manuf_share: use the shortcut **Ctrl + Click** to select non-adjacent columns. (For macs, use **command** instead of **Ctrl**.) Insert a PivotTable. Drag pm10 and manuf_share to Σ VALUES. Drag year to ROWS. In Σ VALUES, using the drop down menu, change the Value Field Settings for pm10 and manuf_share to AVERAGE.

Row Labels	Average of pm10	Average of manuf_share
2003	117.1358026	50.70905873
2004	106.6470591	52.12400018
2005	95.04705894	49.47199988
2006	96.36470612	50.26082375
2007	90.82352941	50.64105876
2008	87.36470588	51.23047056
2009	85.90588235	49.55399991
2010	87.31764706	50.06188234
2011	85.67434109	50.86141158
2012	82.85777613	49.96
Grand Total	93.40216315	50.48747057

Test/exam examples: Zheng and Kahn (2017) has not appeared, but other cases have.

- Questions (3) and (4)(b), [April 2018 Final Exam](#) (with [solutions](#))
- Questions (1), (3)(b), (4)(b), and 5(c), [October 2017 Test #1](#) (with [solutions](#))
- Question (2)(a), [November 2016 Test #1](#) (with [solutions](#)).
- Question (4)(b), [April 2017 Final Exam](#) (with [solutions](#)).

2. Consider that Stata (often used for figures in this handbook) draws histograms differently. Recall Carlin et al. (2017) and [cred_card.xlsx](#) from Module A.1. Use the first 100 observations of the variable age: $n = 100$ with the data as originally sorted by resp_id (i.e. do *not* re-sort in Excel).⁵

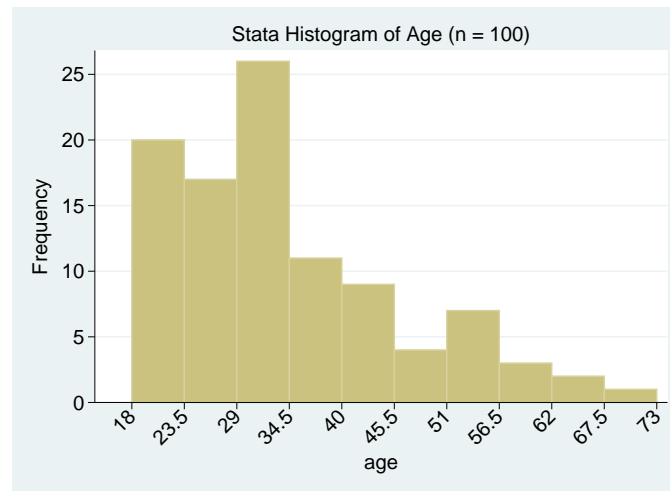
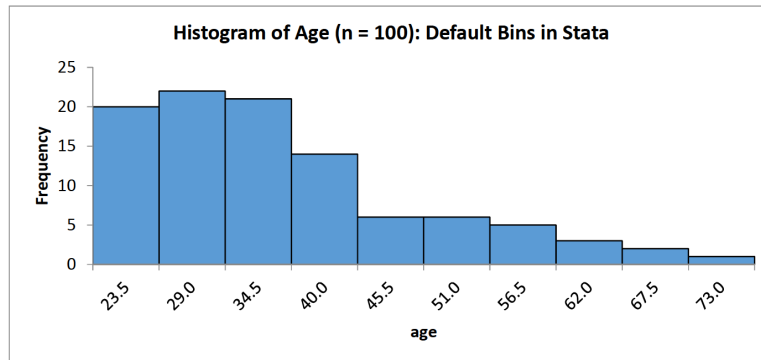
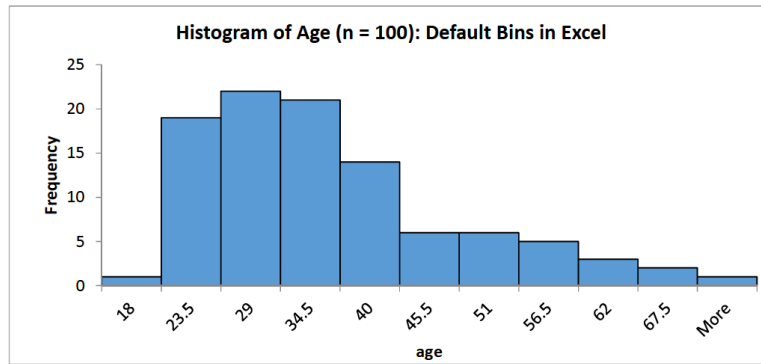
- (a) Review the three frequency histograms below. All three have a bin width of exactly 5.5.

If the data, bin widths, and histogram type match, why aren't the histograms identical?

- When an observation is on the border between two bins, Excel and Stata make different choices, which explains why the second and third histograms differ. Because age is an integer and the bin width is 5.5, this situation arises a lot. In the first 100 observations, five people are 29 years old, which is on the boundary between two bins. Stata puts borderline cases in the bin to the right, whereas Excel chooses left. In the bin from 23.5 to 29, Excel puts 22 observations ($23.5 < age \leq 29$) whereas Stata puts 17 observations ($23.5 \leq age < 29$), which excludes the five 29-year-olds.
- The first histogram has an extra bin ending at 18. Because Excel puts borderline cases in the bin to the left, it has the one 18-year-old (no one is younger). The second Excel histogram uses the default bins in Stata: the first bin ends at 23.5 and has the 20 people who are 18, 19, 20, 21, 22 or 23 years old.

- (b) So which one is correct? There is no such thing as *the* correct histogram (but there are many possible misleading ones). Histograms give an overall visual summary of the distribution of an interval variable. Focus on the *overall* picture presented by a histogram.

⁵Excel and Stata also differ in the rules for sorting. If you re-sort by resp_id in Excel, it will put the observations in a different order. For simple sorts, for example numeric data with no characters, Stata and Excel do the same thing.



Interpretation tips: Given the Stata histogram (the third one), should we say that the shape is bimodal? Multimodal? It would be wrong to conclude that the distribution is bimodal or multimodal: those are just little blips affecting the first bin and seventh bin. There is no systematic evidence supporting the inference that the age distribution has more than one mode. Slightly different, and still reasonable, choices when drawing the histogram make those little blips disappear. The distribution is unimodal: one major peak for those in their late twenties to mid thirties. Also, all three histograms clearly show that the age distribution is positively skewed. Most respondents on Amazon's Mechanical Turk are quite young, which is not surprising as it is a fairly new online technology. In fact, in this case, it is skew that stands out, not modality. A good interpretation correctly focuses on the skew.

3. **Browse** the data file [who-aap-database-may2016.xlsx](#) and verify the pollution levels for the examples (Beijing, Toronto, LA, Rome, Tokyo, and Delhi) given in the preamble to Module A.2 on page 17.

4. *Become familiar* with some additional data sets described below.

- (a) **PENN WORLD TABLES:** In Modules B.2 and B.3 starting on page 46, we consider the paper “Asiaphoria Meets Regression to the Mean” by Pritchett and Summers (2014). Those authors use an important database: the [Penn World Tables \(PWT\)](#). As of April 1, 2024, the most recent version is version 10.01 (published on January 23, 2023).

- [pwt1001.xlsx](#): This is the January 23, 2023 version of the publicly posted PWT 10.01 data. Browse through all three worksheets to familiarize yourself with these data.
- You will encounter earlier versions – including versions 8.0 and 9.0 – elsewhere in this handbook. They have a similar design.

Test/exam examples: PWT data have appeared on many, many tests and exams. These are listed in later modules when you work with these data.

- (b) **WORLD HAPPINESS REPORT:** Since 2012, the [World Happiness Report](#) has been published annually. One important data source it uses is the [Gallup World Poll](#). It has been a frequent example in ECO220Y and is a great chance to continue developing your skill of understanding a wide variety of data. Page 16 of the [2021, World Happiness Report](#) explains how the researchers measure “happiness.” In the section titled “Life evaluations,” they explain the key variable: “The Gallup World Poll, which remains the principal source of data in this report, asks respondents to evaluate their current life as a whole using the image of a ladder, with the best possible life for them as a 10 and worst possible as a 0. Each respondent provides a numerical response on this scale, referred to as the Cantril ladder. Typically, around 1,000 responses are gathered annually for each country.” The posted data have the same structure each year: browse and familiarize yourself with these data.

- [whr_2021.xlsx](#): These are the data posted with the 2021 World Happiness Report.
- [whr_2022.xlsx](#): These are the data posted with the 2022 World Happiness Report.
- [whr_2023.xlsx](#): These are the data posted with the 2023 World Happiness Report.
- [whr_2024.xlsx](#): These are the data posted with the 2024 World Happiness Report.
- Unfortunately, starting in 2025 the World Happiness Report website stopped posting most of the replication data. Fortunately, the University of Toronto has a license for the Gallup World Poll data. You will work with these proprietary data near the end of the Winter term.

Test/exam examples: Analyses of the WHR (happiness) data have appeared in recent tests (but you are not yet ready to solve these test questions).

- Questions (1) and (2), [April 2025 Exam](#) (with [solutions](#))
- Question (6), [March 2025 Test #4](#) (with [solutions](#))
- Question (8), [October 2024 Test #1](#) (with [solutions](#))
- Question (2), [May 2023 Test #1](#) (with [solutions](#))
- Question (2), [March 2023 Test #4](#) (with [solutions](#))

A.0.0 Practice questions for Module A

- Q1.** Recall the caution about missing values in Section 3.2 on page 4. In Module B, we study a major database: the Penn World Tables (PWT). In preparing the data files for this handbook, we cleaned the original data from many sources to make them easier for you to understand and to work with. Some data required a lot of work. The PWT data are excellently prepared: we simply removed missing values and removed some variables we do not use. For this question use [pwt90.xlsx](#), which are the version 9.0 data exactly as downloaded from <https://www.rug.nl/ggdc/productivity/pwt/pwt-releases/pwt9.0> (retrieved April 12, 2018).
- (a) How many unique countries are in these data? How many unique years are in these data?
 - (b) Create a new variable measuring the natural logarithm of the variable `rgdpe`. What is the standard deviation of that new variable?
 - (c) Create a new variable measuring real GDP in billions of 2011US\$ by dividing the variable `rgdpe` by 1,000. Now create a final new variable that is the natural logarithm of your variable measuring GDP in billions. What is the mean of that final new variable?
- Q2.** Recalling Carlin et al. (2017), use [cred_card.xlsx](#). Create a cross tabulation of the variables `confidence` and `easy_choice`.
- (a) What do the variables `confidence` and `easy_choice` measure?
 - (b) Using your cross tabulation, fill in the blanks with the appropriate *number* of respondents. Of the 1,603 respondents, _____ respondents gave the same answer to the questions about confidence and easiness. _____ respondents indicated both having high confidence (6 or higher) and finding the choice quite easy (6 or higher). Among those respondents indicating the highest confidence (7), _____ respondents found the choice at least fairly easy (5 or higher). Among those respondents indicating the choice was very easy (7), _____ respondents were at least fairly confident in their choice (5 or higher).
 - (c) Using your cross tabulation, fill in the blanks with the appropriate *percent* (rounded to the nearest first decimal place). _____ percent of respondents quite strongly disagreed (2 or lower) with both the easiness and confidence questions. Among those respondents that quite strongly disagreed (2 or lower) with the easiness question, _____ percent quite strongly disagreed (2 or lower) with the confidence question. Among those respondents that quite strongly disagreed (2 or lower) with the confidence question, _____ percent quite strongly disagreed (2 or lower) with the easiness question. Among those respondents that were neutral (4) on the confidence question, _____ percent agreed (5 or higher) with the easiness question.
 - (d) Using your cross tabulation, fill in the blanks with the appropriate pair of answers to the respective questions in the format # and # (e.g. 3 and 4). The most common pair of answers to the easiness and confidence questions are _____, respectively. The second most common pair of answers to the easiness and confidence questions are _____, respectively.
 - (e) Using your cross tabulation, fill in the blanks with the appropriate integer. There are _____ pairs of values for the easiness and confidence questions that never occur in these data. There are _____ pairs of values for the easiness and confidence questions that only occur one time in these data.

Q3. Recalling Carlin et al. (2017), use [cred.card.xlsx](#).

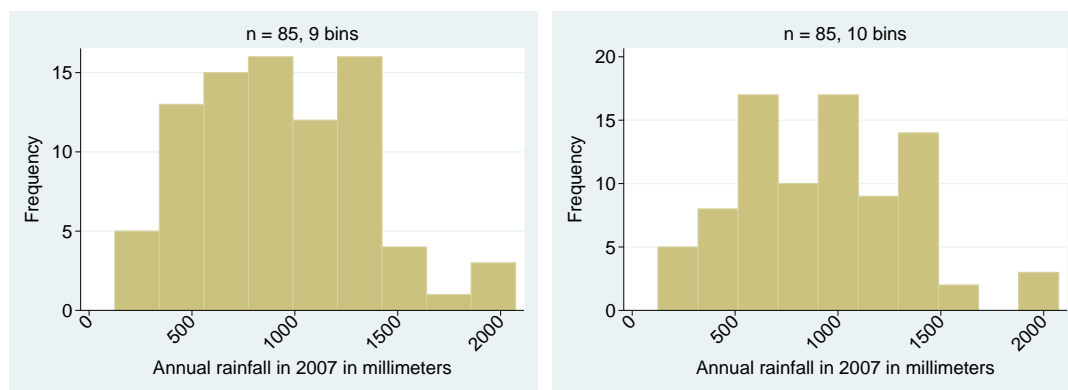
- (a) Complete this table for the variable age.

Descriptive Statistics for Age	
Mean	
S.D.	
Median	
Minimum	
Maximum	
Obs.	

- (b) Identify the outlier and repeat the previous part excluding the outlier.
- (c) Tabulate age. Using your results, fill in the blanks with the appropriate *number* of observations: Of the 1,603 respondents, _____ did not answer the question about age, _____ are under 21 years old, _____ are 24 years old (the mode), _____ are over 30, and _____ are 65 years old or older. (Again, use the original age variable as these summary values are not sensitive to outliers.)
- (d) Draw a histogram of age excluding the outlier. Describe the shape of the histogram.
- (e) Draw a histogram of the natural log of age excluding the outlier. Describe the shape of the histogram.

Q4. Recall Zheng and Kahn (2017) and the data [pol.chn.xlsx](#).

- (a) Which kind of data are these? What is the unit of observation? How many observations are there? How many variables? How many identifier variables? How many nominal (categorical) variables? How many interval variables?
- Suppose we dropped all observations except for the year 2011 and kept all variables. Which kind of data would these be? Unit of observation? Number of observations?
 - Suppose we dropped all observations except for the city of Beijing and kept all variables. Which kind of data would these be? Unit of observation? Number of observations?
- (b) Create a variable measuring GDP *per capita* in *U.S. dollars*. (Use the exchange rate in the data.) What are the mean, median, s.d., minimum, and maximum value of this variable?
- (c) Consider the following two Stata histograms.



- Why is the number of observations 85? In contrast, the number of observations is 850 for the variable `manuf_share` in these same data?

- ii. What is the *suggested* number of bins according to $\left(\text{MIN} \left\{ \sqrt{n}, \frac{10 \ln(n)}{\ln(10)} \right\} \right)$, which happens to be the specific formula used by Stata?
- iii. Which histogram is the correct one: the one with 9 bins or 10 bins?
- iv. Describe the shape of the distribution of 2007 rainfall across the 85 Chinese cities.

Q5. Use [assor_ctor_goog_oecd.xlsx](#).

- (a) Using the data in the worksheet titled OECD (Ren, Ene., CO2, GDP, Oil), construct a variable measuring the percent of the primary energy supply that is *not* renewable energy. Draw a histogram of the percent of the primary energy supply from non-renewable sources using ONLY the data for 2013. Which is the best way to describe the shape of this histogram: A = symmetric, B = positively skewed, C = negatively skewed, D = multimodal?
- (b) Continuing with the worksheet in the previous part, what is the bin WIDTH if we use the formula \sqrt{n} to set the number of bins for a histogram of CO2 emissions using only the observations for Canada (which has the 3-letter country code CAN)?

For extra practice, additional questions, with an ^e superscript (*e* for extra), are next.

Q^e1. Recalling Carlin et al. (2017), use [cred_card.xlsx](#). Do higher income respondents make better choices among the credit card offers? To answer, report the percent choosing the dominant card by income level. Which describes the results: “higher income respondents tend to make better choices,” “lower income respondents tend to make better choices,” or “income levels and choices seem unrelated”?

Q^e2. Recalling Carlin et al. (2017), use [cred_card.xlsx](#). Do the results in Figure 6 vary by sex? Fill in the blanks with the percent (rounded to the nearest first decimal place). Among female respondents _____ percent chose the dominant card. Among male respondents _____ percent chose the dominant card. Among female respondents that saw the implemental video and no superfluous taglines _____ percent chose the dominant card. Among male respondents that saw the implemental video and no superfluous taglines _____ percent chose the dominant card. Among female respondents that saw the baseline video and superfluous taglines _____ percent chose the dominant card. Among male respondents that saw the baseline video and superfluous taglines _____ percent chose the dominant card.

Q^e3. Use [assor_ctor_goog_oecd.xlsx](#).

- (a) Using the data in the worksheet titled City of Toronto (Wellbeing), construct a histogram of home prices. Which is the best way to describe the shape of this histogram: A = symmetric, B = positively skewed, C = negatively skewed, D = multimodal?
- (b) Continuing with the previous part, construct a histogram of the natural logarithm of home prices. Which is the best way to describe the shape of this histogram compared to the shape when the home price is not logged: A = it is now skewed in the opposite direction, B = it is now symmetric, C = it is still positively skewed, but the skew is less extreme, D = it is still negatively skewed, but the skew is less extreme?

- (c) Using the data in the worksheet titled Google Finance (Apple Stock), construct a histogram of the natural log of the price of Apple stock. Which is the best way to describe the shape of this histogram: A = symmetric, B = positively skewed, C = negatively skewed, D = multimodal?
- (d) Continuing with the previous part and if we use all available observations before January 1, 2007, how many bins are suggested by the formula $= MIN \left\{ \sqrt{n}, \frac{10 \ln(n)}{\ln(10)} \right\}$?

Answers for Module A practice questions:

- A1.** (a) There are 182 unique countries and 65 unique years.
 (b) 2.26078 (Note: The correct number of non-missing observations is 9,439.)
 (c) 3.256812 (Note: The correct number of non-missing observations is 9,439.)
- A2.** (a) Referring to the worksheet “readme” in [cred_card.xlsx](#), confidence measures how intensely the respondent agreed with the statement that “Choosing the best credit card was easy” on a 1 to 7 Likert scale where 1 is strongly disagree and 7 is strongly agree. The variable easy_choice measures how intensely the respondent agreed that choosing the best credit card was easy on a 1 to 7 Likert scale where 1 is strongly disagree and 7 is strongly agree.
- (b) For all remaining parts, use a PivotTable to answer. To add helpful row and column labels, go to the Design tab, Report Layout, and select Show in Tabular Form.

	A	B	C	D	E	F	G	H	I
1									
2									
3	Count of confidence	confidence							
4	easy_choice	1	2	3	4	5	6	7	Grand Total
5	1	36	24	16	20	28	6	7	137
6	2	7	61	45	46	79	37	7	282
7	3		6	36	69	169	69	8	357
8	4			14	68	100	40	7	229
9	5	1	2	3	18	110	142	38	314
10	6		1	2	4	24	130	34	195
11	7		1		1	1	14	71	89
12	Grand Total	45	95	116	226	511	438	172	1603
13									

PivotTable Fields

Choose fields to add to report:

Drag fields between areas below:

FILTERS

COLUMNS

confidence

ROWS

easy_choice

VALUES

Count of conf...

Of the 1,603 respondents, 512 respondents ($= 36 + 61 + 36 + 68 + 110 + 130 + 71$) gave the same answer to the questions about about confidence and easiness. 249 respondents ($= 130 + 34 + 14 + 71$) indicated both having high confidence (6 or higher) and finding the choice quite easy (6 or higher). Among those respondents indicating the highest confidence (7), 143 respondents ($= 38 + 34 + 71$) found the choice at least fairly easy (5 or higher). Among those respondents indicating the choice was very easy (7), 86 respondents ($= 1 + 14 + 71$) were at least fairly confident in their choice (5 or higher).

- (c) Answer by referencing the cross tabulation. 8.0 percent of respondents ($= 100 * (36 + 24 + 7 + 61) / 1,603$) quite strongly disagreed (2 or lower) with both the easiness and confidence questions. Among those respondents that quite strongly disagreed (2 or lower) with the easiness question, 30.5 percent ($= 100 * (36 + 24 + 7 + 61) / (137 + 282)$) quite strongly disagreed (2 or lower) with the confidence question. Among those respondents that quite strongly disagreed (2 or lower) with the confidence question, 91.4 percent ($= 100 * (36 + 24 + 7 + 61) / (45 + 95)$) quite strongly disagreed (2 or lower) with the easiness question. Among those respondents that were neutral (4) on the confidence question, 10.2 percent ($= 100 * (18 + 4 + 1) / (226)$) agreed (5 or higher) with the easiness question.
- (d) Answer by referencing the cross tabulation. The most common pair of answers to the easiness and confidence questions are 3 and 5, respectively (this pair occurs 169 times). The second most common pair of answers to the easiness and confidence questions are 5 and 6, respectively (this pair occurs 142 times).
- (e) Answer by referencing the cross tabulation. There are 5 pairs of values for the easiness and confidence questions that never occur in these data. There are 6 pairs of values for the easiness and confidence questions that only occur one time in these data.

A3. To answer, use Descriptive Statistics under Data Analysis.

(a)

Descriptive Statistics for Age

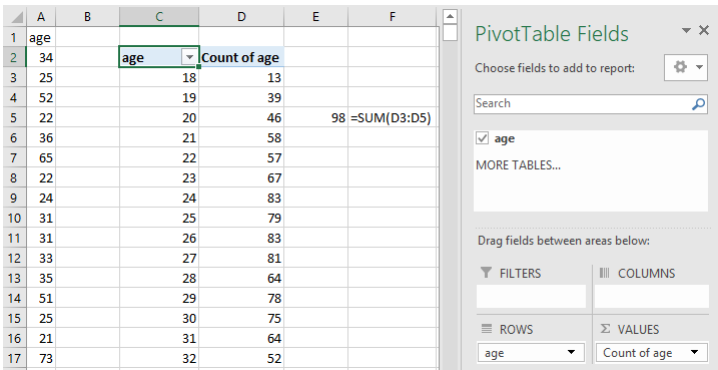
Mean	33.5
S.D.	12.2
Median	30
Minimum	18
Maximum	200
Obs.	1,600

(b) The outlier is the respondent who is 200 years old.

Descriptive Statistics for Age, Excluding Outlier

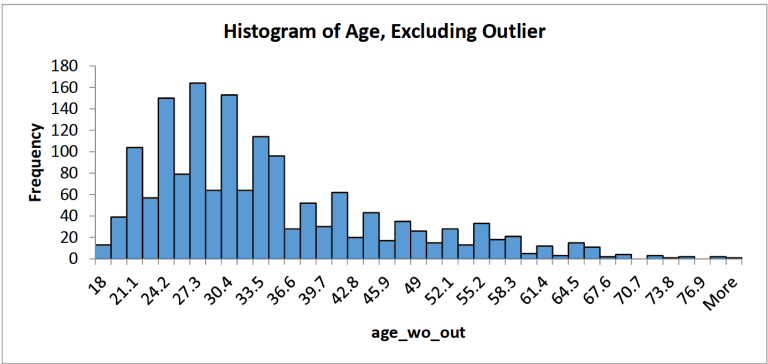
Mean	33.4
S.D.	11.5
Median	30
Minimum	18
Maximum	80
Obs.	1,599

(c) To answer, use a PivotTable of age.

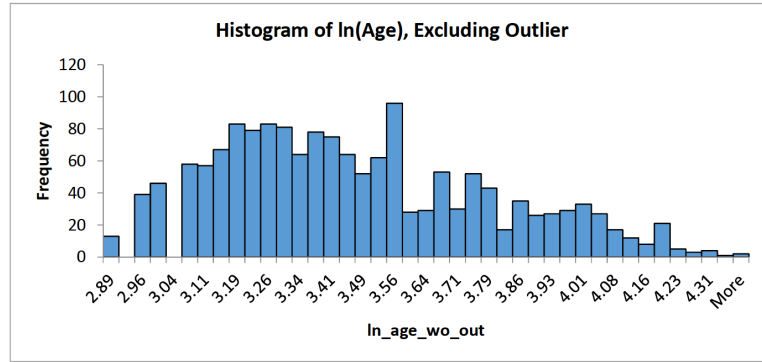


Of the 1,603 respondents, 3 did not answer the question about age, 98 are under 21 years old, 83 are 24 years old (the mode), 777 are over 30, and 27 are 65 years old or older.

(d) The histogram is positively skewed (aka right skewed).



(e) The histogram is somewhat positively skewed, but it is less skewed than the age distribution before the natural log transformation.



- A4.** (a) These data are panel (longitudinal) data. The unit of observation is a particular city in a particular year. There are 850 observations corresponding to 85 cities for each of 10 years. There are 17 variables. There are 7 identifier variables: year, city_id, and the dummy (indicator) variables for the five large cities. There are six nominal (categorical) variables: city_id, shanghai, beijing, tianjin, guangzhou, and shenzhen. There are eleven interval variables: year, pm10, rainfall, longitude, latitude, temp_index, gdp, pop, manuf_share, edu2000, and aexchus.
- These would be cross-sectional data: we have cross-section of cities in 2011. The unit of observation would be a city. There would be 85 observations.
 - These would be time series data: we are following the city of Beijing each year. The unit of observation would be a year. There would be 10 observations.
- (b) The mean GDP per capita is 9,343 U.S. dollars. The median is 7,426 U.S. dollars. The standard deviation is 7,334 U.S. dollars. The minimum value is 1,122 U.S. dollars. The maximum value (city of Shenzhen in 2012) is 63,892 U.S. dollars.
- (c)
 - Recall that rainfall is 2007 annual rainfall in millimeters for each of the 85 cities. These same values are repeated in the data 10 times (for each of the 10 years). It would have been better if the authors had obtained a measure of rainfall in each city and *in each year*, but they did not. Hence, we only actually have 85 observations.
 - 9 bins
 - They are both perfectly reasonable histograms. Remember that there is no one correct histogram. There are many formulas out there to give *suggestions* about the number of bins and they make a range of suggestions. Remember that histograms are meant to give an *overall visual summary* of the distribution of a variable. It is a simplification and there is no single correct way to do that simplification.
 - The shape is fairly symmetric and nearly Normal (Bell). (This overview is unchanged whether we use 9 or 10 bins.)
- A5.** (a) No matter which (standard) formula you use to determine the histogram bins, the distribution of the percent of energy from non-renewable sources in 2013 is clearly *negatively skewed*. (Note: Make sure you sorted the data by year to use only the 26 observations for 2013. Also, make sure you computed the percent from non-renewable sources as 100 minus the percent from renewable sources.)
- (b) There are 15 observations, which suggests 4 bins (applying standard rounding to $\sqrt{(15)}$). Using Descriptive Statistics in Data Analysis for the CO2 emissions variable, we obtain the

minimum value of 15.27 and the maximum value of 16.91, for a range of 1.64 and hence a bin width of 0.41 ($=1.64/4$). (Note: Make sure you sorted the data by country to use only the 15 observations for Canada.)

Answers to the additional questions for extra practice.

A^e1. Use a PivotTable with hh.inc and chosedom. The mean of chosedom tells the share of respondents in each income category picking the dominant card. For example, 48.3% of those in the lowest income category select the dominant card. Income levels and choices seem unrelated.

Row Labels	Average of chosedom
Under \$25,000	0.482666667
\$25,000 - \$49,999	0.489021956
\$50,000 - \$74,999	0.473988439
\$75,000 - \$99,999	0.507614213
\$100,000 - \$149,999	0.510948905
\$150,000 or over	0.489361702
Grand Total	0.488459139

A^e2. Answer all parts efficiently by creating a single PivotTable. Put the variable male in COLUMNS, the variables video and tagline in ROWS, and the average of the variable chosedom in Σ VALUES. Among female respondents 50.5 percent chose the dominant card. Among male respondents 47.4 percent chose the dominant card. Among female respondents that saw the implemental video and no superfluous taglines 67.2 percent chose the dominant card. Among male respondents that saw the implemental video and no superfluous taglines 62.9 percent chose the dominant card. Among female respondents that saw the baseline video and superfluous taglines 40.9 percent chose the dominant card. Among male respondents that saw the baseline video and superfluous taglines 33.5 percent chose the dominant card.

- A^e3.** (a) No matter which (standard) formula you use to determine the histogram bins, the distribution of home prices is clearly *positively skewed*.
- (b) No matter which (standard) formula you use to determine the histogram bins, the distribution of the natural log of home prices is somewhat positively skewed, but much less severely positively skewed than the distribution of (unlogged) home prices.
- (c) No matter which (standard) formula you use to determine the histogram bins, the distribution of the natural log of Apple stock prices is clearly *multimodal*.
- (d) There are 1,380 observations, which suggests 31 bins.

B Module B: Describing Relationships & Asiaphoria

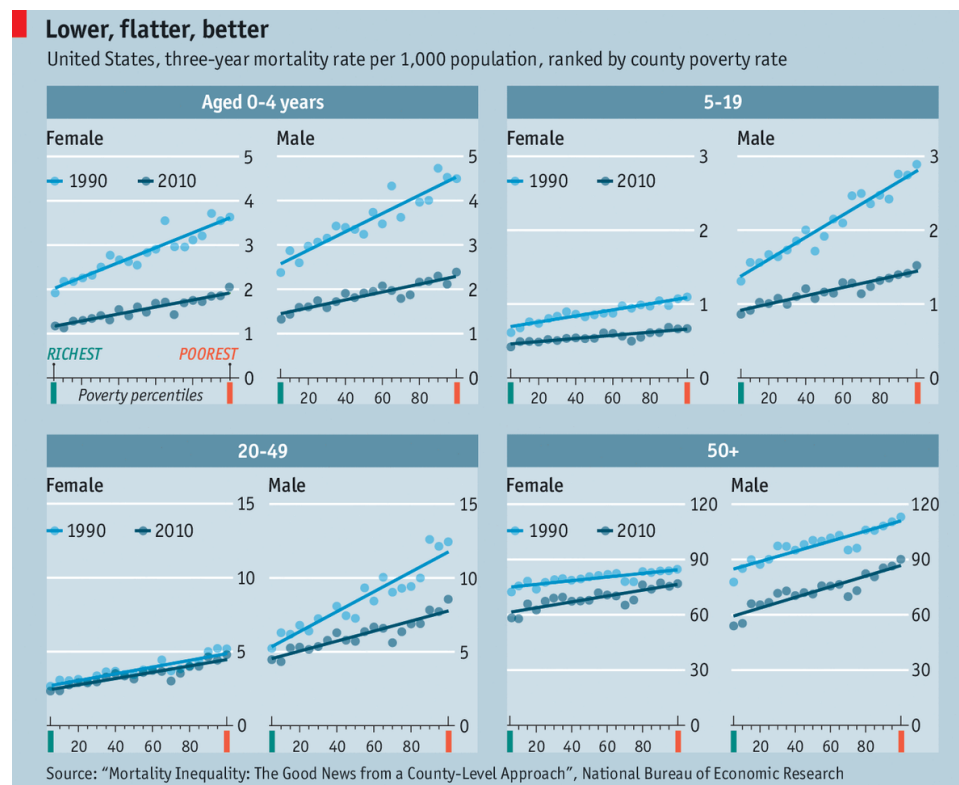
B.1 Module B.1: Association, Correlation, Regression & Composition Effects

Concepts: Describe relationships among quantitative variables using graphs, statistics, and regression. Confront “Simpson’s paradox,” which economists call “composition effects.”

Case studies: We consider the Big Mac index by *The Economist*. We replicate parts of an academic journal article “Mortality Inequality: The Good News from a County-Level Approach,” abbreviated Currie and Schwandt (2016) (featured in *The Economist*).

Required readings: Sections 4.5 (Simpson’s Paradox), 7.1-7.7. Background reading:

- *The Economist* reproduces a key figure from Currie and Schwandt (2016) adding formatting and captions. The title “Lower, flatter, better” and subtitle “United States, three-year mortality rate per 1,000 population, ranked by county poverty rate” help convey the main message. The y-axis is the three-year mortality rate per 1,000 population. See the helpful reminder about the x-axis: counties in the 100th poverty percentile are the **POOREST** (highest poverty) whereas counties in the 1st poverty percentile are the **RICHEST** (lowest poverty).



- A well-written paper explains the take-away messages. In these excerpts, Currie and Schwandt (2016) clarify the y-axis and interpret the top half of the figure. “The three-year mortality rate in 1990 is the ratio of all deaths in a cohort [sex and age group] between April 1, 1990, and March 31, 1993, divided by the 1990 Census population count [for that sex and age group]” (p. 37). Why use three-year mortality rates? Why not one-year? The authors explain that three-year mortality rates “helps to minimize noise due to measurement error and to avoid counties

reporting zero deaths” (p. 36). Mortality has both systematic and chance explanations. For example, in some years in some places there is an inexplicable spike in motorcycle deaths (bad luck). The systematic effect of community-level poverty on mortality rates will persist over the years whereas the random part will not. Combining years reduces the importance of the random part and sharpens focus on the role of poverty.

- “[The figure] shows three-year mortality rates at the level of county groups, with counties ranked by the share of their population below the poverty line, for males and females in four different age groups. In these figures, each marker shows the mortality rate for a bin representing 5 percent of the US population in the relevant year. A slope that becomes steeper over time implies increasing inequality and vice versa” (p. 40).
- “[The figure] shows dramatic reductions in mortality among children aged zero to four between 1990 and 2010. Overall, the reductions in under-five mortality were much greater in poorer counties than in richer ones, and slightly larger for males than for females. For example, the under-five mortality rate for males fell from 4.5 per 1,000 in 1990 to 2.4 per 1,000 in the poorest counties, compared to a decline from 2.4 to 1.3 per 1,000 in the richest counties over the same period. Among children aged 5 to 19, there were large reductions in mortality for males, with more modest reductions for females (from already low levels). Once again, reductions were larger in poorer counties, implying significant reductions in mortality inequality” (p. 40).
- On page 43 is Figure 3 from Currie and Schwandt (2016) (from which *The Economist* created its figure). The note says “Mortality rates in 2000 and 2010 are age-adjusted using the 1990 population, that is, they account for changes in the age structure within age, gender, and county groups since 1990.” What does that mean? They use adjusted mortality to avoid composition effects (aka Simpson’s paradox). Table B.1 illustrates with a made-up county.

Table B.1: Illustrative county showing composition effects (aka Simpson’s paradox)

Age group	Years	Deaths	Population	Mortality per 1,000	Adjusted pop.	Adj. deaths	Adj. mortality per 1,000
40-44	1990-93	300	50000	6.00	50000	300.00	6.00
40-44	2000-03	290	51000	5.69	50000	284.31	5.69
40-44	2010-13	280	52000	5.38	50000	269.23	5.38
45-59	1990-93	4800	300000	16.00	300000	4800.00	16.00
45-59	2000-03	6300	400000	15.75	300000	4725.00	15.75
45-59	2010-13	7800	500000	15.60	300000	4680.00	15.60
40-59	1990-93	5100	350000	14.57	350000	5100.00	14.57
40-59	2000-03	6590	451000	14.61	350000	5009.31	14.31
40-59	2010-13	8080	552000	14.64	350000	4949.23	14.14

- From the numbers in boldface you can compute every other number. Deaths is the number of people in that county and in that age group who died in those years. Population is the total number of people in that age group living in the county in the start year. For example, in the years 2000 to 2003, 6,300 people aged 45-59 died and in the year 2000 there are 400,000 people aged 45-59 in the county. From these we can compute the three-year mortality rate per 1,000 people: from 2000 to 2003, for every 1,000 people aged 45-59, 15.75 people died.
- How to combine the 40-44 and 45-59 age groups into a bigger 40-59 year old age group? It’s not hard to compute the total deaths and population (summing over the two age groups) and

the mortality per 1,000 as above. For example, in the years 2010 to 2013 there are 14.64 deaths per 1,000 people aged 40-59. While this is computationally OK, there is a serious problem if we wish to investigate, like Currie and Schwandt (2016), how mortality is changing over time.

- To see the problem, notice that BOTH age groups have declining mortality rates (i.e. things improve from 1990 to 2010). However, a simple calculation (i.e. not adjusting) paradoxically implies that mortality rates are increasing (i.e. things worsen from 1990 to 2010) for the two groups combined. How can two positives lead to a negative? TWO things are changing: the mortality rates AND the composition of the combined age groups. In Table B.1, age group 45-59 is much larger and has a higher mortality rate. Further, it has a growing population size. Hence, the combined group has an increasing fraction of the older age group (45-59).
- Currie and Schwandt (2016) construct *adjusted* mortality to control for changes in the relative population sizes of groups 40-44 and 45-59. It holds the composition of the combined age group fixed. With adjusted mortality for 40-59 year olds (last column of Table B.1), combining two groups each with declining mortality rates yields a declining mortality rate. It fixes the paradox.
- How to compute adjusted mortality? First, for the 40-44 and 45-59 age groups, construct adjusted population: hold it fixed at the 1990 level. Next, for these two age groups, construct adjusted deaths: adjusted population times mortality per 1,000 divided by 1,000. This is what total deaths would have been if population stayed at the 1990 level. Next, sum the adjusted values for total (adjusted) population and (adjusted) deaths for the combined age group (40-59). Last, compute adjusted mortality per 1,000 using adjusted population and adjusted deaths.
- For the other case study, review Figure 1, which shows the scatter diagram and OLS line for the January 2017 analysis underlying *The Economist*'s construction of the Big Mac index.⁶



Figure 1: *The Economist* online showing the OLS results (red line) underlying their **January 2017** analysis, retrieved June 13, 2017. The regression line: $\hat{y} = 2.487433 + 0.0394057x$.

Datasets: For the Big Mac index: [big_mac_jan_2017.xlsx](#). For Currie and Schwandt (2016): [mort_in_figure_3_table_a3.xlsx](#) and [mort_in_illustrate_composition_effects.xlsx](#) where “mort_in” abbreviates “Mortality Inequality” from the title.

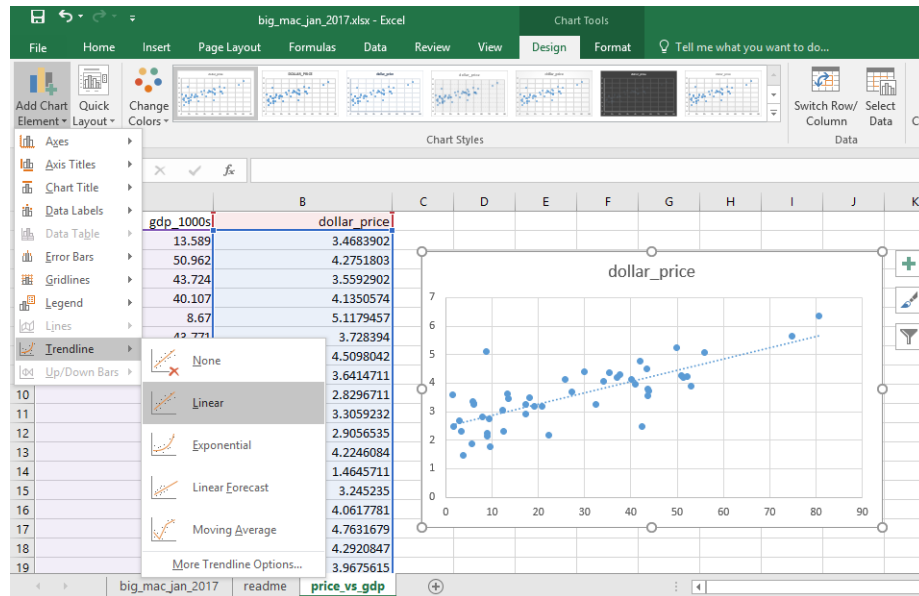
⁶ *The Economist* uses the simple regression in Figure 1 to compute the adjusted Big Mac index that controls for differences in richness across countries. In 2018, *The Economist* posted replication files and updated its Big Mac Index webpage but no longer displays Figure 1.

Interactive module materials for Module B.1:

1. Consider the Big Mac index by *The Economist*:

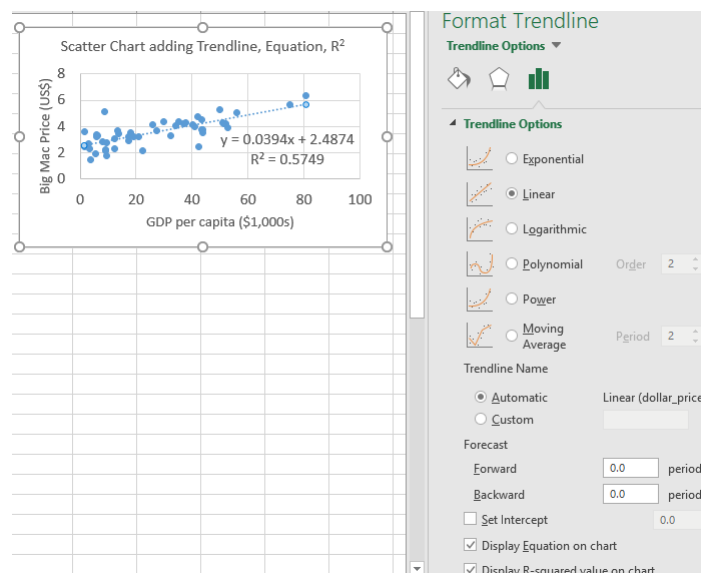
- Using [big_mac_jan_2017.xlsx](#), **create** a variable measuring GDP per capita in thousands of dollars (not dollars), giving it a meaningful name: gdp_1000s.
- Replicate** the scatter plot, including the regression line, in Figure 1 on page 39.

EXCEL TIPS: Insert a new worksheet “price_vs_gdp” and copy gdp_1000s and Paste Special, Values. Next copy and paste dollar_price. Select the entire two columns. Under the Insert tab, select Scatter. (Excel scatter plots treat the first variable as x and the second as y.) To include the regression line, add a Trendline as illustrated below.



- Replicate** the OLS results (regression line equation estimate in the caption of Figure 1).

EXCEL TIPS: To replicate the regression equation, right click anywhere on the trendline in your scatter plot and select Format Trendline. As illustrated below, check the boxes for Display Equation on chart and Display R-squared value on chart.

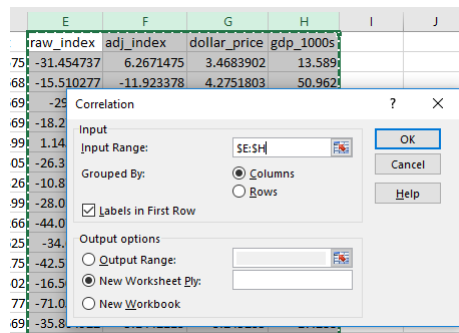


Interpretation tips: What does $\text{price-hat} = 2.4874 + 0.0394 \cdot \text{gdp}$ mean? The intercept has no interpretation because no country has zero GDP. The slope must be interpreted descriptively, not causally, because these are observational data and there are many lurking/unobserved/confounding/omitted variables related to both the price of a Big Mac in a country and its GDP per capita (e.g. level of the minimum wage). Among the 49 countries in The Economist's Big Mac database in January 2017, those with GDP per capita that is \$1,000 higher have Big Mac prices that are on average 4 cents higher.

- (d) **Create** a correlation matrix for the variables measuring the raw index, adjusted index, dollar price of a Big Mac, and GDP per capita. **Verify** with output below. **Note:** The units of GDP per capita (dollars or thousands of dollars) are irrelevant because the coefficient of correlation is unit-free.

	raw_index	adj_index	dollar_price	gdp_1000s
raw_index	1.0000			
adj_index	0.6244	1.0000		
dollar_price	1.0000	0.6244	1.0000	
gdp_1000s	0.7582	-0.0211	0.7582	1.0000

EXCEL TIPS: In the Data tab, click Data Analysis and select Correlation.



Interpretation tips: What does 0.7582 mean? Unsurprisingly, there is a pretty strong positive correlation between GDP per capita and the dollar price of a Big Mac. Countries with GDP per capita that is one standard deviation higher on average have Big Mac prices that are 0.76 standard deviations higher. (See pages 177-178 of our textbook.)

- (e) Estimate the regression in Figure 1 on page 39 again but this time also **compute the residual** for each observation. **Verify** that the residual for observation 1 is 0.445472719.

	A	B	C	D	E	F	G	H	I
	country	gdp_pc_usd_2015	local_price	dollar_ex	raw_index	adj_index	dollar_price	gdp_1000s	
1	Argentina	13589	55	15.8575	-31.454737	6.2671475	3.4683902	13.589	
2	Australia	50962	5.8000002	1.356668	-15.510277	-11.923378	4.2751803	50.962	
3	Austria	43724	3.4000001	0.95524669	-29.6583	-21.704809	3.5592902	43.724	
4	Belgium	40107	3.95	0.95524669	-18.279495	-5.8519607	4.1350574	40.107	
5	Brazil	8670	16.5	3.2239499	1.1451656	67.550941	5.1179457	8.67	
6	Britain	43771	3.0899999	0.82877505	-26.316324	-18.020756	3.728394	43.771	
7	Canada	43332	5.98	1.326	-10.873441	-0.430594	4.5098042	43.332	
8	Chile	13341	2450	672.80499	-28.034166	11.931901	3.6414711	13.341	
9	China	7990	19.6	6.9266	-44.077644	-6.4762373	2.8296711	7.99	
10	Colombia	6084	9900	2994.625	-34.66555	12.273956	3.3059232	6.084	
11	Czech Republic	17257	75	25.81175	-42.576019	-15.036997	2.9056535	17.257	
12	Denmark	52114	30	7.1012502	-16.509716	-13.835538	4.2246084	52.114	
13	Egypt	3740	27.49	18.77	-71.055908	-48.517914	1.4645711	3.74	
14	Estonia	17288	3.0999999	0.95524669	-35.864922	-5.1442113	3.245235	17.288	
15	Euro area	34142	3.8800001	0.95524669	-19.727707	-1.8496463	4.0617781	34.142	
16	Finland	41974	4.5500002	0.95524669	-5.8662548	6.5221009	4.7631679	41.974	
17	France	37675	4.0999999	0.95524669	-15.176186	0.080773718	4.7970847	37.675	

EXCEL TIPS (IMPORTANT!): We now switch to using the Data Analysis ToolPak to run regressions. This is a more powerful tool. Click Data Analysis in the Data tab and

select Regression. Select the range of each variable including the variable name. Check the labels box. Check the boxes for Residuals, Residual Plots, and Line Fit Plots.

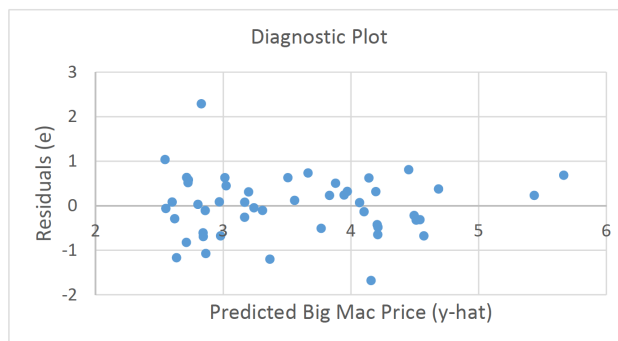
- (f) **Verify** that the mean of the residuals is zero. **Create a histogram** of the residuals. **Verify** that your histogram looks Normal is centered at zero and has a standard deviation of about \$0.68, which is $s_e = \sqrt{\frac{\sum_{i=1}^n e_i^2}{n-2}}$ (rounded from 0.684927656). Notice how the amount of scatter around the OLS line in Figure 1 matches the amount of scatter around zero in your histogram of the residuals.

EXCEL TIPS: For a quick histogram, remember to go to the Insert tab and select the histogram as discussed in part 1d on page 20 in Module A.2. (You can also use Histogram in Data Analysis, but it is more work and we just need a simple histogram of our residuals.)

Interpretation tips: What does 0.684927656 mean? The s_e of \$0.68 USD (i.e. 68 cents)⁷ measures the amount of scatter around the OLS line. Scatter is the variation in Big Mac prices that the regression line *cannot* explain with variation in GDP per capita. In the extreme case of a perfect fit (no scatter), the s_e would be zero. Is \$0.68 big or small? Big Mac prices are on average (across the 49 countries) \$3.55 with a standard deviation of \$1.04. Hence, \$0.68 is large. While GDP per capita helps predict Big Mac prices across countries, there is a lot of variation in Big Mac prices that is simply not explained by variation in GDP per capita. The s_e measures the importance of factors other than GDP per capita and it is quite large: we're missing some key variables to explain Big Mac prices.

- (g) Now **construct** a diagnostic scatter plot of the residuals against the predicted value of y .

EXCEL TIPS: Go to the new worksheet with the output created in part 1e. Under RESIDUAL OUTPUT, select the Predicted dollar_price and Residuals columns and Insert a Scatter plot. (If your worksheet is missing the RESIDUAL OUTPUT, repeat the previous part, this time remembering to check the boxes under Residuals as shown above.)



- (h) **Read** this part about data updates. *The Economist* updates the “The Big Mac Index” twice per year. The January 2021 version has the same structure as earlier versions. Beyond adding more recent data, it adds new countries. Make sure to familiarize yourself with the data set below.

- [big_mac_jan_2021.xlsx](#): The updated version of [big_mac_jan_2017.xlsx](#).

Test/exam examples: The Big Mac index has appeared quite a bit. Selections are below. Right now you are only ready for the October 2015 test questions.

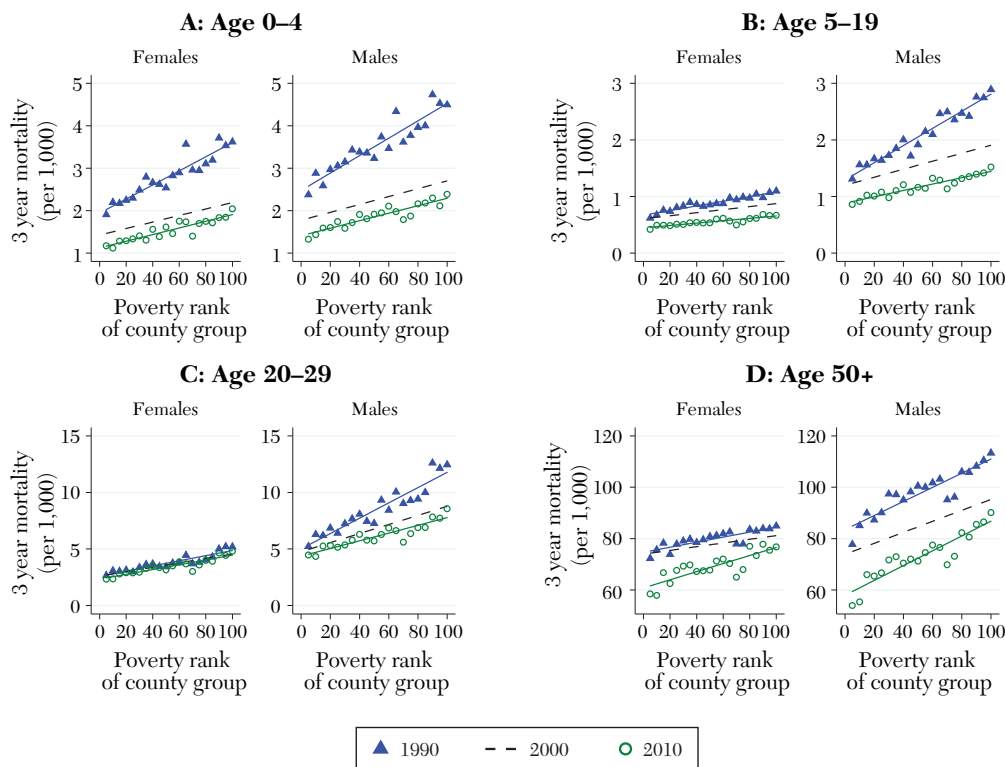
⁷Remember that the s_e is measured in the same units as the y variable. Because Big Mac prices (the y variable) are measured in US dollars, the s_e is also measured in US dollars.

- Question (2), [April 2017 Final Exam](#) (with [solutions](#))
- Questions (11) - (20), [October 2015 Test #1](#) (with [solutions](#))
- Question (2), [March 2016 Test #4](#) (with [solutions](#))

2. **Review** Figure 3 (this is what *The Economist* used to create its figure “Lower, flatter, better” on page 37). **Read** the note below the figure. **Note** the units of the x and y variables.

Figure 3

Three-Year Mortality Rates across Groups of Counties Ranked by their Poverty Rate



Source: Authors using data from the Vital Statistics, the US Census, and the American Community Survey.

Note: Three-year mortality rates for four different age groups are plotted across county groups ranked by their poverty rate. Mortality rates in 2000 and 2010 are age-adjusted using the 1990 population, that is, they account for changes in the age structure within age, gender, and county groups since 1990. Table A3 provides magnitudes for individual mortality estimates and for the slopes of the fitted lines.

Figure 3: Currie and Schwandt (2016), p. 41. **Note:** Panel C should say “Age 20-49,” not “Age 20-29.”

- (a) Using [mort_in_figure_3.table_a3.xlsx](#), **replicate** the regression line shown in Figure 3, Panel A for Females in 1990. First, **create** the y-variable, which is the adjusted number of deaths per 1,000 people of that sex, in that age group, and in that year for a county group. To create the y-variable (adjusted mortality per 1,000 people), use the variables adj_deaths and adj_population. (You will learn what adjusted means next. For now, just use the adjusted variables prepared for you.) For the x-variable, review the worksheet “readme”. **Verify** that you obtain: $\hat{y} = 1.940436 + 0.0166231x$.

EXCEL TIPS: After creating the y-variable, in the Data tab, click the Filter button. Select age_group and Uncheck the box for “(Select All)” and then check 0-4 yrs (see below). Repeat for the variables male (checking the box for the value 0) and year (checking the

box for the value 1990). Next, select the entire worksheet and copy to a new worksheet. Conveniently, it will copy only the filtered rows (i.e. just the variable names and 20 observations for females, aged 0-4 yrs, in 1990). Use Regression in Data Analysis, selecting the input variables from your new worksheet, and output the results to your new worksheet.

Interpretation tips: What does $\hat{y} = 1.940436 + 0.0166231x$ mean? The intercept says the mortality rate in 1990 is about 1.94 deaths per 1,000 female babies aged 0 to 4 years in the richest counties. The positive slope means that in 1990 as poverty increases the death rate rises. For every additional 10 percentile increase in poverty – for example, comparing counties at the 70th poverty percentile versus the 60th percentile – the mortality for female babies rises by an extra 0.17 deaths per 1,000. This is a considerable increase: going from the 60th to 70th poverty percentile, which means living in a somewhat poorer county, is associated with a 5.7% increase in female infant mortality ($5.7 \approx 100 \times \frac{3.104053 - 2.937822}{2.937822}$).

- (b) Recall that the preamble, starting on page 38, discussed the note below Figure 3. Use [mort.in.illustrate.composition.effects.xlsx](#), which already includes the numbers in bold-face, to *replicate Table B.1*, reproduced again below for convenience.

Table B.1: Illustrative county showing composition effects (aka Simpson's paradox)

Age group	Years	Deaths	Population	Mortality per 1,000	Adjusted pop.	Adj. deaths	Adj. mortality per 1,000
40-44	1990-93	300	50000	6.00	50000	300.00	6.00
40-44	2000-03	290	51000	5.69	50000	284.31	5.69
40-44	2010-13	280	52000	5.38	50000	269.23	5.38
45-59	1990-93	4800	300000	16.00	300000	4800.00	16.00
45-59	2000-03	6300	400000	15.75	300000	4725.00	15.75
45-59	2010-13	7800	500000	15.60	300000	4680.00	15.60
40-59	1990-93	5100	350000	14.57	350000	5100.00	14.57
40-59	2000-03	6590	451000	14.61	350000	5009.31	14.31
40-59	2010-13	8080	552000	14.64	350000	4949.23	14.14

Test/exam examples: An example of a Simpson's Paradox question:

- Question (2), [October 2019 Test #1](#) (with [solutions](#))

Test/exam examples: Currie and Schwandt (2016) has appeared a lot. You are ready for the November 2018 test. Save the rest until after Module E.3.

- Question (4), [April 2023 Final Exam](#) (with [solutions](#))
- Question (7), [April 2019 Final Exam](#) (with [solutions](#))
- Question (3), [November 2018 Test #2](#) (with [solutions](#))
- Questions (2) and (3), [April 2018 Test #5](#) (with [solutions](#))
- Question (7), [April 2017 Final Exam](#) (with [solutions](#))

B.2 Module B.2: PWT & Asiaphoria (Part 1 of 2)

Concepts: Measures of the strength of a relationship: correlation, rank correlation, regression slope, and R^2 . Analyzing subsamples.

Case studies: We use a major database – Penn World Tables (PWT) – to replicate parts of an academic working paper “Asiaphoria Meets Regression to the Mean,” abbreviated Pritchett and Summers (2014). Pritchett and Summers (2014) use version 8.0 of the PWT. You will also encounter newer versions.

Required readings: Section 7.8, “Logarithms in Regression Analysis with Asiaphoria” (Quercus). For [Pritchett and Summers \(2014\)](#), read *only* the abstract and Sections 1, 2.1, 6 (pp. 1-11, 56-59). Start with the [NBER digest](#). Below, review Table 1, an excerpt describing it, and some background.

Table 1: Little persistence in cross-national growth rates across decades						
Period 1	Period 2	Correlation	Rank Correlation	Regression Coefficient	R-squared	N
Adjacent decades						
1950-60	1960-70	0.363	0.381	0.378	0.132	66
1960-70	1970-80	0.339	0.342	0.382	0.115	108
1970-80	1980-90	0.337	0.321	0.323	0.114	142
1980-90	1990-00	0.361	0.413	0.288	0.130	142
1990-00	2000-10	0.237	0.289	0.205	0.056	142
One decade apart						
1950-60	1970-80	0.079	0.192	0.095	0.006	66
1960-70	1980-90	0.279	0.312	0.306	0.078	108
1970-80	1990-00	0.214	0.214	0.163	0.046	142
1980-90	2000-10	0.206	0.137	0.143	0.043	142
Two decades apart						
1960-70	1990-2000	0.152	0.177	0.152	0.023	108
1970-80	2000-2010	-0.022	0.005	-0.015	0.001	142
Source: Author’s calculations with PWT8.0 data (Feenstra, Inklaar and Timmer (2013)).						

Figure of Table 1: Pritchett and Summers (2014), p. 9.

- “Table 1 presents four measures of persistence: the correlation, the rank correlation (which reduces the influence of outliers), the regression coefficient of current growth on lagged growth,

and the R^2 (which is of course the square of the correlation coefficient). We use the PWT8.0 data on real GDP and population to compute real GDPPC. We compute least-squares growth rates of natural log GDPPC for 10 and 20 year periods for all countries with sufficient data.” (pp. 7-8)

- Figure 2 below visually summarizes the plan for Modules B.2 and B.3. The unit of observation (each dot) is a country. Module B.2 runs regressions like in Table 1 – see the red line (“second-stage” regression) on the right below – given the growth rates. Module B.3 finds the growth rates (“first-stage” regressions): it makes the data for Module B.2 (i.e. makes the dots).

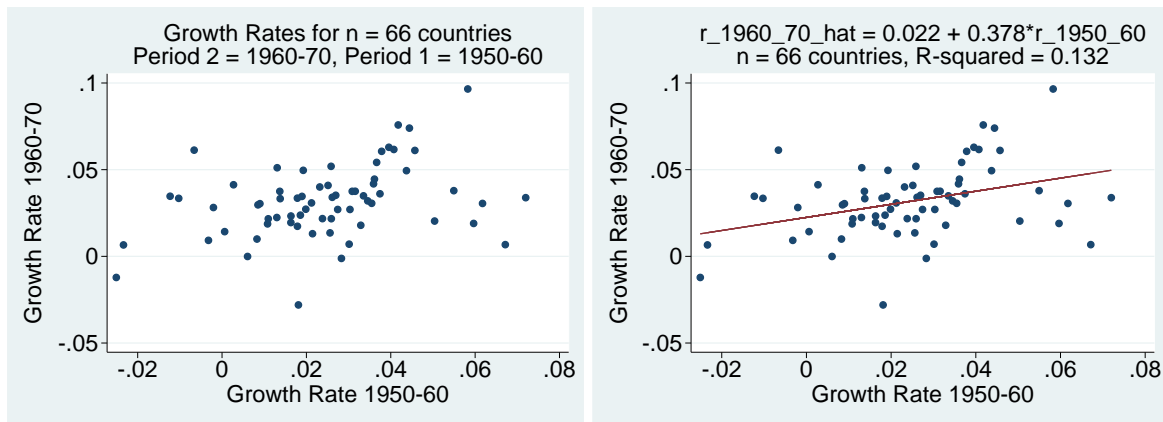
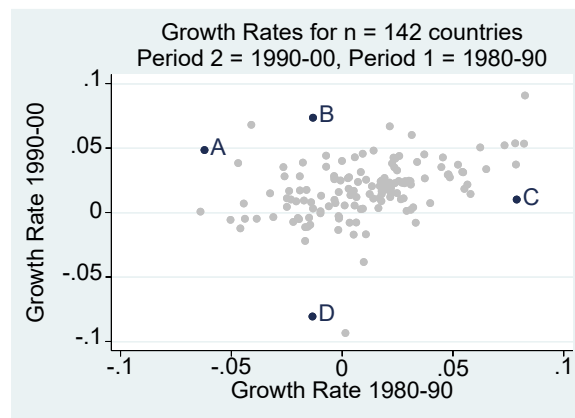


Figure 2: These scatter plots and OLS results visually illustrate the first row of results in Table 1.

Datasets: For Pritchett and Summers (2014): [asiap_rates_pwt.80.xlsx](#), where “asiap” abbreviates “Asiaphoria” from the title and “pwt.80” abbreviates Penn World Tables, version 8.0. Also, “_rates” means that these data contain the estimated growth rates.

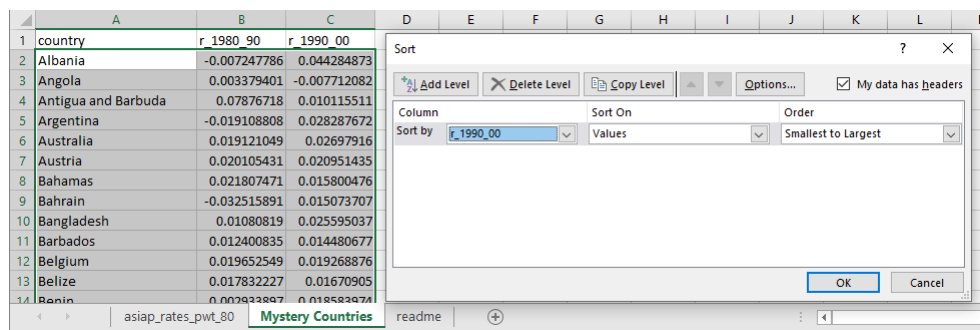
Interactive module materials for Module B.2:

1. **Browse** [asiap_rates_pwt.80.xlsx](#), which are the data used throughout this module. **Examine** the scatter diagram below. **Inspect** the data to find which countries are A, B, C, and D. **Verify** that you get: A=Qatar, B=Iraq, C=Antigua and Barbuda, D=Congo, Dem. Rep.



EXCEL TIPS: Copy the variables country, r_1980_90, and r_1990_00 to a new worksheet named “Mystery Countries.” Noting that B and D stand out for relatively extreme growth

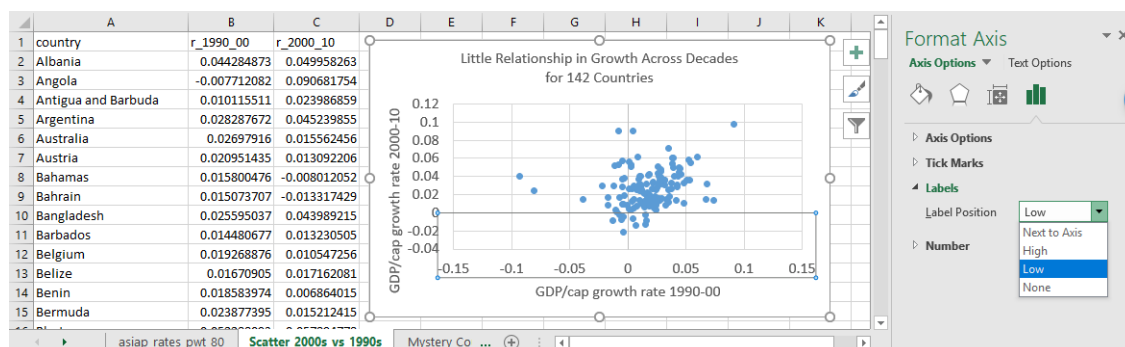
rates in the 1990s, sort the data by the growth rates in the 1990s. Make sure to select all three columns before sorting.



Similarly, to figure out A and C, sort by growth rates in the 1980s.

2. **Create** a scatter plot of the 2000-2010 growth rates (y axis) against the 1990-2000 growth rates (x axis). **Verify** that your graph shows no visible relationship between countries' growth rates in the 1990s and countries' growth rates in the 2000s. **Verify** that the *unit of observation* in your graph is a country (i.e. each dot corresponds to a different country).

EXCEL TIPS: Copy the variables country, r_1990_00, and r_2000_10 to a new worksheet named "Scatter_2000s_vs_1990s." Insert a Scatter Chart. To make it more readable, put the numeric axis labels around the edges (instead of in the middle of the scatter). For each axis, under Format Axis, Labels, select Low for the Label Position.



3. **Review** Table 1 on page 46. For convenience, here are the **numbers** that we replicate next.

Table 1: Little persistence in cross-national growth rates across decades						
Period 1	Period 2	Correlation	Rank Correlation	Regression Coefficient	R-squared	N
Adjacent decades						
1990-00	2000-10	0.237	0.289	0.205	0.056	142
One decade apart						
1970-80	1990-00	0.214	0.214	0.163	0.046	142

- (a) **Replicate** the results **0.205** (regression coefficient) and **0.056** (R-squared).

Interpretation tips: What do 0.205 and 0.056 mean? On average, countries with an additional one percentage point of growth in the 1990s only had an additional 0.2 percentage points of growth in the 2000s: just one-fifth carries over. The slope is positive

but fairly flat. Looking at the cross-section of 142 countries, fast growth in the 1990s is *not* a guarantee of fast growth in the 2000s. Similarly, slow growth in the 1990s is *not* a guarantee of slow growth in the 2000s. The very low value of the R-squared means that only 5.6 percent of the variation across countries in growth rates in the 2000s is explained by variation across countries in growth rates in the 1990s.

(b) **Replicate** the results **0.163** (regression coefficient) and **0.046** (R-squared).

(c) **Replicate** the result **0.237** (correlation).

EXCEL TIPS: As an alternative to using the tools under Data Analysis, Excel has some useful functions. For example, CORREL returns the coefficient of correlation. =CORREL(B:B,C:C) returns 0.2370 if Column B contains r_1990_00 and Column C contains r_2000_10. Note that =CORREL(B1:B143,C1:C143), which specifies the non-missing rows, and =CORREL(B2:B143,C2:C143), which excludes the variable name in the first row, also both return 0.2370. Similarly, for correlations the order of the variables does not matter: =CORREL(C:C,B:B) also returns 0.2370.

(d) **Replicate** the result **0.289** (rank correlation).

EXCEL TIPS: Use the RANK function. Create a new variable named rank_1990_00 with the function =RANK(I2,I:I) (where Column I has the growth rate in the 1990s) that you can copy and paste for the remaining 141 countries. Similarly, create a variable rank_2000_10. Use rank_1990_00 and rank_2000_10 to compute the correlation.

4. **Review** the results in the table below, which explores any differences between OECD member nations versus non-OECD member nations (i.e. two subsets of the original data).

Period 1	Period 2	Correlation	Rank Correlation	Regression Coefficient	R-squared	N
Adjacent decades: OECD member nations						
1990-00	2000-10	0.382	0.215	0.282	0.146	29
Adjacent decades: non-OECD member nations						
1990-00	2000-10	0.282	0.333	0.236	0.079	113

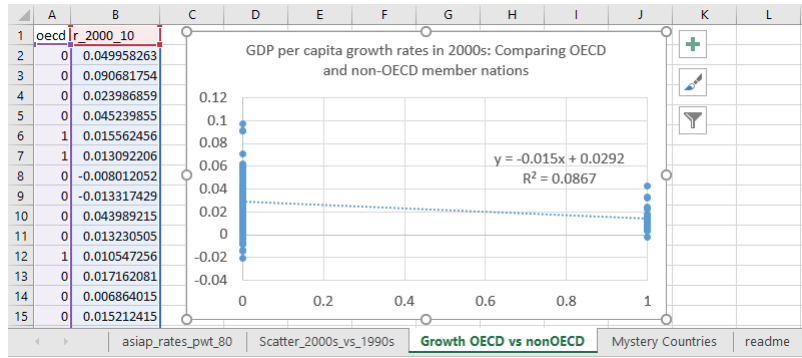
(a) **Replicate** the “Regression Coefficient” and “R-squared” results in the top panel.

EXCEL TIPS: There are several ways to work with a subset of data for a regression analysis. One is to select the entire worksheet and sort based on OECD status choosing descending to put OECD at the top so that you can include labels (for your regression for OECD countries only). Alternatively, you can filter the data and copy the filtered data to a new worksheet. The copying is necessary because filter merely hides irrelevant observations but does not move or delete them. Because regression inputs must be contiguous rows, you cannot run regression directly on the filtered data: you must work with a copy of it.

(b) **Replicate** “Regression Coefficient” and “R-squared” results in the bottom panel.

5. Next, **consider** a *new* question: how do growth rates in the 2000s compare between OECD member nations versus non-OECD member nations? **Run** an appropriate regression to answer.

EXCEL TIPS: Copy the variables oecd and r_2000_10 to a new worksheet titled “Growth OECD vs nonOECD.” Follow the steps starting on page 40 for parts 1b and 1c in Module B.1: insert a scatter chart, add a trend line, and show the equation and R-squared.



Interpretation tips: What do the OLS regression results of $\text{growth-hat} = 0.0292 - 0.015 \cdot \text{oecd}$ mean? For the cross-section of 142 countries with sufficient annual real GDP per capita data in the PWT 8.0 database between 2000 and 2010, on average real GDP per capita grew annually at 2.9% for non-OECD member nations. For OECD member nations, real GDP per capita grew much slower at only 1.4%, which is 1.5 percentage points slower than the non-OECD countries and corresponds to growth that is 51 percent slower. However, the scatter diagram shows that it is not only the mean but also the variance that differs between non-OECD and OECD member nations. There is far more scatter for non-OECD countries, with some experiencing notable contractions and other growing amazingly fast. This is not surprising as the OECD is an organization that gathers fairly similar developed nations as [members](#).

6. **Replicate** all of the “Correlation” and “Rank Correlation” results in part 4.

Note: Because Israel and the United Kingdom have the exact same growth rate in 2000-10 (0.0139203), you may obtain a slightly different rank correlation than 0.215. It depends on how these two identical values are ranked: if they are ranked equal at 12 it comes out to 0.218.

Period 1	Period 2	Correlation	Rank Correlation	Regression Coefficient	R-squared	N
Adjacent decades: OECD member nations						
1990-00	2000-10	0.382	0.215	0.282	0.146	29
Adjacent decades: non-OECD member nations						
1990-00	2000-10	0.282	0.333	0.236	0.079	113

EXCEL TIPS: Select each subset using the filter tool and copy that subset to a new worksheet (e.g. “OECD Countries”). Also, put the outputs of your analyses in the associated worksheet.

Test/exam examples: For many examples of Pritchett and Summers (2014), see page 56.

B.3 Module B.3: PWT & Asiaphoria (Part 2 of 2)

Concepts: Natural logarithms in regression. Using regression analysis to estimate GDP per capita *growth rates* (using panel data measuring annual GDP levels and population for 142 countries).

Case studies: Continue with Pritchett and Summers (2014).

Required readings: Recall readings for Module B.2, including Table 1 on page 46 and the excerpt. In Module B.3, we run twelve simple regressions, where the y variable is the natural log of GDP per capita and the x variable is year. These estimate the growth rate (the “slope” coefficient) in each of six decades for two countries: Canada and China. Table B.2 previews the results. The r_ in the variable names abbreviates rate: for example, r_1970_80 records the GDP growth rate in the 1970s.

Table B.2: Growth rate estimates for Canada and China (boldface numbers appear in Figure 2)

country	countrycode	r_1950_60	r_1960_70	r_1970_80	r_1980_90	r_1990_00	r_2000_10
Canada	CAN	0.01896	0.03459	0.02813	0.01946	0.02056	0.00916
China	CHN	0.05040	0.02037	0.03870	0.08251	0.09092	0.09735

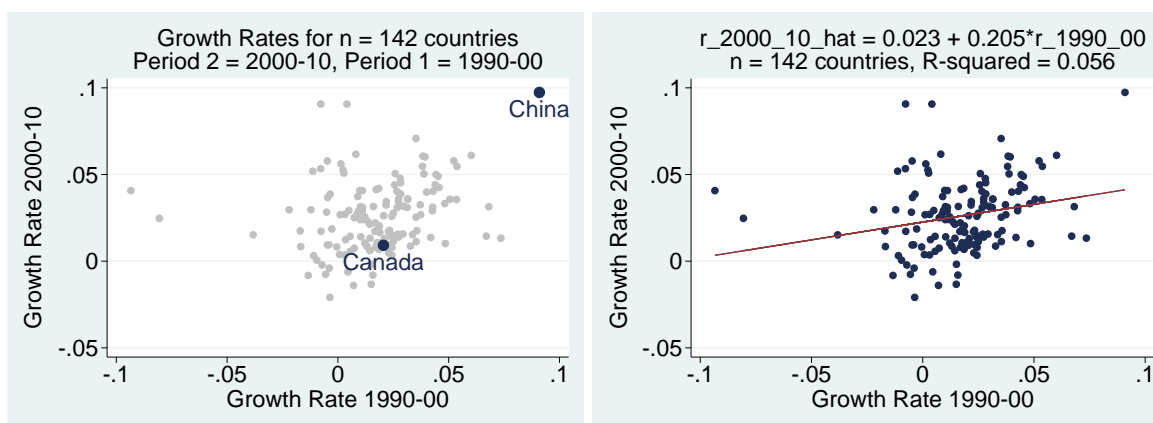


Figure 2: These scatter plots and OLS results visually summarize Modules B.2 and B.3.

- Module B.2 shows how to run regressions like in Table 1: the red line in the right scatter diagram of Figure 2 illustrates such a “second-stage” regression.
- Module B.3 shows how to find each dot in the scatter plots (“first-stage” regressions): together we do this for Canada and China, which are highlighted in the left scatter diagram of Figure 2. This is the *first* step to replicate Table 1 on page 46. We use regression to estimate GDP per capita growth rates for different decades and countries. Today we complete the first step for Canada and China and construct the data outlined in Table B.3.

Table B.3: Part of a data that you construct in Module B.3 (blanks for you to fill in)

country	countrycode	r_1950_60	r_1960_70	r_1970_80	r_1980_90	r_1990_00	r_2000_10
Canada	CAN						
China	CHN						

- Module B.2 replicated Table 1 using all available countries. In other words, in Module B.2 the data provided had far more than two rows (which is all that Table B.3 has).

- Simple regressions, where the y-variable is the natural log of GDP per capita, provide the growth rates for Table B.3. *Why* do Summers and Pritchett (2014) use the natural log instead of just GDP per capita (without the natural log transformation)? There are TWO distinct reasons: (1) growth *rates* can be directly compared across countries (richer versus poorer) and across time (when a country was poorer versus richer) whereas growth *levels* cannot and (2) GDP per capita may have been increasing nonlinearly. We illustrate these reasons interactively.

Datasets: For Pritchett and Summers (2014): [asiap_pwt_80_one_decade.xlsx](#). It includes only the PWT 8.0 data they use in their analysis⁸ leading to Table 1 on page 46. Also, “one_decade” says these are the data they use to estimate annual growth rates each one decade (e.g. 1990s: 1990-2000).

Interactive module materials for Module B.3:

1. To start work on filling in Table B.3, open [asiap_pwt_80_one_decade.xlsx](#).
 - (a) **Create** new variables measuring real GDP per capita (name it `rgdpna_pc`) and the natural log of real GDP per capita (name it `ln_rgdpna_pc`).
 - (b) Starting with Canada, **copy and paste** all observations for Canada into a new worksheet.

EXCEL TIPS: In the Data tab, click the Filter button. Uncheck the box for “(Select All)” and then check Canada. Next, select the entire worksheet and copy to a new worksheet. Conveniently, it will only copy the filtered rows (i.e. just Canada).

File

Home

Insert

Page Layout

Formulas

Data

Review

View

Tell me what you want to do...

From Access

From Web

From Text

From Other Sources

Existing Connections

New Query

Recent Sources

Show Queries

From Table

Refresh All

Connections

Properties

Edit Links

Sort

Filter

Get External Data

Get & Transform

Connections

Sort & Filter

A1

country

	A	B	C	D	E	F	G	H
1	country	countrycode	year	rgdpna	pop	rgdpna_pc	ln_rgdpna	
	Sort A to Z		1970	6819.207269	2.135599	3193.112222	8.068751338	
	Sort Z to A		1971	7092.031262	2.18865	3240.367927	8.08344216	
	Sort by Color		1972	7376.698965	2.241623	3290.784831	8.098881366	
	Clear Filter From "country"		1973	7673.0881	2.294578	3344.008397	8.114925485	
	Filter by Color		1974	7977.759137	2.347607	3398.251555	8.13101633	
	Text Filters		1975	8301.114186	2.400801	3457.643589	8.148342592	
	Search		1976	8635.641987	2.454255	3518.640886	8.165830082	
	<input type="checkbox"/> Brazil		1977	8966.309723	2.508026	3575.04656	8.181733479	
	<input type="checkbox"/> Brunei		1978	9352.011809	2.562121	3650.105443	8.202511335	
	<input type="checkbox"/> Bulgaria		1979	9722.360922	2.61653	3715.746016	8.220334749	
	<input type="checkbox"/> Burkina Faso		1980	10036.51942	2.6713	3757.166705	8.231420416	
	<input type="checkbox"/> Burundi		1981	10613.17583	2.725029	3894.701977	8.267372441	
	<input type="checkbox"/> Cambodia		1982	10926.12154	2.777592	3933.666836	8.277327307	
	<input type="checkbox"/> Cameroon		1983	11046.84843	2.831682	3901.161371	8.269029575	
	<input checked="" type="checkbox"/> Canada		1984	10908.58579	2.891004	3773.2863	8.235701598	
	<input type="checkbox"/> Cape Verde		1985	11102.82795	2.95739	3754.265736	8.230648002	
	<input type="checkbox"/> Central African Republic		1986	11723.99819	3.033393	3864.978322	8.259711352	
			1987	11631.59827	3.116009	3732.851308	8.224927646	
			1988	11466.35754	3.194854	3589.008307	8.185631206	
			1989	12594.71565	3.255859	3868.323429	8.26057647	
			1990	11338.88613	3.289483	3447.011621	8.145262938	

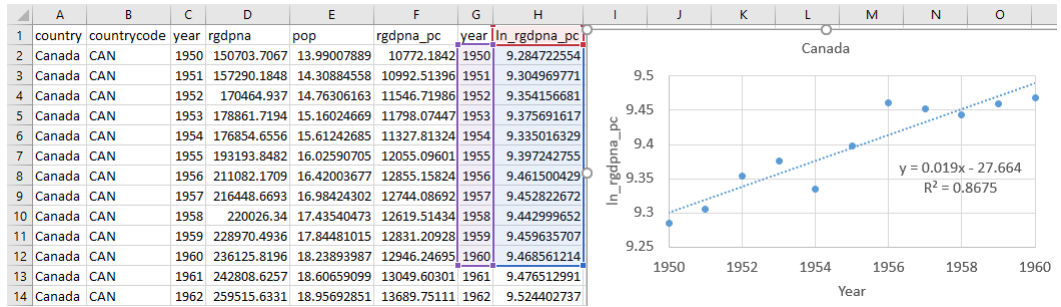
OK

Cancel

- i. **Create** a scatter plot of the natural log of real GDP per capita over the period 1950-60, *including* both endpoints: 1950 and 1960.

EXCEL TIPS: Selecting the variable names and the first 11 rows of data for year and `ln_rgdpna_pc` (see below), insert a Scatter Chart. (You can either insert a copy of the year column before `ln_rgdpna_pc` (shown below) or you can use Ctrl to select two non-adjacent columns.) Recall from parts 1b and 1c of Module B.1 on page 40 that you can add a trendline and the OLS equation and R^2 .

⁸It excludes observations with missing data, labeled as outliers, or in decades with insufficient observations.

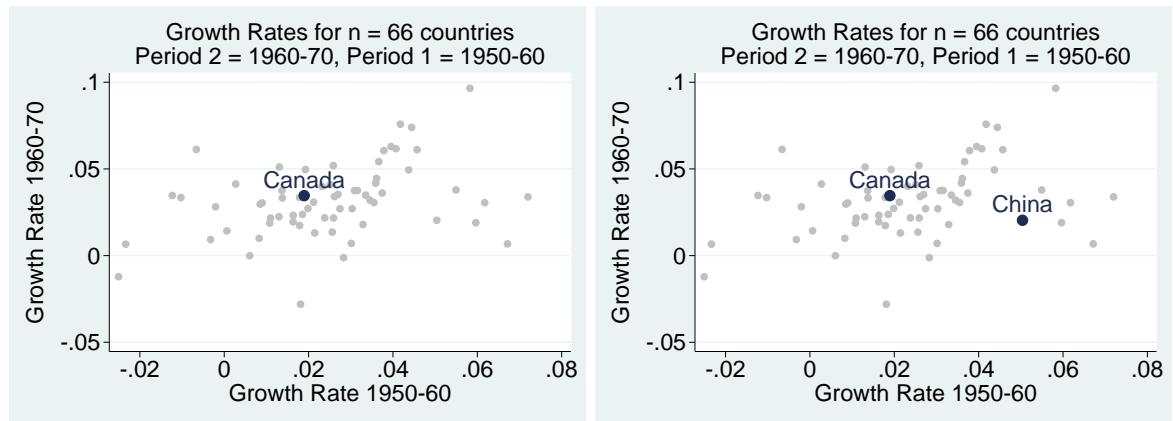


- ii. For the same years (1950-1960), **run a simple regression** of the natural log of real GDP per capita (y variable) on year (x variable). Using the worksheet “Your results Canada and China,” **copy** the “slope” coefficient into the cell for Canada for 1950-1960: r_1950_60, which is the growth rate. **Verify** that you obtain 0.0190: an average GDP per capita growth of 1.9% annually in the 1950s in Canada.

EXCEL TIPS: Use Regression in Data Analysis under the Data tab (exact results).

- iii. For 1960-1970 (including both endpoints), **run a simple regression** of the natural log of real GDP per capita on year. **Copy** the “slope” coefficient into the cell for Canada for 1960-1970: r_1960_70, which is the growth rate. **Verify** that you obtain 0.0346: an average GDP per capita growth of 3.5% annually the 1960s in Canada.

Note: Putting parts 1(b)ii and 1(b)iii together, you have now found *one dot* in the scatter plot in Figure 2 on page 47, which is illustrated below (left).



- (c) Move on to China. **Copy and paste** all observations for China into a new worksheet.

- i. For 1950-1960, **run a simple regression** of the natural log of real GDP per capita (y variable) on year (x variable). Because GDP and population data are only available in China starting in 1952, your regression will have only 9 observations. **Copy** the “slope” coefficient into the cell for China for 1950-1960: r_1950_60, which is the growth rate. **Verify** that you obtain 0.0504: an average GDP per capita growth of 5.0% annually in the 1950s in China.
- ii. **Repeat** part 1(c)i but for 1960-1970. **Verify** that you obtain 0.0204.

Note: Putting parts 1(c)i and 1(c)ii together, you have now found *one more dot* in the scatter plot in Figure 2 on page 47, which is illustrated above (right).

- (d) We have not yet considered the other 64 countries (the light grey dots above) and we are not even done with Canada and China. **Recall** that Table 1 on page 46 also considers

growth rates in other decades. **Note** the blank cells still to be filled.

What we have filled in so far in Table B.3 on page 51

country	countrycode	r_1950_60	r_1960_70	r_1970_80	r_1980_90	r_1990_00	r_2000_10
Canada	CAN	0.0189558	0.0345877				
China	CHN	0.0503967	0.020365				

- i. **Obtain** the values of r_1970_80, r_1980_90, r_1990_00, and r_2000_10, for China. **Verify** that you obtain 0.0387, 0.0825, 0.0909, and 0.0974, respectively.

EXCEL TIPS: The function LINEST is a shortcut. It returns the OLS slope given a range of values for y and the corresponding range of values for x. See the screenshot next: it returns 0.0387. (Similarly, the function RSQ returns the R^2 .)

The screenshot shows an Excel spreadsheet with a data table for China. The formula bar displays the formula `=LINEST(G20:G30,C20:C30)`. The data table has columns for country, countrycode, year, rgdpna, pop, rgdpna_pc, and ln_rgdpna_pc. The data for China is as follows:

country	countrycode	year	rgdpna	pop	rgdpna_pc	ln_rgdpna_pc
China	CHN	1969	317702.5124	778.7267134	407.9769024	6.011210561
China	CHN	1970	354961.2279	799.946841	443.7310203	6.095218569
China	CHN	1971	379808.5139	821.436505	462.3711164	6.136367851
China	CHN	1972	394241.2374	842.515035	467.933771	6.148326771
China	CHN	1973	425386.2952	862.740261	493.064152	6.200639291
China	CHN	1974	435170.18	881.626929	493.5990107	6.201723468
China	CHN	1975	473029.9856	898.891252	526.2371667	6.265751998
China	CHN	1976	465461.5058	914.236509	509.1259223	6.232695378
China	CHN	1977	500836.5803	927.91348	539.7449127	6.291096644
China	CHN	1978	559434.4602	940.448391	594.8592879	6.388324887
China	CHN	1979	601799.6142	952.699898	631.6780504	6.44837985
China	CHN	1980	648989.3819	965.365571	672.2731797	6.510664775

- ii. **Obtain** the values of r_1970_80, r_1980_90, r_1990_00, and r_2000_10 for Canada. **Verify** that you obtain 0.0281, 0.0195, 0.0206, 0.0092, respectively.

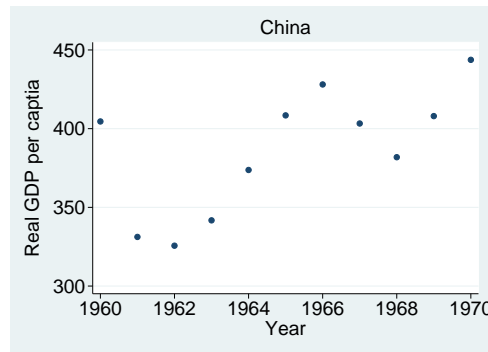
(e) Consider repeating all these steps for each country (not just Canada and China)! In Module B.2 you were simply given the results for the remaining 140 countries. To see that, browse [asiap_rates_pwt_80.xlsx](#). It includes all available countries, not just Canada and China. Verify that 142 simple regressions are necessary to obtain the values of the variable r_1970_80. Verify that a total of 742 simple regressions (like those you ran in B.3) are run to obtain all the data shown in [asiap_rates_pwt_80.xlsx](#). Of course, for efficiency and accuracy, these data are generated by computer code and not interactively, as we did for Canada and China. However, there is an important pedagogical value to doing it yourself interactively (at least for a couple of countries).

2. *Why* did Summers and Pritchett (2014) use the natural log of GDP per capita instead of just GDP per capita (without the log transformation)? There are TWO distinct reasons: (1) growth *rates* can be directly compared across countries (richer versus poorer) and across time (when a country was poorer versus richer) whereas growth *levels* cannot and (2) GDP per capita may have been increasing nonlinearly. Next, we illustrate these reasons interactively.

- (a) To illustrate the first reason for logging GDP per capita (meaningful comparisons of growth across richer and poorer countries), compare growth in Canada and China in the 1960s:
 - i. **Create** a scatter plot of real GDP per capita over the period 1960-1970 for Canada. **Note** that the relationship looks linear.
 - ii. **Run a simple regression** of real GDP per capita on year over the period 1960-1970 for Canada. **Verify** that the slope coefficient you obtain is 526.9533, which corresponds to an average GDP per capita growth of \$527 USD (in constant 2005

dollars) annually in the 1960s in Canada. This is an estimate of the *level* of growth each year.

- iii. **Create** a scatter plot of real GDP per capita over the period 1960-1970 for China. **Verify** that it looks similar to graph below (created by Stata).



Note: You can describe this relationship as *linear*. With only 11 data points equally spaced (annual data), it is easy for “patterns” to appear by chance. It is tempting to see a “W” in the diagram, but there is not sufficient evidence to rule out the most simple explanation: linear with noise. “When you hear hoofbeats, think of horses not zebras” [https://en.wikipedia.org/wiki/Zebra_\(medicine\)](https://en.wikipedia.org/wiki/Zebra_(medicine)).

- iv. **Run a simple regression** of real GDP per capita on year over the period 1960-1970 for China. **Verify** that the slope coefficient you obtain is 7.7167, which corresponds to an average GDP per capita growth of \$8 USD annually (in constant 2005 dollars) in the 1960s in China. This is an estimate of the *level* of growth each year.

Interpretation tips: How do we compare \$8 and \$527? Compared to \$527, \$8 looks tiny, but, remember that China was much poorer per capita than Canada in the 1960s. Using our data we can see exactly how much poorer: in 1965 (the midpoint year in the 1960s) GDP per capita was \$408 USD (in constant 2005 dollars) in China versus \$15,451 USD (in constant 2005 dollars) in Canada. Earlier we found that, in the 1960s, GDP per capita grew at an average annual *rate* of 2.0% in China and 3.5% in Canada. Given that we are comparing a rich and poor country, it makes sense to compare growth *rates*, not growth *levels*. Note that even though we obtained those rate estimates using a regression with a log transformation of GDP per capita, we can get a rough idea of the rates simply by looking at $100 \times 527 / 15,451 = 3.4\%$ (which is very close to 3.5%) for Canada and by looking at $100 \times 8 / 408 = 2.0\%$ (which, rounded, is the same as 2.0%) for China.

- (b) Consider China’s growth in the 2000s to illustrate the second reason for logging GDP per capita (addressing nonlinear growth):
 - i. **Create** a scatter plot of real GDP per capita over the period 2000-2010 for China. **Note** that the relationship is nonlinear: increasing at an increasing rate. If you are having trouble seeing the nonlinearity, **create** a scatter plot of real GDP per capita over the period 1970-2010 for China: the nonlinearity is more visually obvious over a longer time horizon.
 - ii. **Ignoring the nonlinearity, run a simple regression** of real GDP per capita on year over the period 2000-2010 for China. **Verify** that the slope coefficient you obtain

is 542.0147, which corresponds to an average GDP per capita growth of \$542 USD annually in the 2000s in China.

- Note how \$542 overstates growth in the early 2000s and understates growth in the late 2000s: for instance, from 2000 to 2001 GDP per capita actually increased by \$258 ($=\$3667.29 - \3409.736) whereas from 2009 to 2010 GDP per capita actually increased by \$776 ($=\$8727.472 - \7950.976). This is because a straight line systematically fails to fit a curved relationship.
 - iii. **Create** a scatter plot of the natural log of real GDP per capita over the period 2000-2010 for China. **Note** that the relationship now looks linear. **Recall** your results from earlier in this module where you found that GDP per capita grew at an average annual rate of 9.7% in China in the 2000s.
3. **Read** this part about data updates. The [Penn World Tables \(PWT\)](#) are periodically updated. Version 10.01 (published on January 23, 2023) has the same structure as earlier versions. Each new version of the PWT updates and refines previously reported data and sometimes adds more recent data and/or additional countries. In the data sets described below, we have also updated the variables that we add – the dummy for whether the country is a member of the OECD and the name of the continent – to reflect countries that have recently joined the OECD (including Latvia, Lithuania, and Colombia) and to match the countries as included in the PWT 10.0 data. Make sure to read the notes below and familiarize yourself with the data sets below.
- [asiap_pwt_100_all.xlsx](#): These data are an excerpt of the variables (columns) from the publicly posted PWT 10.0 data. Also, it adds the OECD dummy variable and the continent identifier variable. These data include all of the same observations (rows) as the publicly posted data *except* that [asiap_pwt_100_all.xlsx](#) *excludes* rows with missing values.
 - [asiap_rates_pwt_100.xlsx](#): These data include the same countries and use the same methods as described in Module B.2, which starts on page 46, for [asiap_rates_pwt_80.xlsx](#). Remember that those methods focus on subsets of the data and apply regression analysis to produce a new dataset. You can simply think of [asiap_rates_pwt_100.xlsx](#) as the updated version of [asiap_rates_pwt_80.xlsx](#): it continues to focus on the same (subset of) countries and the same years. The dataset [asiap_rates_pwt_100.xlsx](#) enables analyses to produce the results in Table 1 of Pritchett and Summers (2014) (see page 46) and simply has updated numbers for the same countries and the same years.

Test/exam examples: Pritchett and Summers (2014) has appeared many times. You are ready for these (except for Question (5) on the November 2019 test, which includes linear combinations of random variables from later in the first half of the course, and the March 2017 test, which includes multiple regression from the second half of the course).

- Question (7), [March 2025 Test #4](#) (with [solutions](#))
- Question (3), [October 2022 Test #1](#) (with [solutions](#))
- Question (7), [October 2021 Test #1](#) (with [solutions](#))
- Questions (1) and (5), [November 2019 Test #2](#) (with [solutions](#))
- Question (2), [Summer 2019 Test #2](#) (with [solutions](#))

- Question (8), [April 2019 Final Exam](#) (with [solutions](#))
- Questions (4) and (6), [November 2017 Test #2](#) (with [solutions](#))
- Question (4), [November 2017 Test #2](#) (with [solutions](#))
- Question (6), [March 2018 Test #4](#) (with [solutions](#))
- Question (2), [March 2017 Test #4](#) (with [solutions](#))
- Question (6), [November 2016 Test #1](#) (with [solutions](#))
- Question (2), [February 2017 Test #3](#) (with [solutions](#))
- Questions (32) - (35), [October 2015 Test #1](#) (with [solutions](#))
- Question (2), [December 2015 Test #2](#) (with [solutions](#))
- Question (3), [March 2016 Test #4](#) (with [solutions](#))
- Questions (20) - (28), [April 2016 Test #5](#) (with [solutions](#))

B.0.0 Practice questions for Module B

Q1. Following Currie and Schwandt (2016), fill in all of the missing values in the two tables below. The optional Excel tips help you efficiently cut-and-paste these tables into Excel. Alternatively, you can enter the values by hand. Whichever way, make sure to answer.

(a) Fill in all of the missing values in this table:

Age group	Year	Deaths	Population	Mortality per 1,000	Adjusted population	Adj. deaths	Adj. mortality per 1,000
0 yrs	1990	834.00	77718.00				
0 yrs	2000	729.00	89991.33				
0 yrs	2010	598.00	90797.00				
1-4 yrs	1990	289.00	365123.00				
1-4 yrs	2000	237.33	359350.33				
1-4 yrs	2010	203.00	366847.00				
0-4 yrs	1990						
0-4 yrs	2000						
0-4 yrs	2010						

(b) Fill in all of the missing values in this table:

Age group	Year	Deaths	Population	Mortality per 1,000	Adjusted population	Adj. deaths	Adj. mortality per 1,000
0 yrs	1990	900	78000				
0 yrs	1995	880	81000				
0 yrs	2000	760	90000				
0 yrs	2005	700	92000				
0 yrs	2010	600	96000				
0 yrs	2015	510	98000				
1-4 yrs	1990	300	400000				
1-4 yrs	1995	270	380000				
1-4 yrs	2000	220	370000				
1-4 yrs	2005	210	380000				
1-4 yrs	2010	200	360000				
1-4 yrs	2015	180	350000				
0-4 yrs	1990						
0-4 yrs	1995						
0-4 yrs	2000						
0-4 yrs	2005						
0-4 yrs	2010						
0-4 yrs	2015						

EXCEL TIPS (OPTIONAL): You can import a comma-delimited version of the tables above. Next, are comma-delimited versions to select and copy, with how-to tips.

Comma-delimited version for part 1a:

```

,,,,,,Adj.
Age,,,Mortality,Adjusted,Adj.,mortality

```

```

group,Year,Deaths,Population,"per 1,000",population,deaths,"per 1,000"
0 yrs,1990,834.00,77718.00,,,
0 yrs,2000,729.00,89991.33,,,
0 yrs,2010,598.00,90797.00,,,
1-4 yrs,1990,289.00,365123.00,,,
1-4 yrs,2000,237.33,359350.33,,,
1-4 yrs,2010,203.00,366847.00,,,
0-4 yrs,1990,,,,,
0-4 yrs,2000,,,,,
0-4 yrs,2010,,,,,

```

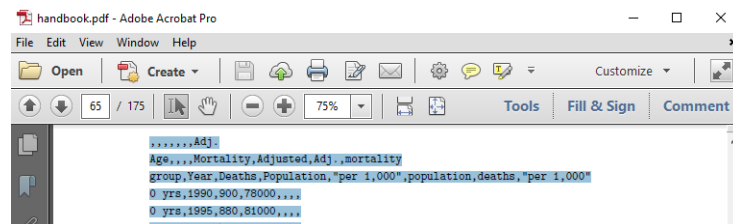
Comma-delimited version for part 1b:

```

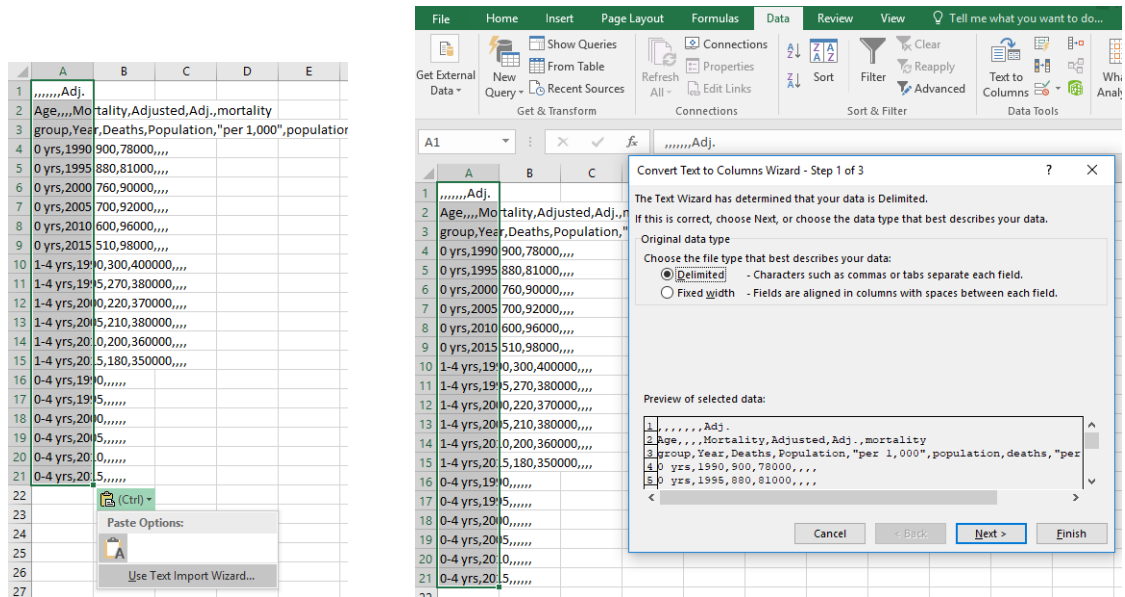
,,,,,,Adj.
Age,,,Mortality,Adjusted,Adj.,mortality
group,Year,Deaths,Population,"per 1,000",population,deaths,"per 1,000"
0 yrs,1990,900,78000,,,
0 yrs,1995,880,81000,,,
0 yrs,2000,760,90000,,,
0 yrs,2005,700,92000,,,
0 yrs,2010,600,96000,,,
0 yrs,2015,510,98000,,,
1-4 yrs,1990,300,400000,,,
1-4 yrs,1995,270,380000,,,
1-4 yrs,2000,220,370000,,,
1-4 yrs,2005,210,380000,,,
1-4 yrs,2010,200,360000,,,
1-4 yrs,2015,180,350000,,,
0-4 yrs,1990,,,,,
0-4 yrs,1995,,,,,
0-4 yrs,2000,,,,,
0-4 yrs,2005,,,,,
0-4 yrs,2010,,,,,
0-4 yrs,2015,,,,,

```

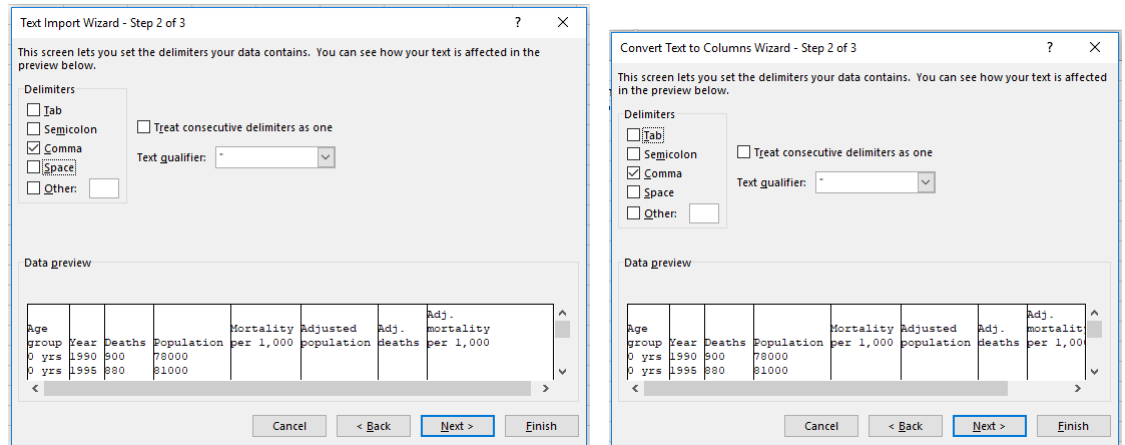
In Adobe, use the Selection tool, which is the arrow icon, to highlight the comma-delimited text. In a PC, use **Ctrl+c** to copy. In a mac, use **Command+c**.



Into a new Excel workbook, click the first cell then paste (**Ctrl+v** or **Command+v**). Next, either select “Use Text Import Wizard...” from the drop-down menu (which appears with the paste if you have a PC) OR select the Text to Columns button under the Data tab (available for both PCs and Macs). Check Delimited then click the Next > button. See screenshots below.



Check Comma under Delimiters (other boxes must be unchecked). Click the Finish button.



The Text to Columns tool (and Text Import Wizard) are flexible and useful: we did an easy case. Refining this skill can enable you to quickly pull data from pdf files and other sources without typing by hand, which can lead to errors (in addition to taking time). An extra trick, which you should *not* need for the simple import above, but is often useful when copying from non-spreadsheet-style sources (e.g. pdf document, web page, etc.), is to paste into a plain text editor first and then copy it again from the plain text editor into Excel. In a PC, use Notepad (the plain-old Notepad that comes standard on all PCs, not + or ++). In a Mac, use the TextEdit app BUT before pasting into it, under the Format tab, select Make Plain Text. This simple trick often makes it possible to cut-and-paste into Excel even when a direct paste into Excel does not work (e.g. puts the entire paste into a single cell, ignoring the line breaks).

Q2. Recall the Big Mac index and Figure 1 on page 39, which shows the scatter diagram and OLS line for the January 2017 analysis. Use [big_mac_jan.2017.xlsx](#) for all subparts.

- (a) Run a regression like Figure 1 *but* measuring GDP per capita in US dollars, not \$1,000s of US dollars. Compare and contrast the intercept, slope, R-squared, SST, SSR, SSE, and s_e

(standard deviation of the residuals) between the regression results when GDP per capita is measured in US dollars versus \$1,000s of US dollars.

- (b) Identify the country with the biggest positive residual (which means that the price of a Big Mac in that country is the furthest above what you may expect given that country's GDP per capita). Is this the country with the highest priced Big Mac?
- (c) Run a regression like Figure 1 on page 39 *but* excluding the one observation for the “Euro area.” What are the values of the OLS intercept and slope? What is the predicted price of a Big Mac for a country with GDP per capita of \$27,341?
- (d) Compute the correlation between the raw index and the adjusted index excluding Switzerland. Compare and contrast the correlation with and without Switzerland.

Q3. Recall Pritchett and Summers (2014) and use [asiap_pwt_80_one_decade.xlsx](#) for all subparts, which focus on obtaining estimates of real GDP per capita growth rates.

- (a) Using appropriate regression analyses, compute the growth rate of real GDP per capita for the periods 1985-1990, 1990-1995, 1995-2000, 2000-2005, 2005-2010 for Brazil. (Note that is just like what Pritchett and Summers (2014) did for each decade and each two-decade period; you are being asked to consider a half-decade period.)
- (b) Identify any and all half-decade periods between 1985 and 2010 where the annual growth rate of real GDP per capita is above 5% for Brazil.

Q4. Recall Pritchett and Summers (2014) and use [asiap_pwt_80_one_decade.xlsx](#) for all subparts.

- (a) Compute the correlation between population and real GDP for Bangladesh for 1960-2010.
- (b) Using an appropriate regression, estimate the population growth rate for Bangladesh between 1960-2010. Answer by filling in the blank: For Bangladesh between 1960-2010, on average the population increased by _____ percent annually.

Q5. Recall Table 1 on page 46 from Pritchett and Summers (2014). Use [asiap_rates_pwt_80.xlsx](#). Replicate the results for the row for “Period 1” equal to 1970-80 and “Period 2” equal to 2000-10. In addition to the results reported in Table 1 (correlation, rank correlation, regression coefficient, and R-squared), also report the intercept, SST, SSR and SSE.

Q6. Recall Pritchett and Summers (2014) and use [asiap_rates_pwt_80.xlsx](#) for all subparts, which focus on looking at the subset of African countries in the context of Table 1 on page 46.

- (a) Regress the growth rates for 2000-2010 on the growth rates for 1990-2000 for the subset of countries in Africa. Report the regression coefficient and R^2 .
- (b) Compute the correlation and rank correlation between the 2000-2010 growth rates and the 1990-2000 growth rates for the subset of countries from Africa. Does a comparison of the correlation and rank correlation suggest any outliers among the African countries?
- (c) How many regressions were run to obtain the 1990-2000 and 2000-2010 growth rates for the African countries used in these analyses?

Q7. Recall Pritchett and Summers (2014) and use [asiap_rates_pwt_80.xlsx](#) for all subparts, which focus on understanding the R-squared values reported in Table 1 on page 46. In particular, recall

the result **0.056** in the column labeled “R-squared” in the panel labeled “Adjacent decades,” and in the row for “Period 1” equal to 1990-00 and “Period 2” equal to 2000-10, which you already replicated in Module B.3. That low R-squared means that only 5.6 percent of the variation across countries in growth rates in the 2000s is explained by variation across countries in growth rates in the 1990s.

- (a) How much variation is there across countries in growth rates in the 2000s? When answering, measure growth as a percent. Answer visually by constructing a histogram. Answer with statistics by computing the standard deviation, range, and coefficient of variation. Describe the shape of the histogram and comment on the size of the standard deviation.
- (b) How much variation is there across countries in growth rates in the 1990s? When answering, measure growth as a percent. Answer visually by constructing a histogram. Answer with statistics by computing the standard deviation, range, and coefficient of variation. Describe the shape of the histogram and comment on the size of the standard deviation.

Q8. Recall Pritchett and Summers (2014) and use [asiap_pwt_90_all.xlsx](#) for all subparts, which contains the PWT Version 9.0 data.

- (a) Consider ten-year periods that allow using the most recent 2014 data. Using appropriate methods, complete the data set below. The variables r_1994_04 and r_2004_14 record the growth rate in real GDP per capita from 1994-2004 and 2004-2014, respectively. (As usual, include the endpoints.)

country	countrycode	r_1994_04	r_2004_14
Canada	CAN		
Mexico	MEX		
United States	USA		

- (b) Using the data set you created above and appropriate methods, complete the table of results below.

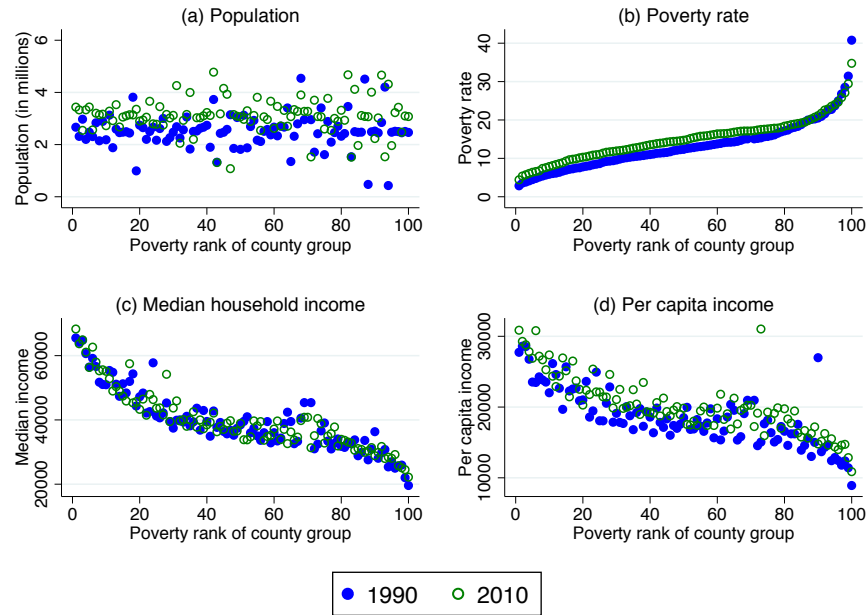
Period 1	Period 2	Correlation	Rank Correlation	Regression Coefficient	R-squared	N
Adjacent decades: Three countries (Canada, Mexico and United States)						
1994-04	2004-14					

- (c) Is there anything unusual about the results compared to Table 1 on page 46 in Pritchett and Summers (2014)?

For extra practice, additional questions, with an ^e superscript (*e* for extra), are next.

Q^e1. Recall Currie and Schwandt (2016). Review Figure A2, including reading the note below it and noting the units of measurement of the x and y variables. Use [mort.in.figure.a2.xlsx](#) for all subparts, which provides the data underlying Figure A2.

Figure A2: County group characteristics



Notes: Median and per capita income are adjusted for inflation and reported in constant 1999 dollars. Median income refers to counties' median income averaged across counties in each county group, weighted by counties' population size. The outliers in panel (d) are driven by New York County, NY, a big county with both a high poverty rate and high per capita income.

Figure A2: Currie and Schwandt (2016), p. 3 of the Appendix.

- Replicate the scatter diagram in Panel (c) in Figure A2 for the year 1990 only.
- Discuss/consider the direction of each relationship (positive, negative, or no relationship) in Figure A2, the type (linear or nonlinear), and the strength; Make sure to consider why, except for Panel (a), we should use the word *association* and not *correlation*.
- Create a scatter plot of median income versus per capita income using only the year 2010 data. How is that relationship best described?
- Regress population in millions on the poverty rank of the county group using only the year 1990 data. What are the values of the SST, SSR, SSE and the R-squared? (Note: Make sure to create a variable to measure population *in millions*, which is what the authors also do in Panel (a) of Figure A2.)

Q^e2. Recall Pritchett and Summers (2014) and use [asiap_pwt.80_one_decade.xlsx](#) for all subparts, which focus on obtaining estimates of real GDP per capita growth rates.

- Using appropriate regression analyses, compute the growth rate of real GDP per capita for the periods 1985-1990, 1990-1995, 1995-2000, 2000-2005, 2005-2010 for Argentina. (Note that is just like what Pritchett and Summers (2014) did for each decade and each two-decade period; you are being asked to consider a half-decade period.)
- Identify any and all half-decade periods between 1985 and 2010 where the annual growth rate of real GDP per capita is above 5% for Argentina.

Q^e3. Recall Pritchett and Summers (2014) and use [asiap_pwt.80_two_decade.xlsx](#) for all subparts, which focus on building the data needed to replicate Table 2.

Table 2: Twenty year periods show modest persistence: hence current growth has little value for predicting future growth						
Period 1	Period 2	Correlation	Rank correlation	Regression Coefficient	R-squared	N
Adjacent two decade periods						
1950-70	1970-1990	0.258	0.318	0.343	0.067	70
1960-80	1980-2000	0.459	0.454	0.494	0.211	108
1970-90	1990-2010	0.327	0.325	0.215	0.107	142
Gap of two decades						
1950-70	1990-2010	0.047	0.015	0.047	0.002	70
Source: Author's calculations with PWT8.0 data (Feenstra, Inklaar and Timmer (2013)).						

Figure of Table 2: Pritchett and Summers (2014), p. 10.

- Review Table 2. In the third row under the panel heading “Adjacent two decade period,” what is the y variable and what is the x variable that correspond to the results reported in the column “Regression Coefficient”?
- In that same row, what does 142 mean?
- For that same row, what is the value of the x variable for Canada? What is the value of the y variable for Canada?
- For the first row under the panel heading “Gap of two decades,” what is the value of the y variable for Canada?

Q^e4. Recall Table 1 on page 46 from Pritchett and Summers (2014). It has three panels: “Adjacent decades,” “One decade apart,” and “Two decades apart.” Use [asiap_rates_pwt.80.xlsx](#).

- Under the panel “Adjacent decades,” why is the sample size only 66?
- Under the panel “Two decades apart,” the authors could report a row of results for “Period 1” equal to 1950-60 and “Period 2” equal to 1980-90. (They chose not to.) What would be the sample size for that row?

Q^e5. Each year, on the website (www.fueleconomy.gov), the U.S. Department of Energy releases a guide and raw data to inform consumers about the fuel economy and greenhouse gas emissions of new vehicles (cars, vans, etc.). Consider the 2017 data on 1,230 makes, models and configurations (e.g. four-door Honda Civic with automatic transmission). These data include

variables measuring the type of engine, fuel efficiency and greenhouse gas emissions. Use [fuel_economy_2017.xlsx](#) and **only the 82 observations for Nissan** (the name of a manufacturer) for all subparts.

- (a) Describe the relationship between CO2 (carbon dioxide) emissions in city driving with the fuel efficiency in city driving.
- (b) Ignoring what you noticed in the previous part, regress CO2 emissions in city driving on fuel efficiency in city driving. Create a diagnostic scatter plot of the residuals (y-axis) versus predicted CO2 emissions (x-axis). (Chapter 7 discusses this diagnostic plot: for example, see Figure 7.4.) Inspect the diagnostic plot: does it show a clear pattern?
- (c) Continuing with the previous part, compute the coefficient of correlation between the residuals and the x variable (fuel efficiency in city driving).
- (d) Regress the natural log of CO2 emissions in city driving on the natural log of fuel efficiency in city driving. Report the “slope” coefficient. Create a diagnostic scatter plot of the residuals (y-axis) versus predicted natural log CO2 emissions (x-axis). Inspect the diagnostic plot: does it show a clear pattern?
- (e) Continuing with the previous part, compute the coefficient of correlation between the residuals and the x variable (the natural log of fuel efficiency in city driving).

Q^e6. University admissions offices use regression. How well do high school marks predict university marks? Use [hs_univ_marks.xlsx](#) for all subparts. (These data are hypothetical, but designed to be realistic.) Each student is identified by a student id number. Beyond that identifier variable, the provided data contain only two variables: cGPA at the end of first year of university and cGPA_hat. cGPA_hat is each student’s predicted cGPA given her/his high school marks using a simple regression where the y variable is university cGPA and x variable is high school average. The provided datafile deliberately excludes the variable measuring high school average.

- (a) Compute the value of the SST, SSE, and SSR.
- (b) Compute the value of the R-squared.
- (c) Using the formula $\sqrt{\frac{\sum e^2}{n-2}}$, compute the value of the s_e (which, loosely speaking, is the standard deviation of the residuals).
- (d) Compute the standard deviation of cGPA.

Answers for Module B practice questions:

- A1.** (a) The complete table is below. Also, you can verify this answer by looking at the original data for: county group at quantile 50 (poverty rate ranking), females, and age group 0-4 yrs.

Age group	Year	Deaths	Population	Mortality per 1,000	Adjusted population	Adj. deaths	Adj. mortality per 1,000
0 yrs	1990	834.00	77718.00	10.73	77718.00	834.00	10.73
0 yrs	2000	729.00	89991.33	8.10	77718.00	629.58	8.10
0 yrs	2010	598.00	90797.00	6.59	77718.00	511.86	6.59
1-4 yrs	1990	289.00	365123.00	0.79	365123.00	289.00	0.79
1-4 yrs	2000	237.33	359350.33	0.66	365123.00	241.15	0.66
1-4 yrs	2010	203.00	366847.00	0.55	365123.00	202.05	0.55
0-4 yrs	1990	1123.00	442841.00	2.54	442841.00	1123.00	2.54
0-4 yrs	2000	966.33	449341.67	2.15	442841.00	870.72	1.97
0-4 yrs	2010	801.00	457644.00	1.75	442841.00	713.91	1.61

- (b) The complete table is below.

Age group	Year	Deaths	Population	Mortality per 1,000	Adjusted population	Adj. deaths	Adj. mortality per 1,000
0 yrs	1990	900	78000	11.54	78000	900.00	11.54
0 yrs	1995	880	81000	10.86	78000	847.41	10.86
0 yrs	2000	760	90000	8.44	78000	658.67	8.44
0 yrs	2005	700	92000	7.61	78000	593.48	7.61
0 yrs	2010	600	96000	6.25	78000	487.50	6.25
0 yrs	2015	510	98000	5.20	78000	405.92	5.20
1-4 yrs	1990	300	400000	0.75	400000	300.00	0.75
1-4 yrs	1995	270	380000	0.71	400000	284.21	0.71
1-4 yrs	2000	220	370000	0.59	400000	237.84	0.59
1-4 yrs	2005	210	380000	0.55	400000	221.05	0.55
1-4 yrs	2010	200	360000	0.56	400000	222.22	0.56
1-4 yrs	2015	180	350000	0.51	400000	205.71	0.51
0-4 yrs	1990	1200	478000	2.51	478000	1200.00	2.51
0-4 yrs	1995	1150	461000	2.49	478000	1131.62	2.37
0-4 yrs	2000	980	460000	2.13	478000	896.50	1.88
0-4 yrs	2005	910	472000	1.93	478000	814.53	1.70
0-4 yrs	2010	800	456000	1.75	478000	709.72	1.48
0-4 yrs	2015	690	448000	1.54	478000	611.63	1.28

- A2.** (a) Only the slope differs: it is 0.0000394 when GDP per capita is measured in dollars versus 0.0394057 when GDP per capita is measured in thousands of dollars. Everything else – the intercept, R-squared, SST, SSR, SSE and s_e – are identical. The intercept is not changed because a GDP per capita of zero is still zero regardless of measurement in dollars or thousands of dollars. The R-squared is a unit-free statistic so it is not changed. The SST, SSR, and SSE are all measured in units y squared and we changed the units of x so they are not affected. The s_e is measured in units y, so it is also not affected.

- (b) Brazil has the largest residual (2.288865): the price of a Big Mac is more than \$2 US above what would be expected given its GDP per capita. However, Brazil is not the country with the highest priced Big Mac: that is Switzerland at \$6.35 US (versus \$5.12 US in Brazil).
- (c) Running the regression with 48 observations (excluding “Euro area”), the intercept is 2.484995 and the slope is 0.039319. (These are extremely similar to the regression results with all 49 observations: this observation is not an outlier or an influential point.) The predicted price of a Big Mac for a country with GDP per capita of \$27,341 (i.e. $x = 27.341$) is \$3.56 (i.e. $\hat{y} = 3.56$).
- (d) Switzerland a somewhat unusual observation: it is a bit of a gray area about whether to label it an outlier. Regardless, we can compute statistics with and without it to see how sensitive they are. The coefficient of correlation is a bit higher without Switzerland: 0.6464 versus 0.6244.

A3. (a) The results for Brazil:

Half decade	Coefficient	n
1985 - 1990	0.0001	6
1990 - 1995	0.0159	6
1995 - 2000	0.0023	6
2000 - 2005	0.0151	6
2005 - 2010	0.0323	6

- (b) There is no half-decade period between 1985 and 2010 where Brazil has an annual growth rate of real GDP per capita above 5%.

A4. (a) The coefficient of correlation is 0.9327.

- (b) For Bangladesh between 1960-2010, on average the population increased by 2.2 percent annually.

A5. For the correlation, rank correlation, regression coefficient, and R-squared see Table 1 on page 46 to check your answers. The intercept is 0.0264828, the SST is 0.059588362, the SSR is 0.000029923 and the SSE is 0.059558439.

A6. (a) The regression coefficient is 0.1115 and the R^2 is 0.0183.

- (b) The correlation is 0.135. The rank correlation is 0.286. The large difference suggests the presence of outliers. (Recall that the rank correlation is robust to outliers, while the correlation is not.) A scatter plot suggests Angola, The Democratic Republic of the Congo, and Sierra Leone as potential outliers.
- (c) This subset includes 47 African countries. Since, for each country, we used regression to compute two growth rates (one for the 1990s and one for the 2000s), a total of 94 ($=2*47$) regressions were required.

A7. (a) The histogram of growth rates is fairly Normal (Bell shaped). The standard deviation is 2.1%, the range is 11.8 percentage points, and the coefficient of variation is 0.79. The mean growth is 2.6% and a s.d. of 2.1% is large in the context of real GDP per capita growth: countries vary a lot with some growing incredibly quickly and some even experiencing negative growth.

- (b) The histogram of growth rates is fairly Normal (Bell shaped). The standard deviation is 2.4%, the range is 18.4 percentage points, and the coefficient of variation is 1.35. The mean growth is 1.8% and a s.d. of 2.4% is very large in the context of real GDP per capita growth: there is even more variation in growth across countries in the 1990s than in the 2000s.

A8. (a) Following the methods in Module B.3:

country	countrycode	r_1994_04	r_2004_14
Canada	CAN	0.0244	0.0055
Mexico	MEX	0.0166	0.0072
United States	USA	0.0224	0.0029

(b) Following the methods in Module B.2:

Period 1	Period 2	Correlation	Rank Correlation	Regression Coefficient	R-squared	N
Adjacent decades: Three countries (Canada, Mexico and United States)						
1994-04	2004-14	-0.6409	-0.5000	-0.3420	0.4107	3

- (c) Yes, these results look unusual compared to Table 1 on page 46. Unlike Table 1 that showed a weak positive correlation, these results show a moderate negative correlation between growth in adjacent decades. However, given that we used a tiny sample of 3 countries, it is not surprising that we obtained very noisy results.

Answers to the additional questions for extra practice.

A^e1. (a) Check that your Excel graph looks like the dark blue dots in Figure A2, Panel (c).

- (b) Panel (a) shows that there is no relationship between population size and the poverty ranking of the county group. As expected, Panel (b) shows that the mean poverty rate is positively associated with the poverty ranking of the county group; further, this relationship is extremely strong (by construction, as county groups with higher poverty rates will automatically be in a higher percentile group of poverty rates) and nonlinear. As expected, Panel (c) shows that the median income is negatively associated with the poverty ranking of the county group; further, this relationship is very strong and nonlinear. As expected, Panel (d) shows that per capita income is negatively associated with the poverty ranking of the county group; further, this relationship is strong and nonlinear. We use *association* for Panels (b) - (d) because the word *correlation* should only be used to describe linear relationships and all of these are clearly nonlinear.
- (c) A scatter diagram of the 100 county groups in 2010 shows that median household income and income per capita have a strong, positive, linear relationship (so we can say correlation). There is one notable outlier: a county group with the highest income per capita level overall but only a middle value for median household income.
- (d) SST = 39.989; SSR = 0.003; SSE = 39.986; R-squared = 0.0001. This is not surprising as we already noted that Panel (a) of Figure A2 shows no evidence of any relationship between population size and the poverty ranking of the county group.

A^e2. (a) The results for Argentina:

Half decade	Coefficient	n
1985 - 1990	-0.0226	6
1990 - 1995	0.0480	6
1995 - 2000	0.0146	6
2000 - 2005	0.0120	6
2005 - 2010	0.0531	6

- (b) In the half-decade period from 2005 and 2010 Argentina has an annual growth rate of real GDP per capita above 5%.

A^e3. (a) The y variable is the growth rate of real GDP per capita over the two-decade period from 1990-2010. The x variable is the growth rate of real GDP per capita over the two-decade period from 1970-1990.

- (b) There are 142 countries in the regression (i.e. that we have growth rates for for both two-decade periods).

- (c) Using the same methods used in Module B.2, we can obtain an estimate of Canada's growth rate from 1970-1990 (remember: include the endpoints) as 0.0209491, which is the value of the x variable for Canada. Similarly, we obtain an estimate of Canada's growth rate from 1990-2010 (remember: include the endpoints) as 0.0185448, which is the value of the y variable for Canada.

- (d) The answer is the same as the previous part: 0.0185448 is the value of the y variable for Canada.

A^e4. (a) The sample size is only 66 because there are only 66 countries with sufficient real GDP per capita data in the decade from 1950 to 1960. All countries with sufficient real GDP per capita data in the decade from 1950 to 1960 also have sufficient real GDP per capita data in the decade from 1960 to 1970, so what is limiting the sample size is data availability for the earliest decade.

- (b) 66

A^e5. (a) There is an extremely strong, negative, and nonlinear association between CO2 emissions and fuel efficiency.

- (b) Yes, the diagnostic plot shows a very clear U-shaped pattern: it is catching the nonlinearity pointed out in the previous part.

- (c) The coefficient of correlation is 0 (exactly). (Given the limits of machine precision, you may obtain an extremely small number instead of the theoretical answer of zero.) In fact, an alternative way to think about OLS (Ordinary Least Squares) is that it returns the intercept and slope that yield a perfect zero correlation between the x variable and residuals. (OLS also returns the intercept and slope that minimize the sum of the squared residuals.)

- (d) The "slope" coefficient is -0.985166. This means that, for Nissan vehicles in 2017, a 1 percent increase in city fuel economy is associated with nearly a 1 percent decrease (a 0.985 percent decrease) in city CO2 emissions on average. The diagnostic scatter plot shows no clear pattern: the natural log transformations of the x and y variables have successfully straightened the scatter plot.

(e) The coefficient of correlation is 0 (exactly).

A^e6. (a) Using Excel as a spreadsheet, compute:

- $SST = 377.6 = \sum_{i=1}^{1000} (cGPA_i - \overline{cGPA})^2$, where \overline{cGPA} is the mean cGPA, which comes out to 2.3885;
 - $SSE = 367.6 = \sum_{i=1}^{1000} (e_i)^2$, where e_i is the residual for each observation ($e_i = cGPA_i - cGPA_hat_i$);
 - $SSR = 10.1 = \sum_{i=1}^{1000} (cGPA_hat_i - \overline{cGPA})^2$.
- (b) $R^2 = 0.027 = \frac{SSR}{SST}$, which means that less than 3 percent of the variation across students in their first-year university marks can be explained by variation in their high school marks. In other words, more than 97 percent of the variation in first-year university marks is explained by other factors. (Generally, it is very hard to forecast student success in university.)
- (c) $s_e = 0.607$
- (d) $s_{cGPA} = 0.615$. This is a pretty big standard deviation, given that cGPA is on a 4-point scale. Also, notice that the s_e is nearly as big as the s_{cGPA} , which is what we'd expect given the very low R-squared value: nearly all of the variation is scatter around the line. The line – high school marks – explains very little variation in cGPA across students.

C Module C: Sampling Distributions & Inference (CI est. & HT)

C.1 Module C.1: Sampling Distributions and Simulations

Concepts: Using simulation to obtain a *sampling distribution*: the distribution of a *statistic* reflecting variation caused by sampling error (aka sampling variability). Interactively illustrating the Central Limit Theorem for the sampling distribution of the sample mean \bar{X} .

Case studies: We use the *population* of all Ontario public sector employees with salaries of \$100K+ in the “Universities” or “Colleges” sectors in 2016, abbreviated ON Univ. & Col. (2016).

Required readings: Chapter 10 and the bullets below. (Make sure to read Chapter 10 first.)

- Section 10.3 does a dice-rolling simulation with 10,000 *simulation draws* for each of $n = 1$, $n = 2$, $n = 3$, $n = 5$, and $n = 20$. Figures 10.3 to 10.7 illustrate some simulation results. For example, for $n = 3$, toss three dice 10,000 times and record the mean for each toss of three. Figure 10.5, “Three-Dice Average,” shows how the sample mean (\bar{X}) for $n = 3$ varies across the 10,000 samples. Figures 10.3 and 10.7 show the simulation results for $n = 1$ and $n = 20$, respectively. Notice the *sample size* (e.g. $n = 3$) versus the number of *simulation draws*, which is 10,000 in all of these figures. We use n for the sample size and m for the number of simulation draws. We simulate sampling distributions for $n = 10, 25$, and 100 and $m = 500$ and 10,000.

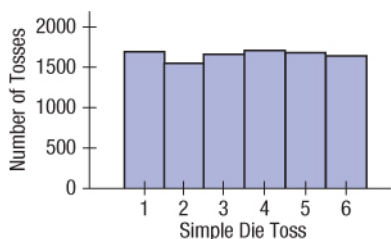


Figure 10.3 Simple die toss.

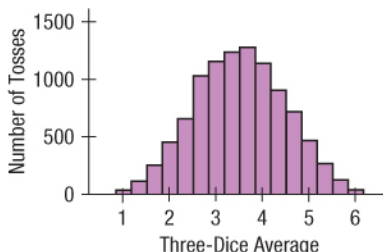


Figure 10.5 Three-dice average.

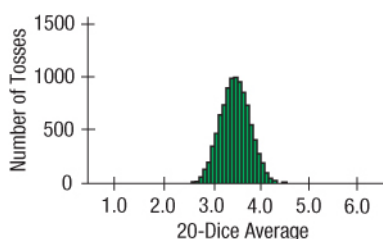
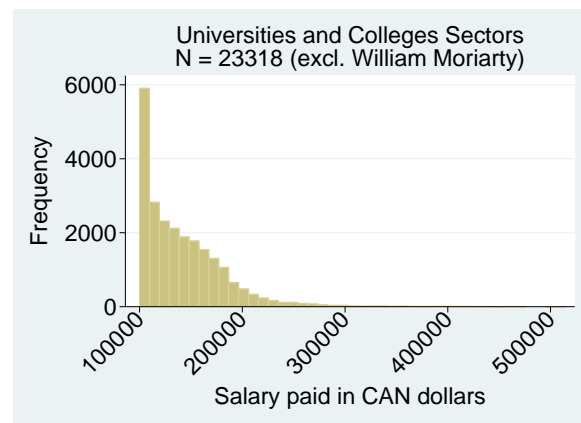
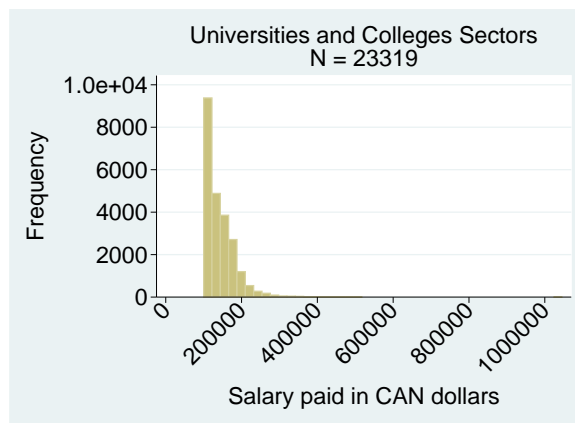


Figure 10.7 Twenty-dice average.

- What is the point of simulations like those illustrated in Figures 10.3, 10.5 and 10.7? These help us answer an abstract question about how a sample statistic *would* vary across samples. In other words, simulations help us figure out sampling distributions.
- *Sample statistics* are *not* perfect measures of *population parameters*. In general, $\bar{X} \neq \mu$ because a random sample of n observations will not be perfectly representative: there is sampling error.
- But, how big of a deal is sampling error? In other words, how accurate is \bar{X} as a measure of μ ? To answer we need to know the *sampling distribution* of \bar{X} , which tells how \bar{X} can differ from μ . \bar{X} has a *distribution* because of sampling error. In an imaginary world where samples were always perfectly representative, there would be no sampling error and \bar{X} would have no distribution: it would always be equal to μ (which, as a parameter, is a constant). Of course in real life there is sampling error and sample statistics are *random variables*. Further, we usually observe *only one* random sample and *only one* sample mean. Computer simulation shows us how \bar{X} *would* vary if we went to the trouble and expense of collecting other random samples. With a computer and a population, it is cheap and quick to repeatedly collect random samples and to directly observe how sample statistics (like \bar{X}) vary across samples. In other words, simulations help us find sampling distributions.

- Figure 10.7 shows that the sampling distribution of the sample mean (\bar{X}) is approximately Normal for $n = 20$. While a toss of $n = 20$ dice could theoretically result in an average value below 2.5 or above 4.5, that is very unlikely. It is likely that we'll get a sample mean (\bar{X}) that is near the population mean (μ), which is 3.5. In contrast, Figure 10.5 shows that a smaller sample size ($n = 3$) causes a lot more sampling error: the sample mean can be way off.
- To do such simulations we need access to a population. In general it is rare to have data on an entire population. Fortunately, there are exceptions (such as censuses). Luckily, the Public Sector Salary Disclosure Act of 1996 mandates that *all* organizations receiving funding from the Province of Ontario publicly disclose the name, job title, salary and taxable benefits of *all* employees paid \$100,000 (CAN) or more in the previous calendar year.
- This includes a broad range of employees, such as police officers, executives in TIFF (Toronto International Film Festival), registered nurses, school teachers, judges, wastewater technicians, university professors, nuclear operators, members of the provincial parliament, directors within the Canadian Red Cross, and TTC engineers. Almost 125,000 ON public sector employees are in the disclosure for the 2016 calendar year. The number of disclosed salaries increases substantially every year because there has been no change in the threshold since 1996. While in 1996 a salary of \$100,000 was notable, two decades later (given inflation), 100K is not as remarkable. Hence, a *much bigger* proportion of ON public sector employees are now having their salaries publicly disclosed compared to when the law was originally passed.
- The disclosure is divided into twelve sectors, listed here from largest to smallest: “Municipalities and Services,” “School Boards,” “Universities,” “Ministries,” “Hospitals and Boards of Public Health,” “Ontario Power Generation,” “Other Public Sector Employers,” “Colleges,” “Crown Agencies,” “Judiciary,” “Legislative Assembly,” and “Seconded.”
- A histogram of the 2016 salaries for the “Universities” and “Colleges” sectors, shows the population is extremely positively (right) skewed. The second histogram drops an outlier (\$1M+).



- But, practically speaking, *how* do you draw a random sample from a population? One method is to assign a random number to each observation. Next, pick all the observations with a random number in a specific interval (e.g. between 10 and 100): that would be a *random sample*. A narrow interval yields a small random sample (small n) and a wide interval yields a big random sample (big n). To be sure to get an exact sample size (e.g. $n = 10$), sort by the random

number and then select the first n observations. To summarize, these are the steps we use to draw random samples from a population:

1. Assign a random number to each observation in the population. We use `RAND()`.
2. Sort by the column of random numbers, which puts the observations in *random* order.
3. Select the first n observations to get a *random sample* with a sample size of n .

Datasets: For ON Univ. & Col. (2016): [on_univ_col_16.xlsx](#), where “on_univ_col_16” abbreviates the Ontario disclosure of 2016 salaries of employees in the “Universities” or “Colleges” sectors.

Interactive module materials for Module C.1:

1. Open [on_univ_col_16.xlsx](#) and **browse** the worksheet “Raw Data.” It includes *all* ON public sector employees in the university or college sectors earning \$100,000 CAN or more in 2016. **Verify** that the population size is $N = 23,319$ (a large population). **Compute** the mean and standard deviation. **Verify** that $\mu = \$141,859.79$ and $\sigma = \$41,434.96$. These are parameters because they describe a population.

EXCEL TIPS: Use the functions COUNT, AVERAGE, and STDEV.P. (The function STDEV.S does the degrees of freedom correction for a *sample*.)

2. Simulating a sampling distribution requires repeatedly drawing random samples from a population. **Browse** the worksheet “500 random samples, n=100,” which has space for 500 random samples ($m = 500$), labeled #1, #2, ..., #500, each with 100 observations ($n = 100$). To save time, 497 random samples are already drawn from the population. **Follow** these steps to draw three more random samples (for a total of $3 + 497 = 500$ samples):

EXCEL TIPS: Use the worksheet “Drawing random samples,” which guides you.

- (a) **Generate** a variable filled with random numbers for each sample you wish to draw.

EXCEL TIPS: Column A in the worksheet “Drawing random samples” has all 23,319 salaries. Column B has random numbers from a Uniform[0,1] distribution. Column C has a *live* random number generator using the `RAND()` function (every time you edit the sheet, it generates a fresh column of random numbers). Select *all* values in Column C (23,319) and copy-and-paste, selecting “Paste Special” and “Values” to Column D (Random #1). Recall the shortcut **Ctrl + Shift + ↓** on page 10 (part 1a of Module A.1). Repeat for Columns E and F (Random #2 and Random #3). Paste *only the values* and *not* the live random number generator (otherwise you cannot sort or retrace your steps). Random #1, #2, and #3 should each contain *different* random numbers.

- (b) **Sort** the population by the variable Random #1.

EXCEL TIPS: Select the entire worksheet “Drawing random samples” and sort.

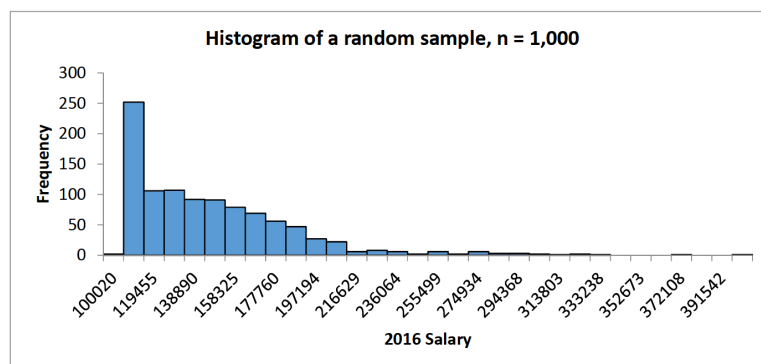
- (c) **Select and copy** the first 100 salaries, which is a *random sample* of size $n = 100$.

EXCEL TIPS: Copy the first 100 salaries and paste (the values) into the worksheet “500 random samples, n=100” in Column #1.

- (d) **Repeat** the previous two steps sorting by Random #2 and pasting into Column #2 and sorting by Random #3 and pasting into Column #3.

3. Before continuing with our simulation, **predict** the shape of a sample with $n = 1,000$. **Check** your prediction by drawing a random sample with $n = 1,000$ from the population and **plotting** a histogram of the sample. **Verify** that your histogram looks similar to the one below. (If your sample happens to include William Moriarty it will have a much longer right tail.)

EXCEL TIPS: Because you already have the population sorted in a random order (if you've completed the previous steps), you can simply copy the first 1,000 salaries and paste these into a new worksheet "Hist random sample 1,000."



- **Note:** If you incorrectly predicted Normal: the question asks about the distribution of a *sample* with $n = 1,000$, *not* the sampling distribution of \bar{X} with $n = 1,000$.

Interpretation tips: What does the positive skew mean? The above histogram mirrors the extreme positive skew of the population. Most of the 1,000 sampled employees of Ontario colleges and universities have 2016 salaries only modestly above the \$100,000 reporting threshold but some have very high salaries (more than 2 or 3 times the threshold), creating the long right tail. This is *not* surprising because the large sample size ($n = 1,000$) means little sampling error: the sample should be highly representative of the population, which is positively skewed (page 72). These basic topics appear early in our course: for example, in our textbook, Section 3.1 (random sampling and sampling error) and Sections 5.1 - 5.2 (histogram of a sample). Chapter 10 is *not about the distribution of a sample*. Instead, it dives into a challenging and abstract topic: the distribution of a *sample statistic*. A *sampling distribution* shows how a sample statistic would vary from one sample to another sample. Do not confuse the (plain-old) distribution of a sample – like the histogram above – with the (fancier) sampling distributions.

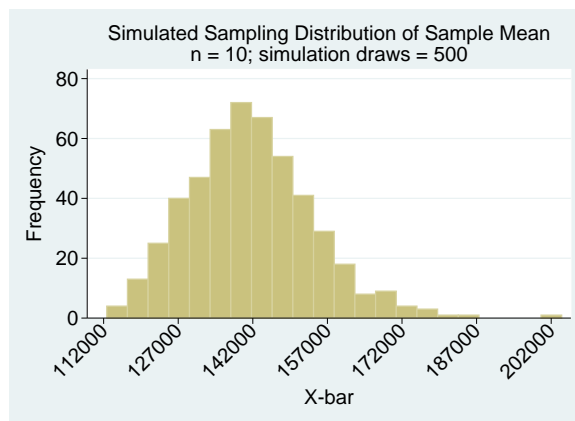
4. We have already drawn 500 random samples ($m = 500$) and are now ready to simulate a sampling distribution. **Consider** $n = 10$: the first 10 observations for each random sample.

- (a) **Compute** \bar{X} for each sample ($n = 10$). **Repeat** for all 500 random samples.

EXCEL TIPS: The worksheet " $n = 10, m = 500$ " automatically pulls the first 10 observations from each of the 500 random samples in worksheet "500 random samples, n=100." In the worksheet " $n = 10, m = 500$," use the AVERAGE function. To efficiently copy and paste, after creating the average in cell B25 [1st screenshot], jump to the last column of data (using the arrow shortcuts on page 10) [2nd screenshot], go up two cells [3rd screenshot], and select back to the initial cell (using the arrow shortcuts on page 10) [4th screenshot]. Now with that area selected, hit **Ctrl + R** (pc) or **command + R** (mac).

This autofills the formulas. A slower (but simpler) method is to copy cell B25, click and drag over the remaining 499 cells, and paste.

- (b) **Plot** a histogram of the 500 sample means. **Verify** that it looks similar to this histogram.



EXCEL TIPS: To save time, the worksheet “ $n = 10, m = 500$ ” is set up to automatically draw a histogram once the previous steps are complete. Note that your histogram will not be identical because your simulation has three random samples generated interactively and because Stata and Excel draw histograms a bit differently.

- (c) **Compute** the mean of the 500 sample means. **Compute** the standard deviation of the 500 sample means. **Verify** that your simulation results are similar to what theory predicts: $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. (Recall that you computed μ and σ in part 1 on page 73.)

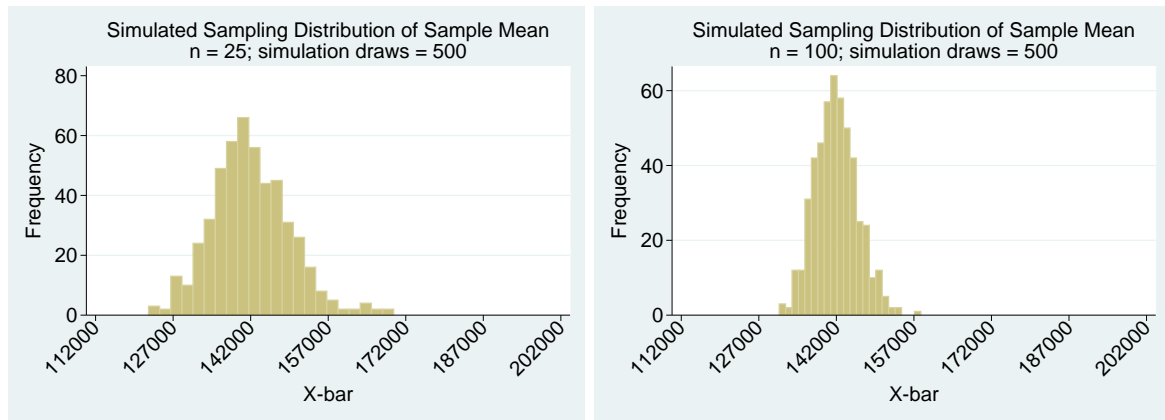
Interpretation tips: What do the results in parts 4b and 4c mean? These results tell us about the *sampling distribution* of the sample mean (\bar{X}) (i.e. the fancy Chapter 10 material). The histogram in part 4b shows the *shape* of the sampling distribution of \bar{X} , which we found using a computer simulation. In this particular case, a sample size of $n = 10$ is not sufficiently large to apply the Central Limit Theorem: you can clearly see some lingering positive skew in the histogram, even though it is much closer to Normal than the extremely positively skewed population. In part 4c, we use well-known theoretical results to find the mean and standard deviation of \bar{X} . The results that $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ are true regardless of the sample size. The purpose of the simulation is to get the shape (part 4b). However, seeing if the mean of the 500 sample means is about equal to μ and if the standard deviation of the 500 sample means is about equal to $\frac{\sigma}{\sqrt{n}}$ is a nice way to check our simulation: if the answer from our simulation differs substantially from what theory predicts, we know we made a programming error.

5. Using the worksheet “ $n = 25, m = 500$,” **repeat** the steps you did for $n = 10$ but now with $n = 25$. Keep the scale of the x-axis the same as for $n = 10$. This is what your textbook does in

Figures 10.3 to 10.7 reviewed on page 71. Keeping the scale of the x-axis constant, even as the sampling distribution itself is less spread out, helps people notice that as the sample size goes up, sampling error goes down. Use the (left) histogram below to (roughly) check your work.

EXCEL TIPS: Note that you will have many empty bins. The bins in “ $n = 25, m = 500$ ” are set up to keep the range of the x-axis constant.

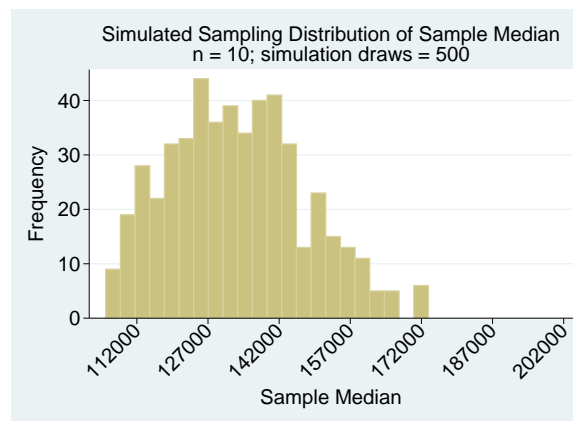
6. Using the worksheet “ $n = 100, m = 500$,” **repeat** the steps you did for $n = 10$ but now with $n = 100$. Use the (right) histogram below to (roughly) check your work.



7. What about the sampling distribution of the *sample median*? In ECO220Y, we do not cover theoretical results for the *median*. Hence, simulation is particularly useful. Let’s repeat the simulation we did for the sample mean ($n = 10$ with $m = 500$) for the sample median.

- (a) **Compute** the sample median for each sample ($n = 10$). **Repeat** for all 500 random samples. **Verify** that your histogram looks similar to the one below.

EXCEL TIPS: Go to the worksheet “Median $n = 10, m = 500$.” Using the MEDIAN function, compute the sample median of each of the 500 samples (with $n = 10$).



- (b) **Compute** the mean of the 500 sample medians. **Verify** it’s *roughly* \$133,000. **Compute** the standard deviation of the 500 sample medians. **Verify** it’s *roughly* \$14,500.
- (c) **Compute** the population median. **Verify** it’s *exactly* \$131,769.30.

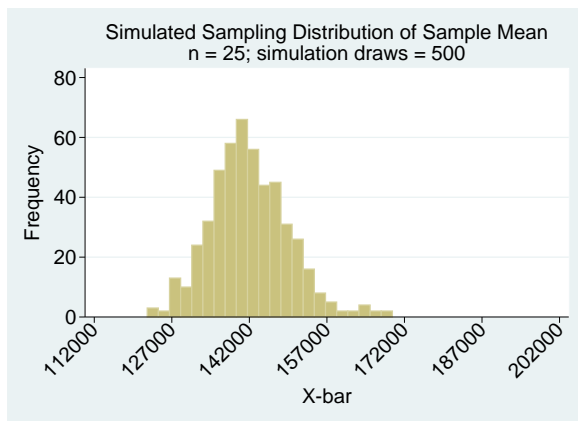
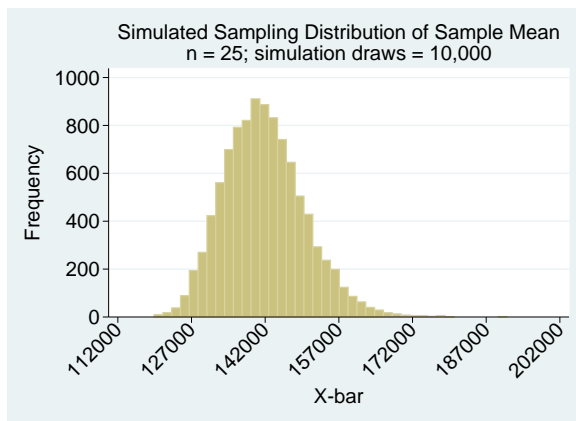
Interpretation tips: What do the results in parts 7a, 7b, and 7c mean? Parts 7a and 7b show the sampling distribution of the sample median, obtained using a computer simulation.

The shape is mildly positively skewed. As expected, the sample median (a statistic) is not a perfect measure of the population median (a parameter). In fact, with the small sample size of $n = 10$ there is substantial sampling error: notice the spread-out histogram in part 7a and the big standard deviation of the sample median in part 7b. For example, it is not uncommon to see sample medians as low as \$120,000 or as high as \$150,000, which is quite far off from the population median of \$131,769.30 in part 7c. Hence, with a small sample of only 10 observations, the sample median is a noisy measure of the population median. On average across many samples, the sample median is fairly close to the population median but the simulation results suggest it may suffer a bit of upward bias (i.e. \$133,000 > \$131,729.30).

8. **Consider** that you have varied the sample size from $n = 10$ to $n = 25$ to $n = 100$ for \bar{X} . What about changing the number of simulation draws? **Browse** the worksheet “ $n = 25, m = 10000$,” **noting** that each *row* shows a random sample of 25 observations. There are 10,000 rows. Hence, $n = 25$ and $m = 10,000$. **Compute** the sample mean of each sample and **plot** the simulated sampling distribution of the sample mean for $n = 25$ and $m = 10,000$. **Verify** that it looks similar to the histogram below.

EXCEL TIPS: In the column labeled X-bar, compute the sample mean for each row using the AVERAGE function. (Exclude Column A: it is not a salary.) Recall the shortcut to double click on the bottom right of the cell (see screenshot below) to autofill the column. (Autofilling columns is easier than rows: Excel tip in part 4a on page 74.) Worksheet “ $n = 25, m = 10000$ ” is set-up to create a histogram with the same bins as earlier for a direct comparison. (Note: More simulation draws would ordinarily imply more bins: see Stata histograms below.)

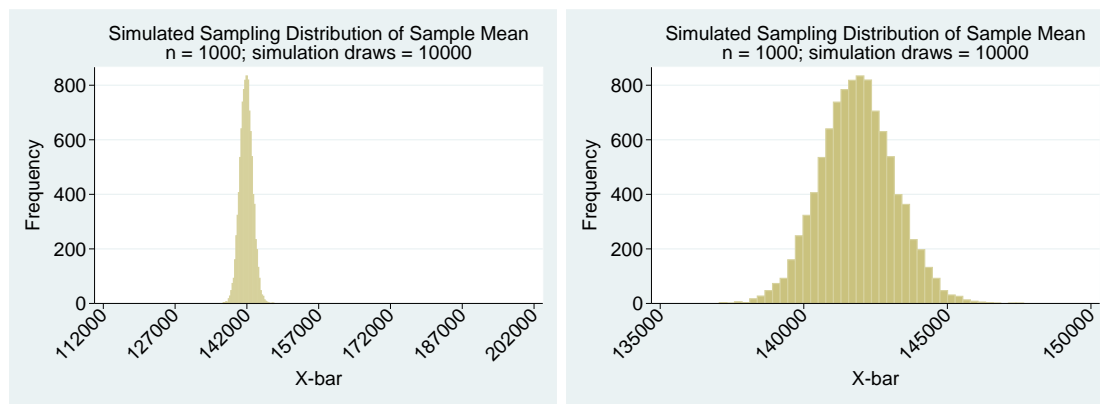
	W	X	Y	Z	AA	AB
1	salary22	salary23	salary24	salary25	X-bar	
2	142160	187061.9	108561.5	146246.412		
3	120648.1	163185.5	166612	100191.1		
4	108773.6	159841.6	130415.9	116584.4		



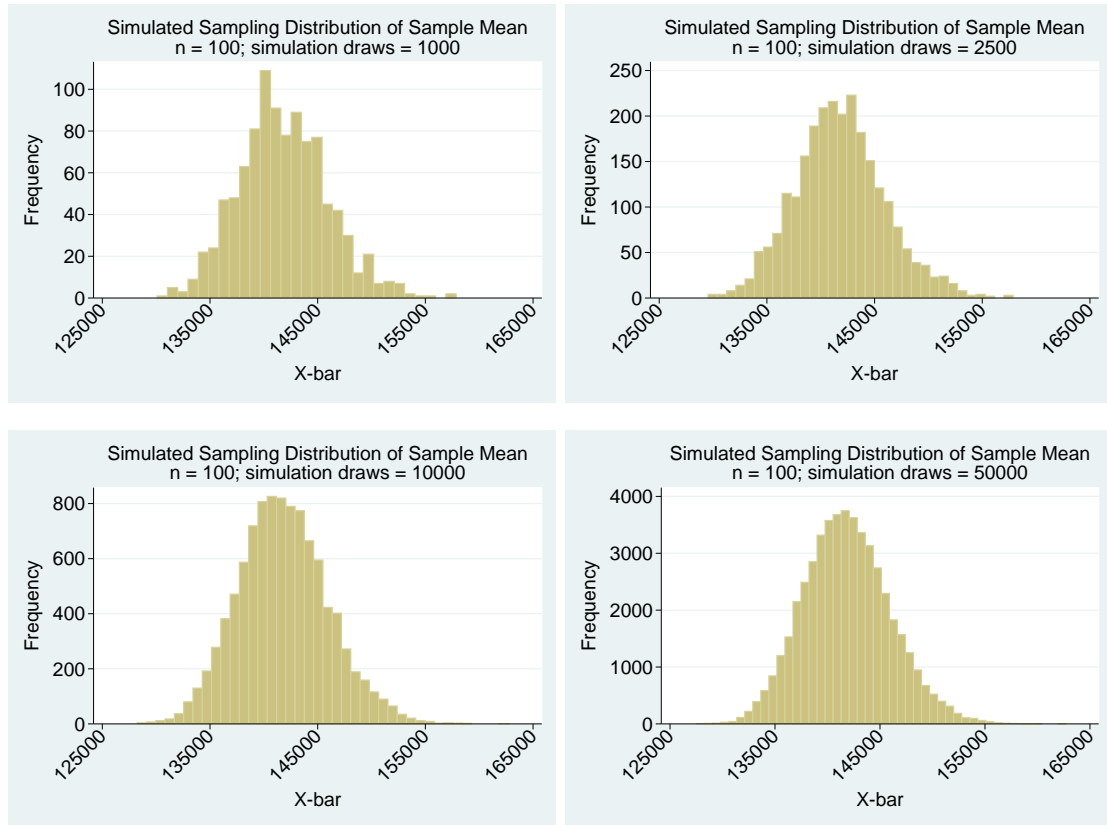
Interpretation tips: What does the comparison between the above two histograms mean? Both show the sampling distribution of the sample mean, obtained via simulation, when $n = 25$. The more powerful simulation (left histogram) with $m = 10,000$ gives a picture of the sampling distribution of \bar{X} that is very similar to the smaller $m = 500$ simulation (right histogram). The more powerful simulation gives a *slightly* clearer picture of the sampling distribution.

9. **Study** the following points, which require reading, thinking, and synthesizing what you have learned, but no further actions in Excel.

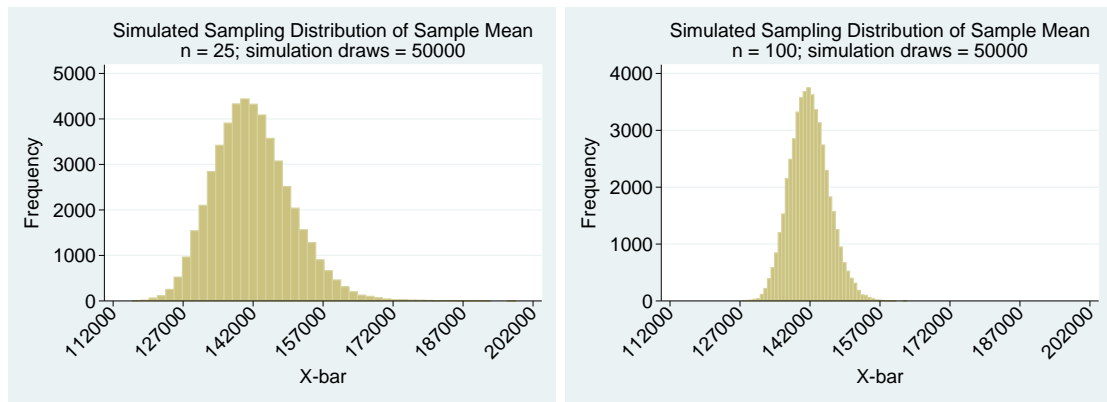
- (a) Recall that a simulation helps learn the *shape* of the sampling distribution of \bar{X} when we cannot use the Central Limit Theorem (CLT). The CLT, a useful theoretical result about the shape of the sampling distribution of \bar{X} , requires a *sufficiently large* sample size. However, because the salary population is extremely skewed, this module showed that a sample size of $n = 10$ is not sufficiently large (positive skew is visible in the simulated sampling distribution of \bar{X}). In fact, even $n = 25$ and $n = 100$ show traces of skew. However, regardless of the shape, we have clear theoretical results for the mean of the means and the standard deviation of the means: $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.
- (b) Consider $n = 1,000$. This simulation would be cumbersome in Excel. However, it helps complete the story of how the *shape* of the sampling distribution of \bar{X} approaches Normal as the sample size increases. Even if the population is very skewed, a sample size of $n = 1,000$ is surely sufficiently large so that the CLT guarantees the sampling distribution of \bar{X} will be Normal. Review the histograms below to see this. Each shows $n = 1,000$ and $m = 10,000$. (Note: Given that we are not doing these simulations interactively in Excel, we do not need to limit ourselves to a small simulation of $m = 500$, even though, we saw above that $m = 500$ gives a good picture that is only a little clearer with $m = 10,000$.) In the graph on the left, the horizontal axis is scaled for easy comparison to the simulated sampling distributions of \bar{X} for $n = 10$, $n = 25$, and $n = 100$. Notice that with $n = 1,000$ there is very little sampling error: the simulated sampling distribution is much less spread out. This is as expected from theory: $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. Also, notice the *shape* with $n = 1,000$ is Normal: there is no longer any trace of positive skew.



- (c) Consider further the important distinction between n (the sample size) and m (the number of samples drawn for the simulation). As seen in this module, changing n fundamentally changes the sampling distribution. In contrast, changing m simply affects how clearly we can see the sampling distribution. A photography analogy: changing n is like pointing your camera away from a giant panda and at a penguin, whereas increasing m is like focusing your camera to get a clearer photo of the giant panda. The graphs below illustrate how increasing m gives a clearer picture of the sampling distribution, but the sampling distribution itself is not changing (i.e. we are not changing n).



In contrast to the four graphs above that vary only m and are very similar (i.e. varying focus but all of a giant panda), the graphs below show the dramatic difference from varying the sample size n (i.e. point at a penguin instead of a giant panda).



10. **Familiarize** yourself with more recent data. Consider more recent ON public sector salary data and for all sectors, not just the “Universities” and “Colleges” sectors. Each year the Government of Ontario posts salaries for those public sector employees paid \$100,000 or more <https://www.ontario.ca/page/public-sector-salary-disclosure>.

- [on_sal.2024.xlsx](#): These data contain the 2024 salaries as disclosed in 2025. These data have had some minor cleaning. It includes sector, employer, fullname, salary, taxben, and jobtitle, random1, random2, random3, and random4. Sorting the data by any of the variables named random1 to random4 puts the observations in a random order, which

facilitates drawing random samples. For how random1 to random4 are created, see Module D.2, which starts on page 112.

- [on_sal_2023.xlsx](#): These data contain the 2023 salaries as disclosed in 2024 and have the same structure as the 2024 salary data. These data have had some minor cleaning.
- [on_sal_2022.xlsx](#): These data contain the 2022 salaries as disclosed in 2023 and have the same structure as the 2024 salary data. These data have had some minor cleaning.
- [on_sal_2021.xlsx](#): These data contain the 2021 salaries as disclosed in 2022 and have the same structure as the 2024 salary data. These data have had some minor cleaning.
- [on_sal_2020.xlsx](#): These data contain the 2020 salaries as disclosed in 2021 and have the same structure as the 2024 salary data. These data have had some minor cleaning.
- [on_sal_2019.xlsx](#): These data contain the 2019 salaries as disclosed in 2020 and have the same structure as the 2024 salary data. These data have had some minor cleaning.

Test/exam examples: The Ontario salary disclosure data appears often.

- Question (7), [November 2024 Test #2](#) (with [solutions](#))
- Question (7), [June 2023 Test #2](#) (with [solutions](#))
- Question (2), [December 2022 Test #2](#) (with [solutions](#))
- Question (9), [December 2021 Test #2](#) (with [solutions](#))
- Question (7), [January 2020 Test #3](#) (with [solutions](#))
- Question (6), [January 2018 Test #3](#) (with [solutions](#))
- Question (4), [January 2017 Test #2](#) (with [solutions](#))
- Question (6), [December 2015 Test #2](#) (with [solutions](#))
- Question (1), [February 2016 Test #3](#) (with [solutions](#))
- Question (2), [June 2015 Test #2](#) (with [solutions](#))
- Part 2, Questions (3) - (4), [April 2015 Final Exam](#) (with [solutions](#))

C.2 Module C.2: Proportions & Confidence Intervals

Concepts: Analyzing categorical data. Single proportions and the difference between two proportions. Computing and interpreting a standard error (SE), margin of error (ME), and CI estimate.

Case studies: We replicate parts of an academic journal article “Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment,” abbreviated Karlan and List (2007).

Required readings: Chapter 11. Background reading for Karlan and List (2007):

- “We take advantage of a capital campaign in which more than 50,000 prior donors to a US organization received direct mail solicitations seeking contributions. Individuals were randomly assigned to either a control group or a matching grant treatment group, and within the matching grant treatment group individuals were randomly assigned to different matching grant rates, matching grant maximum amounts, and suggested donation amounts.” (p. 1775)
 - “Capital campaign” means a charity’s fundraising effort.
 - “Prior donors” means that researchers had the list of names and addresses of people who had donated at least once to this charity in the past.
 - “US organization” is intentionally vague: the charity asked to remain anonymous.
 - “Direct mail” means physical letters mailed to a person’s home.
- “Control group” received the usual letter the charity sends when fundraising.
- “Treatment group” received a modified letter offering a match funded by a grant. Here is sample letter. The underlined items are randomly varied across people in the treatment group.

MATCHING GRANT: NOW IS THE TIME TO GIVE!

Troubled by the continued erosion of our constitutional rights, a concerned member has offered a matching grant of \$50,000 to encourage you to contribute to [our charity] at this time. To avoid losing the fight to defend our religious freedom, this member has announced the following match: \$2 for every dollar you give. So, for every \$25 you give, [our charity] will actually receive \$75. Let’s not lose this match – please give today!

- Five “matching grant maximum amounts”: N/A (control group), \$25,000, \$50,000, \$100,000, and unstated (no limit to the grant indicated in the letter), which Karlan and List (2007) also refer to as the “match threshold.” The sample letter above shows \$50,000.
- Four “matching grant rates”: no match (control group), 1 to 1, 2 to 1, and 3 to 1, which Karlan and List (2007) also refer to as the “match ratio.” The sample letter above shows a 2:1 match ratio: if you donate \$25, the charity will get \$75 ($=\$25 + 2*\25).
- Four “suggested donation amounts”: N/A (control group), “low,” “medium,” and “high,” which Karlan and List (2007) also refer to as the “match example amount.” This is the person’s highest previous donation amount (“low”), 25% more than that (“medium”), or 50% more than that (“high”). The sample letter above shows someone whose highest previous donation amount is \$25 and who is randomly assigned “low.”

- Review Table C.1. It shows how the 50,083 prior donors are *randomly divided* into the control group (16,687 observations) or one of the many treatment groups (33,396 observations).

Table C.1: Summary of experimental design: Sample sizes in each group

Maximum size of matching grant (match threshold)	Match example amount	Match ratio			
		No match (0:1)	1:1	2:1	3:1
No grant, N/A	No example, N/A	16,687	-	-	-
\$25,000	Low	-	928	927	927
\$25,000	Medium	-	929	929	928
\$25,000	High	-	927	929	926
\$50,000	Low	-	929	925	927
\$50,000	Medium	-	928	928	927
\$50,000	High	-	925	928	928
\$100,000	Low	-	929	927	929
\$100,000	Medium	-	926	928	927
\$100,000	High	-	928	928	928
Unstated	Low	-	929	929	928
Unstated	Medium	-	927	928	928
Unstated	High	-	928	928	926

- Review Table 2A, which shows that in this capital campaign soliciting further money from prior donors, 1.8% of those in the control group donated and 2.2% of those in a treatment group donated. Karlan and List (2007) focus on the match ratio, combining the treatment groups with the same match ratio but different maximum grants and suggested donation amounts.

TABLE 2A—MEAN RESPONSES
(Mean and standard errors)

	Control	Treatment	Match ratio		
			1:1	2:1	3:1
Implied price of \$1 of public good:	1.00	0.36	0.50	0.33	0.25
<i>Panel A</i>	(1)	(2)	(3)	(4)	(5)
Response rate	0.018 (0.001)	0.022 (0.001)	0.021 (0.001)	0.023 (0.001)	0.023 (0.001)
Dollars given, unconditional	0.813 (0.063)	0.967 (0.049)	0.937 (0.089)	1.026 (0.089)	0.938 (0.077)
Dollars given, conditional on giving	45.540 (2.397)	43.872 (1.549)	45.143 (3.099)	45.337 (2.725)	41.252 (2.222)
Dollars raised per letter, not including match	0.81	0.97	0.94	1.03	0.94
Dollars raised per letter, including match	0.81	2.90	1.87	3.08	3.75
Observations	16,687	33,396	11,133	11,134	11,129

Figure of Table 2A: Karlan and List (2007), p. 1781, Panel A only.

Optional: In Karlan and List (2007), see pp. 1774-1782: abstract, introduction, and sections “I. Experimental Design,” “A. Price Ratio,” “B. Maximum Size of the Matching Grant,” “C. Ask Amount,” “D. Heterogeneous Treatment Effects,” and first part of “II. Experimental Results.”

Datasets: For Karlan and List (2007): [char_give.xlsx](#), where “char_give” abbreviates “Charitable Giving” from the title.

Interactive module materials for Module C.2:

- EXCEL TIPS:** Use the IF function. If Column A has the variable treatment, you can make a new variable named group with =IF(A2=0,"Control","Treatment"). If Column G has the variable gave, make a new variable named give_status with =IF(G2=0,"Did not give","Gave"). Add these to the original data (i.e. put them in Columns T and U). Also, recall the convenient Excel shortcut to autofill these functions in part 8 of Module C.1 on page 77.

- (a) **Replicate** the response rates of $0.018 \left(\frac{298}{16,687} = 0.017858 \right)$ and $0.022 \left(\frac{736}{33,396} = 0.022039 \right)$.

	G	H	I	J	K	L		T	U
1	gave	mos_last	high_prev	num_prev	yrs_init	don_2005	fe	group	give_status
2	0	31	45	2	4	1		Control	Did not give
3	0	5	25	2	3	0		Control	Did not give
4	0	6	50	3	2	0		1 Treatment	Did not give
5	0	1	50	15	8	0		0 Treatment	Did not give
6	0	24	25	42	95	1		1 Treatment	Did not give
7	0	3	90	20	10	0		0 Control	Did not give
8	0	4	100	12	8	0		0 Treatment	Did not give
9	0	4	65	13	16	0		1 Treatment	Did not give
10	0	6	200	28	19	0		0 Treatment	Did not give
11	0	35	125	4	7	1		1 Treatment	Did not give
12	0	41	100	1	3	1		0 Treatment	Did not give
13	0	8	5	1	1	1		0 Treatment	Did not give
14	0	28	25	2	6	1		0 Control	Did not give
15	0	15	25	80	19	1		0 Treatment	Did not give

Create PivotTable

Choose the data that you want to analyze

☒ Select a table or range

Table/Range: char_give\$A\$1:U\$50084

☐ Use an external data source

Choose Connections...

Connection name:

☐ Use this workbook's Data Model

Choose where you want the PivotTable report to be placed

☐ New Worksheet
 ☒ Existing Worksheet

Location: CI Est. of Resp. Rate!\$F\$3

Choose whether you want to analyze multiple tables

☐ Add this data to the Data Model

OK

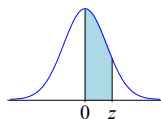
Cancel

(b) Using $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ from our aid sheets, **replicate** the standard errors of 0.001 in parentheses, Columns (1) and (2). **Verify** that you get 0.00102522 and 0.00080335, respectively.

	A	B	C	D	E	F	G	H	I
1	Template for using Excel to compute the Confidence Interval (CI) estimate for a single proportion:								
2									
3	X =	298				Count of give_status	Column Labels		
4	n =	16687				Row Labels	Control	Treatment	Grand Total
5	p-hat =	0.017858				Did not give	16389	32660	49049
6	SE(p-hat) =	0.001025				Gave	298	736	1034
7	alpha =					Grand Total	16687	33396	50083
8	alpha/2 =								
		char_give	readme			CI Est. of Resp. Rate CONTROL		CI Est. of Resp. Rate TREATMENT	...

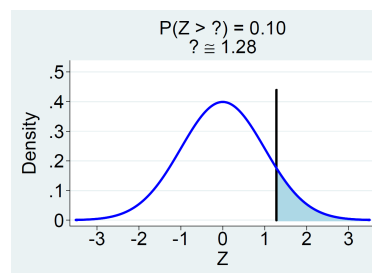
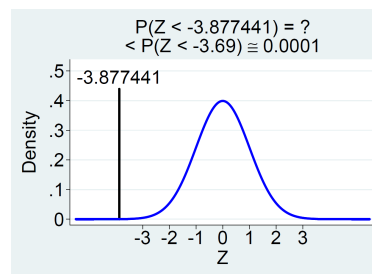
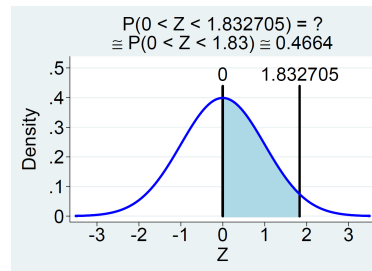
Interpretation tips: What do the values 0.017858213 and 0.00102522 mean? Of the 16,687 people in the control group, who received the regular letter the charity had used before Karlan and List (2007), 1.8 percent responded with a donation. The point estimate of 1.8 percent is measured extremely precisely: the huge sample size ($n = 16,687$) means little sampling error. The standard error, a measure of sampling error, is only 0.1 percent.

4. **Recall** the Normal table. In the three graphs below, **find** the *approximate* value of “?” in each using the Normal table. Next, we will learn to use software for exact answers.



The Standard Normal Distribution:

z	Second decimal place in z									
	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999



(a) Obtain the *exact* value of “?” for each of the above three graphs using software.

- i. **Solve** $P(0 < Z < 1.832705) = ?$ using software. **Verify** that you get 0.466577.

EXCEL TIPS: Use the NORM.S.DIST function, which returns a probability. It has two inputs: a z value and a logical value. Specifying TRUE as the logical value means the function returns the cumulative area *to the left* of the z value: $P(Z < z)$. For example, NORM.S.DIST(-1.96, TRUE) returns 0.024998: $P(Z < -1.96) = 0.024998$.

- ii. **Solve** $P(Z < -3.877441) = ?$ using software. **Verify** that you get 0.000053.

EXCEL TIPS: Again, use the NORM.S.DIST function.

- iii. **Solve** $P(Z > ?) = 0.10$ using software. **Verify** that you get 1.281552.

EXCEL TIPS: Use the NORM.S.INV function, which returns a z value. It has one input: a probability, which is the area *to the left* under the Normal density function. If the area to the right is 0.10, that implies the area to the left is 0.90.

5. Return to [char_give.xlsx](#) and the charitable giving case study.

- (a) **Compute** the margin of error (ME) for the response rate of the treatment group. Use a 95% confidence level. Recall the ME is $z_{\alpha/2} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$ from our aid sheets. The term $z_{\alpha/2}$, which the textbook calls z^* , means $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = (1 - \alpha)$, which also implies that $P(Z < -z_{\alpha/2}) = \alpha/2$ and $P(Z > z_{\alpha/2}) = \alpha/2$. **Verify** the ME is 0.001575.

EXCEL TIPS: Use “CI Est. of Response Rate TREATMENT.” Use the NORM.S.INV function for $z_{\alpha/2}$. Either NORM.S.INV(0.975) or ABS(NORM.S.INV(0.025)) return the correct value of $z_{\alpha/2}$ for $\alpha = 0.05$, which is 1.959963985.

- (b) Continuing, **compute** the 99% confidence interval estimate of the response rate for the treatment group. **Verify** that you get [0.019969, 0.024108].

Interpretation tips: What does [0.019969, 0.024108] mean? We are 99% confident that among all donors (i.e. the population) somewhere between 2.0 and 2.4 percent would send a donation to the charity in response to a letter offering a dollar-for-dollar match. This combines match ratios of 1:1, 2:1, and 3:1 where the first number says how many dollars a wealthy donor will send to the charity for every dollar the person receiving the letter donates: e.g. 2:1 means that a donation of \$50 results in total donation of \$150 (the original \$50 plus a \$100 match). While the point estimate – a 2.2 percent response rate – may seem dismal, fundraising via direct mail is notoriously difficult. In this context, a response rate around two percent is not bad. This response rate estimate is quite precise – the margin of error is only 0.2 percent (even with a 99% confidence level) – which is not surprising given the very large sample size: $n_T = 33,396$.

- (c) **Repeat** Part 5b but for the *control group* and with a 90% confidence level. Verify that you get [0.0161719, 0.019545].

6. How much does the match *improve* the response rate? Answering requires an inference about the *difference* in proportions, not each proportion by itself (like the previous parts). **Recall** the CI estimate of the difference: $(\hat{P}_2 - \hat{P}_1) \pm z_{\alpha/2} \sqrt{\frac{\hat{P}_2(1-\hat{P}_2)}{n_2} + \frac{\hat{P}_1(1-\hat{P}_1)}{n_1}}$. To estimate how much the match *improves* the response rate, the treatment group is 2 and the control group is 1.

- (a) Using your solution to part 3a, **compute** the point estimate of the difference in response rates between the treatment group and control group. **Verify** that you get 0.00418035.

EXCEL TIPS: Use “CI Est. of Diff. in Resp. Rates” to organize your work.

- (b) Continuing, **compute** the standard error (SE) of the difference in response rates. Recall that the SE is $\sqrt{\frac{\hat{P}_2(1-\hat{P}_2)}{n_2} + \frac{\hat{P}_1(1-\hat{P}_1)}{n_1}}$. **Verify** that you get 0.00130248.

- (c) Continuing, **compute** the 95% confidence interval estimate of the difference in response rates. Recall that the ME is $z_{\alpha/2} \sqrt{\frac{\hat{P}_2(1-\hat{P}_2)}{n_2} + \frac{\hat{P}_1(1-\hat{P}_1)}{n_1}}$ and **verify** that you get 0.00255281. Similarly, **verify** that you obtain [0.00162755, 0.00673316] as the lower confidence limit (LCL) and upper confidence limit (UCL), respectively.

Interpretation tips: What does [0.00162755, 0.00673316] mean? We are 95% confident that among all potential donors, if the charity switched from its standard letter to a letter offering a match, this would cause the response rate to increase by between 0.2 to 0.7 percentage points. Building on the interpretation with part 5b, these results are economically significant: if response rates are only around 2 percent, the interval estimate corresponds to roughly a 10 to 35 percent boost ($10 = 100 * 0.2/2$ and $35 = 100 * 0.7/2$).

7. Next, consider Panels B and C in Table 2A, which check for “heterogeneous treatment effects.”

TABLE 2A—MEAN RESPONSES
(Mean and standard errors)

	Control	Treatment	Match ratio		
			1:1	2:1	3:1
Implied price of \$1 of public good:	1.00	0.36	0.50	0.33	0.25
<i>Panel A</i>	(1)	(2)	(3)	(4)	(5)
Response rate	0.018 (0.001)	0.022 (0.001)	0.021 (0.001)	0.023 (0.001)	0.023 (0.001)
Dollars given, unconditional	0.813 (0.063)	0.967 (0.049)	0.937 (0.089)	1.026 (0.089)	0.938 (0.077)
Dollars given, conditional on giving	45.540 (2.397)	43.872 (1.549)	45.143 (3.099)	45.337 (2.725)	41.252 (2.222)
Dollars raised per letter, not including match	0.81	0.97	0.94	1.03	0.94
Dollars raised per letter, including match	0.81	2.90	1.87	3.08	3.75
Observations	16,687	33,396	11,133	11,134	11,129
<i>Panel B: Blue states</i>					
Response rate	0.020 (0.001)	0.021 (0.001)	0.021 (0.002)	0.022 (0.002)	0.021 (0.002)
Dollars given, unconditional	0.897 (0.086)	0.895 (0.059)	0.885 (0.102)	0.974 (0.110)	0.826 (0.091)
Dollars given, conditional on giving	44.781 (2.914)	42.444 (1.866)	42.847 (3.356)	44.748 (3.456)	39.635 (2.838)
Dollars raised per letter, not including match	0.90	0.89	0.88	0.97	0.83
Dollars raised per letter, including match	0.90	2.66	1.77	2.92	3.30
Observations	10,029	19,777	6,634	6,569	6,574
<i>Panel C: Red states</i>					
Response rate	0.015 (0.001)	0.023 (0.001)	0.021 (0.002)	0.024 (0.002)	0.026 (0.002)
Dollars given, unconditional	0.687 (0.093)	1.064 (0.085)	0.987 (0.157)	1.103 (0.148)	1.101 (0.135)
Dollars given, conditional on giving	47.113 (4.232)	45.490 (2.607)	47.667 (5.848)	46.110 (4.392)	43.161 (3.507)
Dollars raised per letter, not including match	0.69	1.06	0.99	1.10	1.10
Dollars raised per letter, including match	0.69	3.23	1.97	3.31	4.40
Observations	6,648	13,594	4,490	4,557	4,547

Figure of Table 2A: Karlan and List (2007), p. 1781.

Interpretation tips: What do Karlan and List (2007) mean by “heterogeneous treatment effects”? It means checking if the effectiveness of the new fundraising letter formats varies across types of people. “Hetero” means different. Table 2A checks if people in Red states responded differently to the treatments (the letters offering a match) compared to people in Blue states. There is some evidence that they did. This is important in marketing: the effectiveness of marketing strategies often varies across groups (e.g. younger versus older people).

- (a) **Replicate** the response rates of 0.020 and 0.021 in Columns (1) and (2), Panel B, *Blue states*, Table 2A. **Verify** that you obtain: $\frac{201}{10,029} = 0.020041879$ and $\frac{417}{19,777} = 0.021085099$.

EXCEL TIPS: Insert a PivotTable with the variables: group, give_status, and blue_state. Drag the variables group and give_status to ROWS, blue_state to COLUMNS, and drag another copy of group to Σ VALUES. (Note that while you can use give_status instead of group in Σ VALUES you *cannot* use blue_state because it contains some missing values.) Also, for a clearer table, under the Design tab under PivotTable Tools, select Show in

Tabular Form under Report Layout.

The screenshot shows an Excel spreadsheet with a PivotTable. The PivotTable is titled 'Count of group' and is set to 'Show in Tabular Form'. The data is organized by 'blue_state' (0 and 1) and 'group' (Control and Treatment). The 'give_status' is also included as a column header. The PivotTable Fields task pane on the right shows the fields 'blue_state', 'group', and 'give_status' assigned to the report.

blue_state	group	give_status	Count of group
0	Control	Did not give	6551
0	Control	Gave	97
0	Treatment	Did not give	13276
0	Treatment	Gave	318
1	Control	Did not give	9828
1	Control	Gave	201
1	Treatment	Did not give	19360
1	Treatment	Gave	417
Grand Total			
	Control		6648
	Treatment		13594
	Grand Total		20242

- (b) **Replicate** the standard errors of the response rates in Columns (1) and (2), Panel B, *Blue States*, Table 2A. Next, **compute** the **99%** confidence interval estimate of the difference in the response rate between the control and treatment groups for prior donors living in Blue states. **Verify** that your point estimate of the difference is 0.00104322 with SE 0.00173263 and ME 0.00446296, which yields the 99% CI estimate $[-0.00341974, 0.00550618]$.
- (c) **Replicate** the response rates of 0.015 and 0.023 and their associated standard errors of 0.001 and 0.001 in Columns (1) and (2), Panel C, *Red states*, Table 2A. Next, compute the **90%** confidence interval estimate of the difference in the response rate between the control and treatment groups among donors living in Red states. Verify that your point estimate of the difference is 0.00880182 with SE 0.00196043 and ME 0.00322463, which yields the 90% CI estimate $[0.00557719, 0.01202645]$.

Test/exam examples: Karlan and List (2007) and similar studies (e.g. responses of potential employers to (fictitious) resumes) have appeared frequently. You are ready for the first now, and the rest after completing Module D.

- Question (6), [June 2023 Test #2](#) (with [solutions](#))
- Question (3), [February 2023 Test #3](#) (with [solutions](#))
- Questions (3) and (4), [March 2022 Test #3](#) (with [solutions](#))
- Question (3), [Summer 2019 Test #3](#) (with [solutions](#))
- Question (5), [January 2017 Test #2](#) (with [solutions](#))
- Question (3), [February 2017 Test #3](#) (with [solutions](#))
- Part 1, Question (2), [April 2016 Final Exam](#) (with [solutions](#))
- Question (3), [June 2015 Test #2](#) (with [solutions](#))
- Question (4), [January 2015 Test #3](#) (with [solutions](#))
- Question (2), [January 2014 Test #2](#) (with [solutions](#))

C.3 Module C.3: Comparing Two Groups & Hypothesis Testing

Concepts: Hypothesis testing for comparing two proportions. Recognizing situations that call for comparing proportions versus comparing means. Conditional versus unconditional means.

Case studies: We continue with Karlan and List (2007) from Module C.2.

Required readings: Section 10.5, Chapter 12, and the background on Karlan and List (2007) “Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment” in Module C.2 starting on page 81. Next, review this excerpt of Table 2A and the discussion of it.

TABLE 2A—MEAN RESPONSES
(Mean and standard errors)

	Control	Treatment	Match ratio		
			1:1	2:1	3:1
Implied price of \$1 of public good:	1.00	0.36	0.50	0.33	0.25
<i>Panel A</i>	(1)	(2)	(3)	(4)	(5)
Response rate	0.018 (0.001)	0.022 (0.001)	0.021 (0.001)	0.023 (0.001)	0.023 (0.001)
Dollars given, unconditional	0.813 (0.063)	0.967 (0.049)	0.937 (0.089)	1.026 (0.089)	0.938 (0.077)
Dollars given, conditional on giving	45.540 (2.397)	43.872 (1.549)	45.143 (3.099)	45.337 (2.725)	41.252 (2.222)
Dollars raised per letter, not including match	0.81	0.97	0.94	1.03	0.94
Dollars raised per letter, including match	0.81	2.90	1.87	3.08	3.75
Observations	16,687	33,396	11,133	11,134	11,129
<i>Panel B: Blue states</i>					
Response rate	0.020 (0.001)	0.021 (0.001)	0.021 (0.002)	0.022 (0.002)	0.021 (0.002)

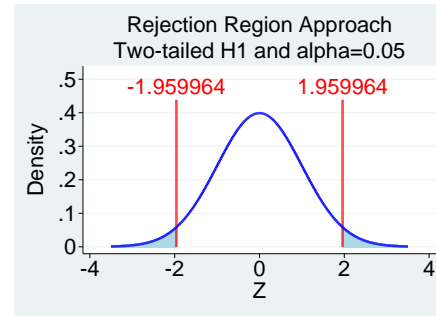
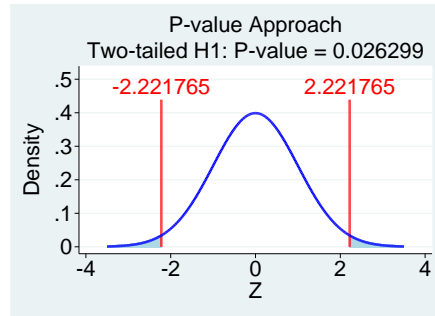
Figure of Table 2A: Karlan and List (2007), p. 1781, Panel A and first row of Panel B only.

- Consider the first row of results in Panel A, Columns (1) and (2) in Table 2A. Is there a statistically significant difference in the response rates between the control and treatment group?
 - Translating to formal hypotheses and standard notation yields $H_0 : (p_T - p_C) = 0$ versus $H_1 : (p_T - p_C) \neq 0$, with T for treatment group and C for control group.
 - For inferences about *proportions*, use z test statistics and the Normal distribution.
 - From our aid sheets: $z = \frac{(\hat{P}_2 - \hat{P}_1) - 0}{\sqrt{\frac{\bar{P}(1-\bar{P})}{n_2} + \frac{\bar{P}(1-\bar{P})}{n_1}}}$, where $\bar{P} = \frac{X_1 + X_2}{n_1 + n_2}$. \bar{P} is the pooled proportion.

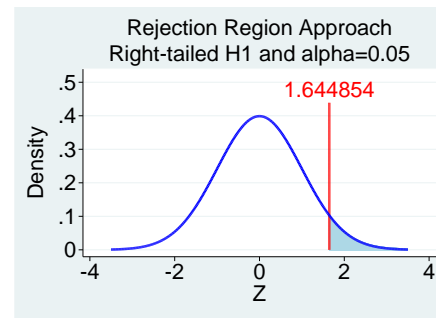
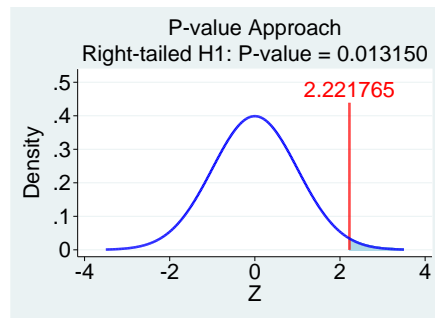
Because H_0 says there is no difference between the population proportions and hypothesis testing starts with the presumption the null is true, we pool the two samples together for a single estimate (\bar{P}). Notice that we do *not* use \bar{P} for *confidence interval estimation*.

 - Conventional significance levels are an α of 0.01, 0.05, or 0.10 (and sometimes $\alpha = 0.001$).
 - Two methods of hypothesis testing: P-value approach and rejection region approach:
 - * With the P-value approach, use the test statistic and the direction of the research hypothesis ($>$, $<$, \neq) to compute a probability called the P-value. The smaller the P-value, the stronger the evidence *in favor of* the research hypothesis (and *against* the null). If the P-value is less than α , the result is statistically significant at that α .

- * With the rejection region (aka critical value) approach, use the significance level (α) and the direction of the research hypothesis ($>$, $<$, \neq) to compute the critical value (edge of the rejection region). Compare your test statistic with the critical value.
- For example, test $H_0 : (p_T - p_C) = 0$ versus $H_1 : (p_T - p_C) \neq 0$. Of the 1,000 people in the treatment group, 80 donate. Of the 3,000 people in the control group, 180 donate. $\bar{P} = \frac{80+180}{1,000+3,000} = 0.065$ and $z = \frac{(0.08-0.06)-0}{\sqrt{\frac{0.065(1-0.065)}{1,000} + \frac{0.065(1-0.065)}{3,000}}} = \frac{0.02}{0.00900185} = 2.221765$.



- * The small P-value of 0.026299 (left graph) supports H_1 . It is less than 0.05, but bigger than 0.01: the best conventional significance level we meet is $\alpha = 0.05$.
- * The z test statistic of 2.221765 falls in the rejection region (right graph) for $\alpha = 0.05$.
- * With either the P-value or rejection region approach, there is a statistically significant difference in the fraction donating (control vs. treatment) at a 5% significance level.
- What if testing $H_0 : (p_T - p_C) = 0$ versus $H_1 : (p_T - p_C) > 0$ (a one-tailed test)?



- After “Response rate,” Table 2A on page 88 lists “Dollars given, unconditional” and “Dollars given, conditional on giving” in Panel A. Unconditional means using all observations, including those who gave zero, when computing the mean. Conditional means ignoring the zeros. Note that amount given is an interval variable, in contrast to donate or not, which is a categorical (nominal) variable. But *why* do we need *three* rows? Consider three *different* questions.
 1. Overall, does the charity raise more money by offering a match? If so, how much?
 2. Does offering a match change the chance someone will donate? If so, how much?
 3. Does offering a match make the donations received more generous? If so, how much?
- “Dollars given, unconditional” addresses question 1. “Dollars given, conditional” addresses question 3. “Response rate” addresses question 2. Question 1 combines 2 and 3. All three questions are interesting. We’d like to know if the mean goes up because more people are giving (averaging in fewer zeros) and/or because people make more generous contributions.

- Table 2A shows making inferences about *both* differences in proportions (categorical variables) and differences in means (interval variables).
 - * To answer question 2, make an inference about the difference in proportions ($p_T - p_C$): textbook Sections 11.6 and 12.8. Start with the sample proportions ($\hat{P}_T - \hat{P}_C$). Use hypothesis testing to assess if there is a difference (yes or no). Use confidence interval (CI) estimation to assess how large the difference is.
 - * To answer questions 1 and 3, make an inference about the difference in means ($\mu_T - \mu_C$) with independent samples: textbook Sections 14.1 - 14.5. Start with the sample means ($\bar{X}_T - \bar{X}_C$). You also need the sample standard deviations, s_T and s_C , to compute standard errors. When computing the sample statistics, either use the subset of the data with a positive (non-zero) donation or all of the data, depending on whether you are doing a conditional comparison. Like above, hypothesis testing answers the first part of the questions (yes/no) and CI estimation answers the second part (how much).

Datasets: For Karlan and List (2007): [char_give.xlsx](#)

Interactive module materials for Module C.3:

1. For Karlan and List (2007), use [char_give.xlsx](#). Continuing with part 2 of Module C.2 on page 83, **create** a *new* variable for blue_state that records the category as a string (text). Name the new variable state_type. (Note: If you did not already create the variables group and give_status in part 2 of Module C.2, do that first by following the Excel tip on page 83.)

EXCEL TIPS: The blue_state variable is tricky because there are some missing values. However, if Column Q has the variable blue_state, make a new variable named state_type with the function `=IF(Q2=1,"Blue State",IF(ISBLANK(Q2),"Missing","Red State"))`. This is a nested if statement. If cell Q2 is 1 it evaluates to “Blue State.” If cell Q2 is not 1, it evaluates to “Missing” if cell Q2 is blank and to “Red State” otherwise. Note the use of the logical function ISBLANK, which returns a value of TRUE if cell Q2 is blank and FALSE otherwise. Add this variable onto the original data (i.e. put it in Column V).

2. **Consider** the hypothesis test $H_0 : (p_T - p_C) = 0$ versus $H_1 : (p_T - p_C) \neq 0$ for the first row of results in Panels A and B, Columns (1) and (2) in Table 2A on page 88.

- (a) For Panel A, **find** the values of X_T , n_T , X_C , and n_C in $\hat{P}_T = \frac{X_T}{n_T}$ and $\hat{P}_C = \frac{X_C}{n_C}$.

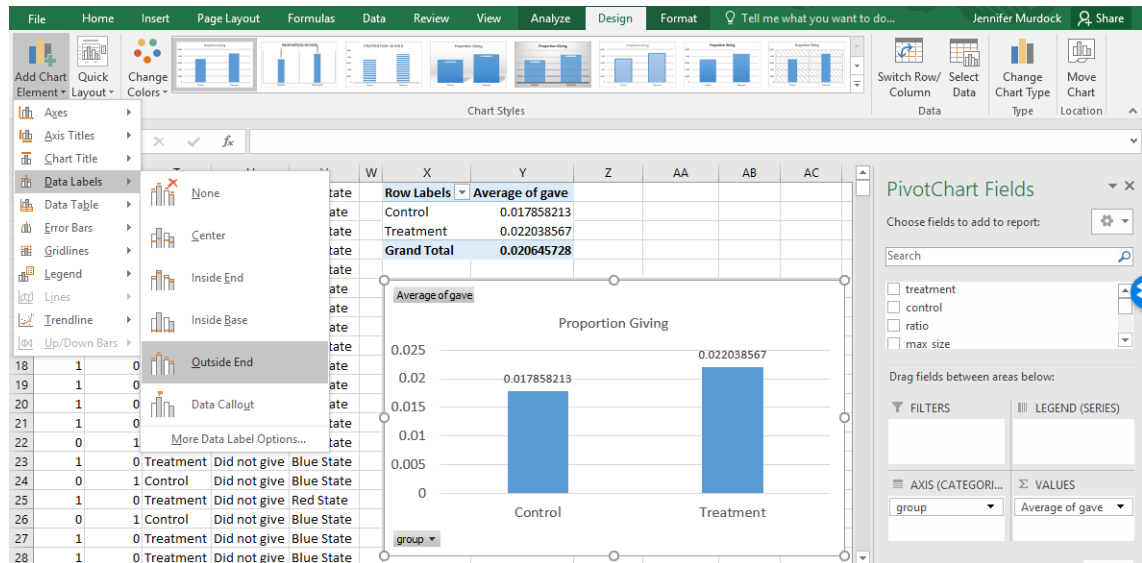
EXCEL TIPS: Use a PivotTable (see below).

	Q	R	S	T	U	V	W	X	Y	Z	AA
1	blue_state	red_cty	blue_cty	group	give_status	state_type					
2	1	0	1	Control	Did not give	Blue State		Count of give_status	group		
3	1	1	0	Control	Did not give	Blue State		give_status	Control	Treatment	Grand Total
4	1	0	1	Treatment	Did not give	Blue State		Did not give	16389	32660	49049
5	1	1	0	Treatment	Did not give	Blue State		Gave	298	736	1034
6	0	0	1	Treatment	Did not give	Red State		Grand Total	16687	33396	50083
7	0	1	0	Control	Did not give	Red State					
8	0	1	0	Treatment	Did not give	Red State					
9	1	0	1	Treatment	Did not give	Blue State					
10	0	1	0	Treatment	Did not give	Red State					
11	1	0	1	Treatment	Did not give	Blue State					
12	1	1	0	Treatment	Did not give	Blue State					
13	1	1	0	Treatment	Did not give	Blue State					
14	0	1	0	Control	Did not give	Red State					
15	0	1	0	Treatment	Did not give	Red State					
16	0	1	0	Control	Did not give	Red State					

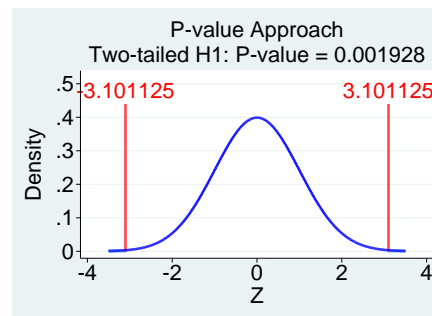
PivotTable Fields
 Choose fields to add to report:
 Search
☒ group
☒ give_status
☐ state_type
 Drag fields between areas below:
 FILTERS
 COLUMNS
 ROWS
 VALUES
 give_status
 Count of give...

- (b) **Make** a figure comparing the response rate for the control and treatment groups. This replicates the point estimates in Columns (1) and (2) of the first row in Panel A.

EXCEL TIPS: Use a PivotChart. Add data labels, with the heights of the bars, as shown below.



- (c) **Recall** the background starting on page 88, which includes the following formulas that you will need in the next step: $z = \frac{(\hat{P}_2 - \hat{P}_1) - 0}{\sqrt{\frac{\bar{P}(1-\bar{P})}{n_2} + \frac{\bar{P}(1-\bar{P})}{n_1}}}$ where $\bar{P} = \frac{X_1 + X_2}{n_1 + n_2}$.
- (d) **Compute** the P-value for the hypothesis test $H_0 : (p_T - p_C) = 0$ versus $H_1 : (p_T - p_C) \neq 0$. **Verify** that you obtain: $\bar{P} = 0.02064573$, $(\hat{P}_T - \hat{P}_C) = 0.00418035$, a SE of the difference presuming H_0 is true of 0.00134801, and a z test statistic of 3.1011251. **Verify** that you obtain a P-value of 0.00192787. For the P-value, remember it's a *two-tailed* test: see the figure below.



EXCEL TIPS: Use the worksheet “HT Diff. in Resp. Rates” to organize your work. Recall the NORM.S.DIST function (page 84).

Interpretation tips: What does 0.00192787 mean? Comparing the control and treatment groups in the charitable giving field experiment, there is a statistically significant difference, at all conventional significance levels including $\alpha = 0.01$, in the proportion donating. With this tiny P-value, we have strong evidence supporting the conclusion that including a match offer in the letter causes a change in the response rate.

- (e) For Panel B in Table 2A, which repeats the analysis using *only* the subset of the data in

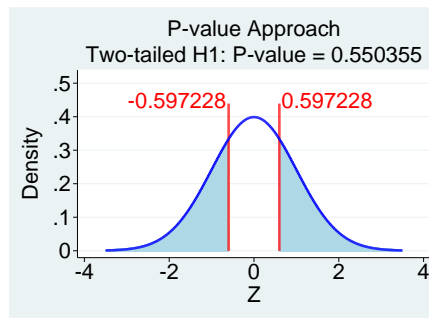
Blue states, **find** the values of X_T , n_T , X_C , and n_C in $\hat{P}_T = \frac{X_T}{n_T}$ and $\hat{P}_C = \frac{X_C}{n_C}$. **Verify** that you obtain $X_T = 417$, $n_T = 19,777$, $X_C = 201$, and $n_C = 10,029$.

EXCEL TIPS: Use a PivotTable: drag the variable state_type to COLUMNS, drag the variables group and give_status to ROWS, and drag another copy of the variable group to Σ VALUES.

Count of group				
Row Labels	Blue State	Missing	Red State	Grand Total
Control	10029	10	6648	16687
Did not give	9828	10	6551	16389
Gave	201		97	298
Treatment	19777	25	13594	33396
Did not give	19360	24	13276	32660
Gave	417	1	318	736
Grand Total	29806	35	20242	50083

PivotTable Fields				
Choose fields to add to report:				
Search				
<input type="checkbox"/>	red_cty			
<input type="checkbox"/>	blue_cty			
<input checked="" type="checkbox"/>	group			
<input checked="" type="checkbox"/>	give_status			
<input checked="" type="checkbox"/>	state_type			
Drag fields between areas below:				
FILTERS		COLUMNS		
		state_type		
ROWS		VALUES		
group		Count of group		
give_status				

- (f) Continuing, in Blue states, is there a statistically significant difference in the response rates between the control and treatment group? If so, at which conventional significance levels? In answering, **compute** the P-value. **Verify** that you obtain: $\bar{P} = 0.02073408$, $(\hat{P}_T - \hat{P}_C) = 0.00104322$, a SE of the difference presuming the null is true of 0.00174677, a z test statistic of 0.59722846, and a P-value of 0.55035486.



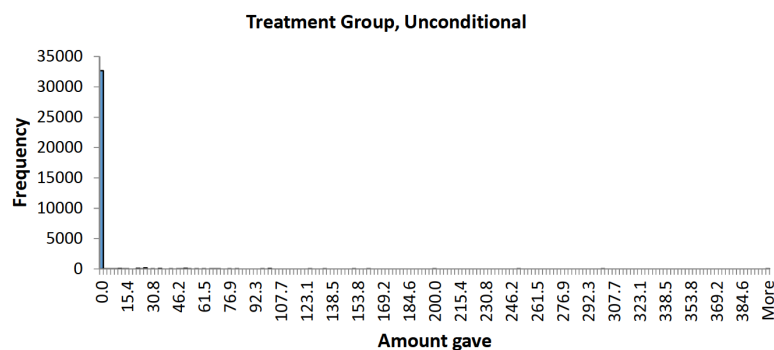
EXCEL TIPS: Make a copy of the worksheet “HT Diff. in Resp. Rates” and update it.

Interpretation tips: What does 0.55035486 mean? Focusing on the subset of the sample living in Blue states (lean toward the U.S. Democratic party), there is not a statistically significant difference at any conventional significance level, including $\alpha = 0.10$, in the proportion donating between the control and treatment groups. With this huge P-value, we definitely do *not* have the necessary evidence to support the conclusion that including a match offer in the letter causes a change in the response rate in Blue states.

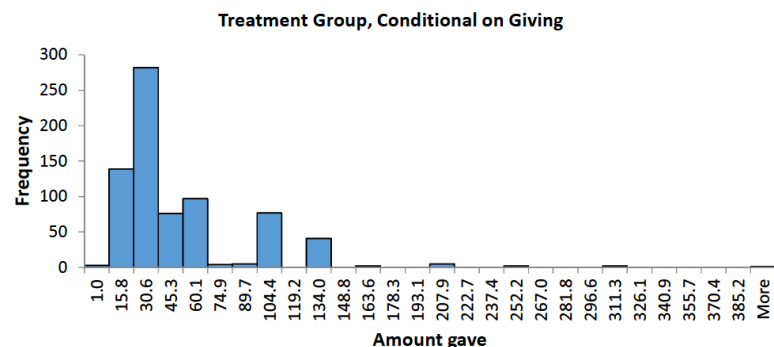
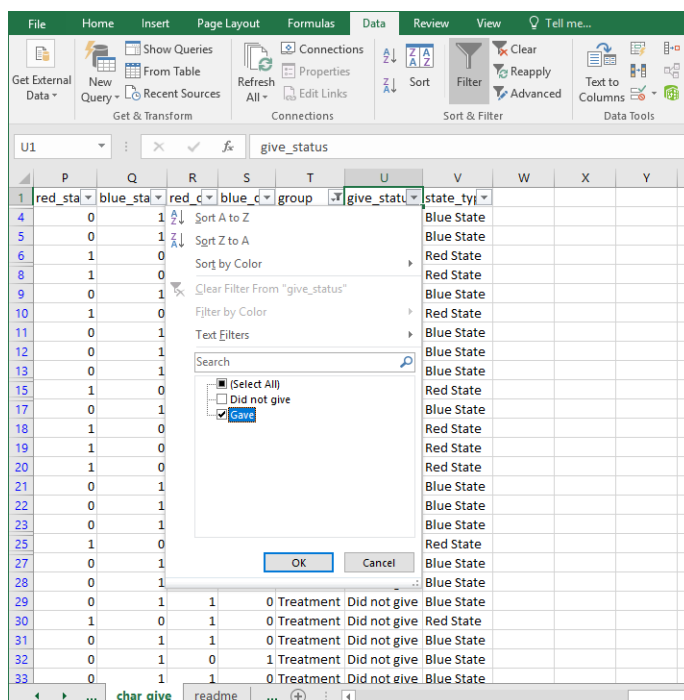
- In Table 2A, notice again the rows “Dollars given, unconditional” and “Dollars given, conditional on giving.” To illustrate the huge difference between unconditional and conditional, **construct** two histograms of the amount given for the treatment group: one unconditional and the other conditional on giving something. **Verify** they are similar to those below.

EXCEL TIPS: Use the Filter tool and make a new worksheet for each case (unconditional and conditional). For the “unconditional” case, just Filter by group (selecting Treatment). (See

page 43 or 52 for a refresher.) Select Histogram under Data Analysis. (Refresher on page 22.)



EXCEL TIPS: For the “conditional” case, again Filter by group (selecting Treatment) and then Filter by give_status (selecting Gave).



- Next, replicate the second and third rows of results in Panel A of Table 2A on page 88 for the control group and treatment group.

EXCEL TIPS: A single PivotTable can do nearly all of the work for this replication. Insert a PivotTable in a new worksheet including the variables amount, group, and give_status. Drag

the group variable to COLUMNS, drag the give_status variable to ROWS, and drag the variable amount to Σ VALUES. If you chose the entire columns (not just the range of observations), in the PivotTable fields environment, unselect blanks (there are no missing values for any of these three variables). From the PivotTable fields environment, drag a second and third copy of the variable amount to Σ VALUES (for the sample mean and the sample standard deviation). To make the chart even more readable, drag the Σ Values field (this field is just called Values in a mac) to rows instead of columns.

	T	U	V	W	X	Y	Z	AA	AB
1	amount	group	give_status	state_type					
2		0	Control	Did not give	Blue State				
3		0	Control	Did not give	Blue State				
4		0	Treatment	Did not give	Blue State				
5		0	Treatment	Did not give	Blue State				
6		0	Treatment	Did not give	Red State				
7		0	Control	Did not give	Red State				
8		0	Treatment	Did not give	Red State				
9		0	Treatment	Did not give	Blue State				
10		0	Treatment	Did not give	Red State				
11		0	Treatment	Did not give	Blue State				
12		0	Treatment	Did not give	Blue State				
13		0	Treatment	Did not give	Blue State				
14		0	Control	Did not give	Red State				
15		0	Treatment	Did not give	Red State				
16		0	Control	Did not give	Red State				
17		0	Treatment	Did not give	Blue State				
18		0	Treatment	Did not give	Red State				
19		0	Treatment	Did not give	Red State				
20		0	Treatment	Did not give	Red State				

Row Labels	Control	Treatment	Grand Total
Did not give			
Average of amount	0	0	0
StdDev of amount2	0	0	0
Count of amount3	16389	32660	49049
Gave			
Average of amount	45.54026846	43.871875	44.35270793
StdDev of amount2	41.3798214	42.01611301	41.82056653
Count of amount3	298	736	1034
Total Average of amount	0.813267813	0.966873278	0.915693948
Total StdDev of amount2	8.17648194	8.963208795	8.709199288
Total Count of amount3	16687	33396	50083

(a) After implementing the Excel tip above, reference your results to answer the following.

- Find** the unconditional mean of \$0.813 for the control group and **verify** that 16,687 observations are used to compute that mean. **Find** the unconditional *standard deviation* (not the standard error) for the control group and **verify** it is $s = \$8.176482$.

EXCEL TIPS: Just read the values from the PivotTable you produced.

- **Replicate** the standard error of \$0.063 using Excel to program this formula from our aid sheets: $\frac{s}{\sqrt{n}}$.
- Find** the conditional mean of \$45.540 for the control group and **verify** that 298 observations are used to compute that mean. **Find** the conditional *standard deviation* for the control group and **verify** it is $s = \$41.37982$.
 - **Replicate** the standard error of \$2.397.
 - Find** the unconditional mean of \$0.967 for the treatment group and **verify** that 33,396 observations are used to compute that mean. **Find** the unconditional *standard deviation* for the treatment group and **verify** it is $s = \$8.963209$.
 - **Replicate** the standard error of \$0.049.
 - Find** the conditional mean of \$43.872 for the treatment group and **verify** that 736 observations are used to compute that mean. **Find** the conditional *standard deviation* for the treatment group and **verify** it is $s = \$42.01611$.
 - **Replicate** the standard error of \$1.549.

Test/exam examples: For many examples related to Karlan and List (2007), see page 87.

C.0.0 Practice questions for Module C

- Q1.** Use [mod_c_sims.xlsx](#). Worksheets “Simulation 1” and “Simulation 2” contain two separate simulations. In each, 10,000 samples ($m = 10,000$), each with a sample size of 30 ($n = 30$), are drawn from undisclosed populations. For example, the first row of data in “Simulation 1” shows the first random sample of 30 observations drawn from Population 1. As another example, the last row of data in “Simulation 2” shows the ten-thousandth random sample of 30 observations drawn from Population 2. Sample statistics are computed as usual (for example, $\bar{X} = \frac{\sum_{i=1}^{30} x_i}{30}$).
- (a) Using worksheet “Simulation 1,” make an inference about the shape of Population 1.
 - (b) Using worksheet “Simulation 1,” construct the simulated sampling distribution of \bar{X} for a sample size of 30 ($n = 30$) using 10,000 random samples ($m = 10,000$). What is the shape of the sampling distribution of \bar{X} ? Using your simulation results, what is the mean of \bar{X} and the s.d. of \bar{X} ?
 - (c) Repeat the previous part, but with only 1,000 random samples ($m = 1,000$) (use the first 1,000). How does this change your answers to the previous part?
 - (d) Using worksheet “Simulation 1,” construct the simulated sampling distribution of the sample MEDIAN for a sample size of 3 ($n = 3$) (the first 3 of the 30) using 10,000 random samples ($m = 10,000$). What is the shape of the sampling distribution of the sample median? Using your simulation results, what is the mean of the sample median and the s.d. of the sample median?
- Q2.** Recall ON Univ. & Col. (2016). Use the worksheet “n=25,m=10000” in [on_univ_col_16.xlsx](#). It gives 10,000 samples ($m = 10,000$), each with a sample size of 25 ($n = 25$), drawn from the population of all ON public sector employees in the universities or colleges sector with a 2016 salary of at least \$100K.
- (a) Using worksheet “n=25,m=10000,” construct the simulated sampling distribution of \bar{X} for a sample size of 16 ($n = 16$) (the first 16 of the 25) using 10,000 random samples ($m = 10,000$). What is the shape of the sampling distribution of \bar{X} ? Using your simulation results (not theory), what is the mean of \bar{X} and the s.d. of \bar{X} ?
 - (b) Sometimes we struggle to visually detect skew. For the simulated sampling distribution of \bar{X} for $n = 16$, let’s make the determination of skew more quantitative and less subjective.
 - i. What fraction of standardized data lie between -1 and 0 for a perfect Normal distribution? How about 0 and 1? (Use the appropriate Excel function.)
 - ii. In the simulated sampling distribution of the sample mean for $n = 16$, what fraction of the standardized sample means lie between -1 and 0? How about 0 and 1?
 - iii. What fraction of standardized data lie between -2 and -1 for a perfect Normal distribution? How about 1 and 2? (Use the appropriate Excel function.)
 - iv. In the simulated sampling distribution of the sample mean for $n = 16$, what fraction of the standardized sample means lie between -2 and -1? How about 1 and 2?
 - v. What fraction of standardized data lie between -2 and -3 for a perfect Normal distribution? How about 2 and 3? (Use the appropriate Excel function.)
 - vi. In the simulated sampling distribution of the sample mean for $n = 16$, what fraction of the standardized sample means lie between -3 and -2? How about 2 and 3?

- Q3.** Recall Karlan and List (2007) and use [char_give.xlsx](#). Replicate Table C.1, which shows the experimental design. How many people are randomly assigned to the control group? How many to a treatment group? How many are randomly assigned to the treatment group that receives a letter that offers a 2 to 1 match ratio, states that the maximum size of the matching grant is \$50,000, and shows a low match example amount?
- Q4.** Recall Karlan and List (2007) and use [char_give.xlsx](#). Review Table 1.

TABLE 1—SUMMARY STATISTICS—SAMPLE FRAME
(Mean and standard deviations)

	All (1)	Treatment (2)	Control (3)
<i>Member activity</i>			
Number of months since last donation	13.007 (12.081)	13.012 (12.086)	12.998 (12.074)
Highest previous contribution	59.385 (71.177)	59.597 (73.052)	58.960 (67.269)
Number of prior donations	8.039 (11.394)	8.035 (11.390)	8.047 (11.404)
Number of years since initial donation	6.098 (5.503)	6.078 (5.442)	6.136 (5.625)
Percent already donated in 2005	0.523 (0.499)	0.523 (0.499)	0.524 (0.499)
Female	0.278 (0.448)	0.275 (0.447)	0.283 (0.450)
Couple	0.092 (0.289)	0.091 (0.288)	0.093 (0.290)
... [Part of the table has been excluded] ...			
<i>State and county</i>			
Red state—proportion that live in red state	0.404 (0.491)	0.407 (0.491)	0.399 (0.490)
Red county—proportion that live in red county	0.510 (0.500)	0.512 (0.500)	0.507 (0.500)

Figure of Table 1: Karlan and List (2007), p. 1778.

- (a) Replicate the first two numbers in Column (1) of Table 1: 13.007 and (12.081).
- (b) What if Table 1 included Column (1a) summarizing all who *did give* a donation in this campaign and Column (1b) for all who *did not give* a donation in this campaign. Compute the numbers that fill in the blanks: the first two numbers in Column (1a) would be _____ and (_____) and the first two numbers in Column (1b) would be _____ and (_____).
- (c) Replicate the two numbers in Column (3) of Table 1 in the row “Red state – proportion that live in red state” under the heading “*State and county*”: 0.399 and (0.490).
- (d) What if the first row of Table 1 under “*State and county*” were “Blue state – proportion that live in blue state”? Compute the numbers that fill in the blanks: the values in Column (3) would be _____ and (_____).
- Q5.** Recall Karlan and List (2007) and use [char_give.xlsx](#). Compute the **98%** confidence interval estimate of the difference in the response rate between the treatment group offered a 1 to 1 match versus those offered a 3 to 1 match. Report the point estimate of the difference, the standard error of the difference, and the margin of error of the difference. Further, fill in the blanks: we are 98% confident that the response rate among *all* previous donors to this charity would be between _____ percentage points lower to _____ percentage points higher if a 3 to 1 match

were offered instead of a 1 to 1 match. This is a wide interval that spans the possibilities that a higher match ratio substantially hurts our response rate to substantially helps our response rate. It is safe to say that a 3 to 1 match offers no clear benefit in response rates and may even be detrimental to response rates compared to a more modest 1 to 1 match.

- Q6.** Recall Karlan and List (2007) and use [char_give.xlsx](#). Review Table 2B from Karlan and List (2007), p. 1782, which is reproduced below.

TABLE 2B—MEAN RESPONSES
(Mean and standard errors)

	Match							
	Threshold					Example amount		
	Control	\$25,000	\$50,000	\$100,000	Unstated	Low	Medium	High
Implied price of \$1 of public good:	1.00	0.36	0.36	0.36	0.36	0.36	0.36	0.36
<i>Panel A</i>	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Response rate	0.018 (0.001)	0.022 (0.002)	0.022 (0.002)	0.022 (0.002)	0.022 (0.002)	0.021 (0.001)	0.022 (0.001)	0.023 (0.001)
Dollars given, unconditional	0.813 (0.063)	1.060 (0.109)	0.889 (0.091)	0.903 (0.084)	1.015 (0.106)	0.914 (0.080)	1.004 (0.091)	0.983 (0.084)
Dollars given, conditional on giving	45.540 (2.397)	49.172 (3.522)	39.674 (2.900)	41.000 (2.336)	45.815 (3.475)	43.107 (2.557)	45.239 (2.932)	43.251 (2.542)
Dollars raised per letter, not including match	0.81	1.06	0.89	0.90	1.01	0.91	1.00	0.98
Dollars raised per letter, including match	0.81	3.32	2.63	2.65	2.99	2.83	2.92	2.96
Observations	16,687	8,350	8,345	8,350	8,351	11,134	11,133	11,129
<i>Panel B: Blue states</i>								
Response rate	0.020 (0.001)	0.020 (0.002)	0.022 (0.002)	0.022 (0.002)	0.020 (0.002)	0.019 (0.002)	0.022 (0.002)	0.022 (0.002)
Dollars given, unconditional	0.897 (0.086)	0.884 (0.115)	0.912 (0.127)	0.900 (0.110)	0.884 (0.116)	0.796 (0.094)	0.950 (0.108)	0.939 (0.102)
Dollars given, conditional on giving	44.781 (2.914)	43.204 (3.716)	41.091 (4.227)	41.236 (3.093)	44.469 (3.806)	41.516 (3.283)	43.194 (3.364)	42.503 (3.063)
Dollars raised per letter, not including match	0.90	0.88	0.91	0.90	0.88	0.80	0.95	0.94
Dollars raised per letter, including match	0.90	2.83	2.72	2.50	2.60	2.38	2.78	2.82
Observations	10,029	5,035	4,954	4,856	4,932	6,574	6,550	6,653
<i>Panel C: Red states</i>								
Response rate	0.015 (0.001)	0.023 (0.003)	0.023 (0.003)	0.022 (0.002)	0.025 (0.003)	0.024 (0.002)	0.022 (0.002)	0.024 (0.002)
Dollars given, unconditional	0.687 (0.093)	1.330 (0.212)	0.856 (0.127)	0.874 (0.124)	1.206 (0.199)	1.086 (0.141)	1.082 (0.158)	1.023 (0.141)
Dollars given, conditional on giving	47.113 (4.232)	57.156 (6.485)	37.649 (3.643)	39.584 (3.462)	47.330 (6.039)	44.929 (4.005)	48.097 (5.234)	43.519 (4.318)
Dollars raised per letter, not including match	0.69	1.33	0.86	0.87	1.21	1.09	1.08	1.02
Dollars raised per letter, including match	0.69	4.08	2.51	2.80	3.57	3.48	3.11	3.11
Observations	6,648	3,309	3,385	3,487	3,413	4,549	4,579	4,466

- Replicate the two numbers in Panel A, Column (2) of Table 2B in the row “Response rates.” Record your answers accurate to at least five decimal places.
- Replicate the two numbers in Panel C, Column (2) of Table 2B in the row “Response rates.” Record your answers accurate to at least five decimal places.
- Compute the 95% confidence interval estimate of the difference in the response rate between those whose letter illustrated the match with a high example amount (a big donation) versus those whose letter illustrated the match with a low example amount (a more modest

donation). Report the point estimate of the difference, the standard error of the difference, the margin of error of the difference, and the LCL and UCL.

- (d) Compute the P-value in testing if there is a statistically significant difference in the response rate between those whose letter illustrated the match with a high example amount (a big donation) versus those whose letter illustrated the match with a low example amount (a more modest donation).

Q7. Recall Carlin et al. (2017) from Module A.1 about credit card choice. Another researcher repeats a simplified version of the experiment. In the simplified version, all participants watch the full instructional video (“implemental video”) about choosing a credit card. Of the 2,201 participants, 1,104 see the four credit card offers without misleading advertising (“no taglines”). The remaining 1,097 see the four credit card choices with misleading advertising (“superfluous taglines”). The researcher correctly computes and interprets the 98% confidence interval. S/he is 98% confident that misleading ads reduce the proportion picking the best credit card by 10.3 to 20.1 percentage points, which is a large negative impact on people’s ability to make good choices among credit cards. If 483 people in the superfluous taglines group picked the best card, then how many people in the no taglines group picked the best card?

For extra practice, additional questions, with an ^e superscript (e for extra), are next.

Q^e1. Recall the background given with question 1 on page 95 and use [mod_c_sims.xlsx](#).

- Using worksheet “Simulation 2,” make an inference about the shape of Population 2.
- Using worksheet “Simulation 2,” construct the simulated sampling distribution of \bar{X} for a sample size of 3 ($n = 3$) (the first 3 of the 30) using 10,000 random samples ($m = 10,000$). What is the shape of the sampling distribution of \bar{X} ? Using your simulation results, what is the mean of \bar{X} and the s.d. of \bar{X} ?
- Repeat the previous part, but with a sample size of 30 ($n = 30$). How does this change your answers to the previous part?
- Using worksheet “Simulation 2,” construct the simulated sampling distribution of the sample STANDARD DEVIATION for a sample size of 15 ($n = 15$) (the first 15 of the 30) using 10,000 random samples ($m = 10,000$). What is the shape of the sampling distribution of the sample standard deviation? Using your simulation results, what is the mean of the sample standard deviation and the s.d. of the sample standard deviation?

Q^e2. Recall the dice example (Section 10.3, which Module C.1 reviewed). Use [mod_c_dice_roll.xlsx](#).

- How can we use `RAND()`, which gives a random draw from a Uniform distribution with $a = 0$ and $b = 1$, to simulate the roll of a die? In other words, how can we translate a continuous Uniform distribution that ranges from 0 to 1 to the discrete outcome from the roll of a die that ranges from 1 to 6? Use the worksheet “Active Die Roll” to verify that `=ROUND((0.5+6*RAND()),0)` provides a solution. Using these same ideas, what would be the Excel function to generate tosses of a fair coin if we mark a head as 1 and a tail as zero? Similarly, what about an unfair coin with a 40% chance of heads?

- (b) Use the worksheet “n=20, m=10,000” to replicate the “Simple Die Toss” figure at the beginning of Module C.1, which shows the simulated sampling distribution of \bar{X} for $n = 1$ (the first 1 of the 20). How many unique values of \bar{X} occur in your simulation? How frequent is a value between 3 and 4 including those endpoints?
- (c) Use the worksheet “n=20, m=10,000” to replicate the “Three-Dice Average” figure at the beginning of Module C.1, which shows the simulated sampling distribution of \bar{X} for $n = 3$ (the first 3 of the 20). How many unique values of \bar{X} occur in your simulation? How frequent is a value between 3 and 4 including those endpoints?
- (d) Use the worksheet “n=20, m=10,000” to replicate the “20-Dice Average” figure at the beginning of Module C.1, which shows the simulated sampling distribution of \bar{X} for $n = 20$. How many unique values of \bar{X} occur in your simulation? How frequent is a value between 3 and 4 including those endpoints?

Q^e3. Recall question 2 on page 95 and use the worksheet “n=25,m=10000” in [on_univ_col.16.xlsx](#).

- (a) Using worksheet “n=25,m=10000,” construct the simulated sampling distribution of the sample median for a sample size of 16 ($n = 16$) (the first 16 of the 25) using 10,000 random samples ($m = 10,000$). What is the shape of the sampling distribution of the sample median? Using your simulation results (not theory), what is the mean of the sample median and the s.d. of the sample median?
- (b) According to the simulation results, which of the two sample statistics – the sample mean (from question 2) or the sample median – is less affected by sampling error as judged by the standard deviation? As judged by the range?

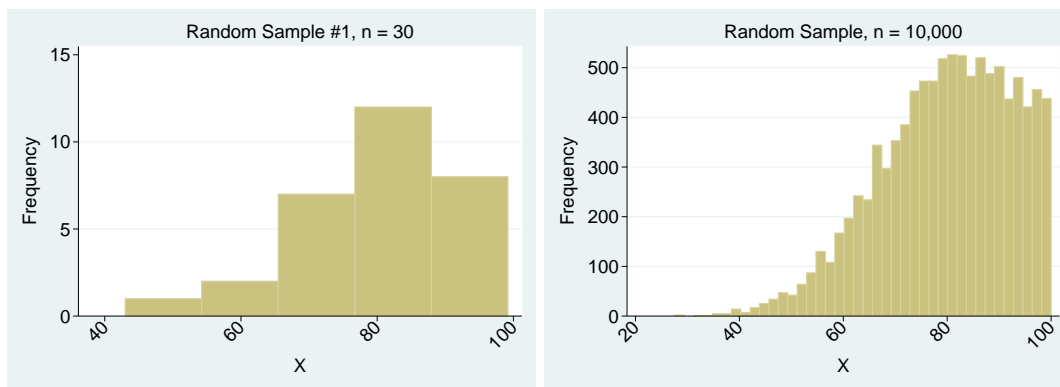
Q^e4. Recall ON Univ. & Col. (2016). Use the worksheet “Drawing random samples” in [on_univ_col.16.xlsx](#).

A method of selecting random samples discussed is to generate a column of random numbers, sort by it, and then select the first n observations. Consider a random sample of size $n = 1,000$. A simulation would require drawing many (e.g. $m = 10,000$) such samples. The combination of $n = 1,000$ and $m = 10,000$ is too cumbersome in Excel. However, doing the first few of the 10,000 draws is no problem. Use the column “Random Example (Values)” to complete the first row of the data file below. Create your own columns of random numbers for the second and third rows using the RAND function and copy-and-paste values.

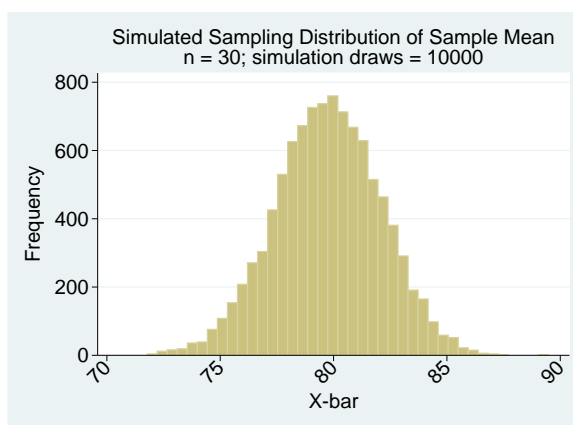
Draw #	Sample Mean
1	
2	
3	
...	...
9,999	
10,000	

Answers for Module C practice questions:

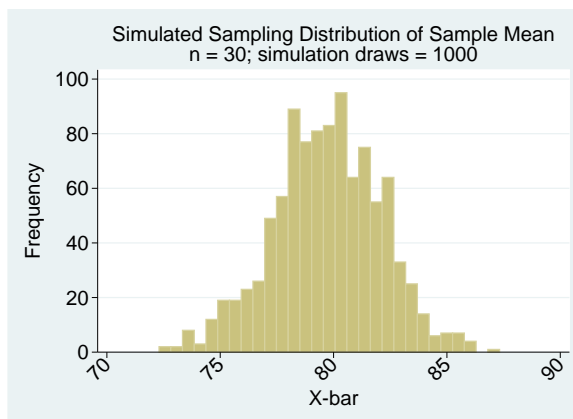
- A1.** (a) You could make a rough inference using Sample #1. However, in worksheet “Simulation 1” we have 300,000 ($=30 \times 10,000$) observations randomly drawn from Population 1 so it is easy to make an inference about that population using this very large sample. In fact, we do not need to use all 300,000 observations. For ease, we can use the first 10,000 draws from the population (i.e. x_1). We can confidently infer with this large sample and the clear pattern in the histogram that the population is negatively skewed.



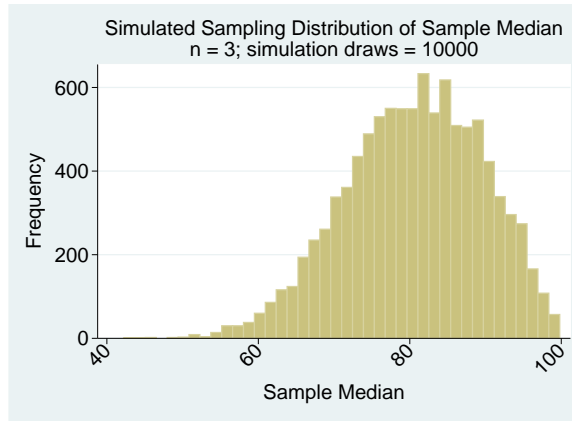
- (b) The mean of the 10,000 sample means is 79.7475 and the s.d. of the 10,000 sample means is 2.327985. The shape looks Normal (see histogram below).



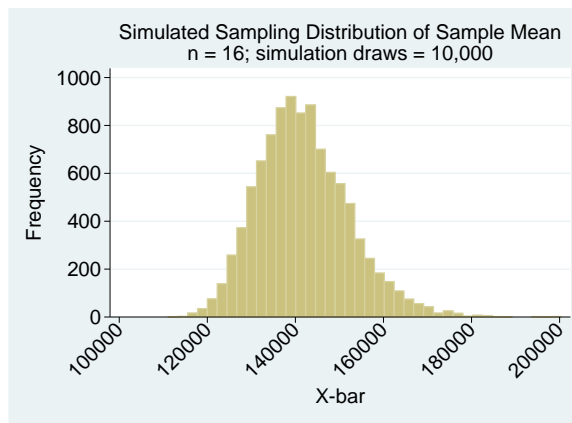
- (c) It does not change our answers in any meaningful way. The shape looks the same and the mean and s.d. of the sample mean are comparable. The mean of the 1,000 sample means is 79.71958 and the s.d. of the 1,000 sample means is 2.383696.



- (d) The mean of the 10,000 sample medians is 80.54002 and the s.d. of the 10,000 sample medians is 9.037415. The shape is somewhat negatively skewed (see histogram below).



- A2.** (a) The mean of the 10,000 sample means is 141882.7 and the s.d. of the 10,000 sample means is 10404.93. The shape is somewhat positively skewed.



- (b) i. $=\text{NORM.S.DIST}(1,\text{TRUE})-\text{NORM.S.DIST}(0,\text{TRUE}) = 0.341344746$ (Given the symmetry of the Normal distribution: $P(-1 < Z < 0) = P(0 < Z < 1)$.)
- ii. 0.3819 between -1 and 0 and 0.3157 between 0 and 1. There are many ways to work this out. All start with computing the sample mean for a sample size of 16: for example, for the first simulation draw use `=AVERAGE(B2:Q2)` and copy through for all 10,000 simulation draws. All approaches also require computing the mean and standard deviation of that new column of 10,000 means and then creating a new column that gives the standardized (Z score) values created by subtracting that mean and dividing by the standard deviation. For example, `=(AA2-AA10003)/AA10004` if the cell AA10003 has the mean of the sample means (which comes out to 141882.7213) and AA10004 has the standard deviation of the sample means (which comes out to 10404.93362). After that, you have multiple options to figure out how many of the Z scores (the new column) are between -1 and 0 and are between 0 and 1. The quickest is to use the COUNTIF function. You obtain that there are 3,819 observations where the Z score is above -1 but below 0 using `=COUNTIF(AB2:AB10001,">-1")-COUNTIF(AB2:AB10001,">0")` where the Z scores are in Column AB. Similarly, you obtain that there are 3,157 observations where the Z score is above 0 but below 1 using `=COUNTIF(AB2:AB10001,">0")-COUNTIF(AB2:AB10001,">1")`. However,

you could do more crude methods like sorting by the Z scores and selecting the cells manually to see how many are within the asked ranges. You could also using the Filter tool and copy the filtered cells to a new worksheet to count.

- iii. $=\text{NORM.S.DIST}(2,\text{TRUE})-\text{NORM.S.DIST}(1,\text{TRUE}) = 0.135905122$ (Given the symmetry of the Normal distribution: $P(-2 < Z < -1) = P(1 < Z < 2)$.)
- iv. 0.1434 between -2 and -1 and 0.1157 between 1 and 2
- v. $=\text{NORM.S.DIST}(3,\text{TRUE})-\text{NORM.S.DIST}(2,\text{TRUE}) = 0.021400234$ (Given the symmetry of the Normal distribution: $P(-3 < Z < -2) = P(2 < Z < 3)$.)
- vi. 0.0082 between -3 and -2 and 0.0280 between 2 and 3

A3. Table C.1 is easily replicated with a PivotTable in Excel (three variables: ratio, max_size, and example_amt). It shows that 16,687 people are randomly assigned to the control group. 33,396 people are randomly assigned to a treatment group. 925 people are randomly assigned to the treatment group receiving a letter that: offers a 2 to 1 match ratio, states that the maximum size of the matching grant is \$50,000, and shows a low match example amount.

Count of ratio	Column Labels				
Row Labels	0	1	2	3	Grand Total
N/A	16687				16687
N/A	16687				16687
25000		2784	2785	2781	8350
low		928	927	927	2782
medium		929	929	928	2786
high		927	929	926	2782
50000		2782	2781	2782	8345
low		929	925	927	2781
medium		928	928	927	2783
high		925	928	928	2781
100000		2783	2783	2784	8350
low		929	927	929	2785
medium		926	928	927	2781
high		928	928	928	2784
Unstated		2784	2785	2782	8351
low		929	929	928	2786
medium		927	928	928	2783
high		928	928	926	2782
Grand Total	16687	11133	11134	11129	50083

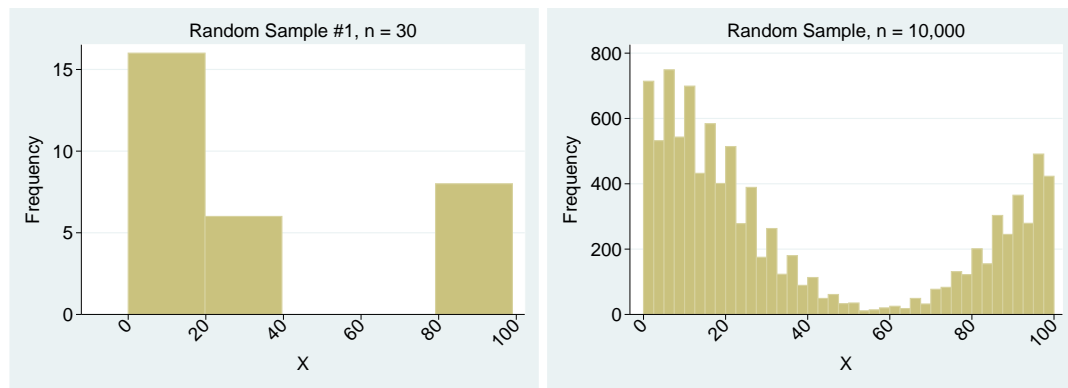
- A4.** (a) Check that your mean and standard deviation match those given in Table 1. Note that Table 1 clearly indicates right below the title that the reported values are means and standard deviations (not standard errors).
- (b) The first two numbers in Column (1a) would be 7.140232 and (8.360157) and the first two numbers in Column (1b) would be 13.13095 and (12.11711).
- (c) Check that your mean and standard deviation match those given in Table 1.
- (d) The values in Column (3) would be 0.6013672 and (0.4896316). (Note: You did not need to use Excel for these because the proportion in blue states is simply the compliment of the proportion in red states.)
- A5.** The point estimate of the difference is: 0.00198428. The standard error of the difference is: 0.00195483. The margin of error of the difference is: 0.00454761. The LCL is -0.00256334 and

the UCL is 0.00653189. We can fill in the blanks to offer an interpretation of the interval: We are 98% confident that the response rate among *all* previous donors to this charity would be between 0.3 percentage points lower to 0.7 percentage points higher if a 3 to 1 match were offered instead of a 1 to 1 match. (See question for further elaboration.)

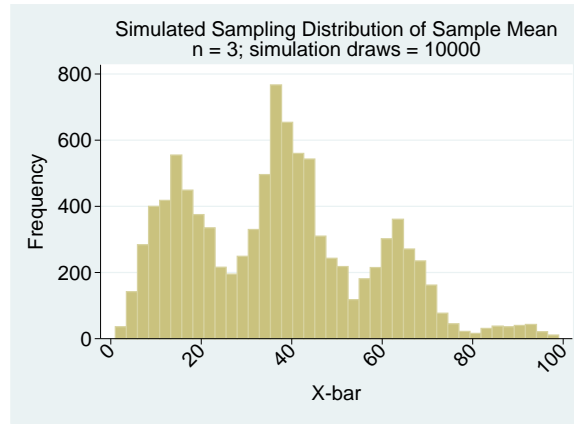
- A6.** (a) 0.02155689 and (0.00158934)
 (b) 0.02326987 and (0.00262081)
 (c) The point estimate of the difference is: 0.00153706. The standard error of the difference is: 0.00196461. The margin of error of the difference is: 0.00385064. The LCL is -0.00231349 and the UCL is 0.00538762.
 (d) Test $H_0 : (p_{high} - p_{low}) = 0$ versus $H_1 : (p_{high} - p_{low}) \neq 0$. Obtain a z test statistic of 0.782372982 and a P-value of 0.433995408.
- A7.** The point estimate of the difference in proportions, $(\hat{P}_{ST} - \hat{P}_{NT})$, where *ST* abbreviates superfluous taglines and *NT* abbreviates no taglines, is in the center of the confidence interval estimate of the difference. Hence, $(\hat{P}_{ST} - \hat{P}_{NT})$ is $-0.152 = \frac{(-0.201 + -0.103)}{2}$. We are given that $\hat{P}_{ST} = \frac{483}{1097}$. Hence, $(\hat{P}_{ST} - \hat{P}_{NT}) = -0.152 = (\frac{483}{1097} - \hat{P}_{NT})$. Thus, $\hat{P}_{NT} = (\frac{483}{1097} + 0.152)$. Finally, find X_{NT} in $\hat{P}_{NT} = \frac{X_{NT}}{n_{NT}}$ given that $n_{NT} = 1104$, which yields $X_{NT} = 654$. Hence, it must be that 654 people in the no taglines group chose the best card. (You can double-check your answer by plugging the numbers into the CI formula for the difference in proportions, using Excel to give you the exact value of $z_{\alpha/2}$ for $\alpha = 0.02$, to make sure you get the same LCL and UCL as in the question.)

Answers to the additional questions for extra practice.

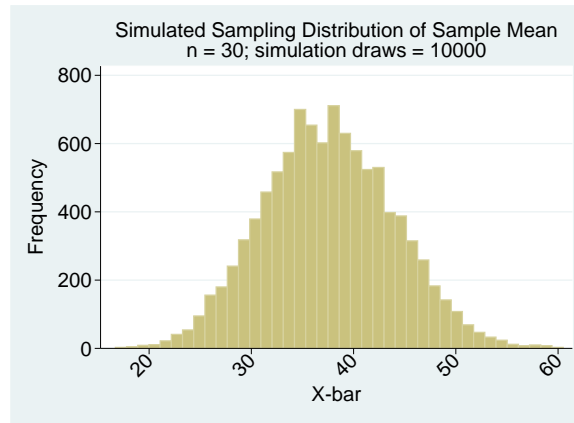
- A^e1.** (a) You could make a rough inference using Sample #1. However, in worksheet “Simulation 2” we have 300,000 ($=30 \times 10,000$) observations randomly drawn from Population 2 so it is easy to make an inference about that population using this very large sample. In fact, we do not need to use all 300,000 observations. For ease, we can use the first 10,000 draws from the population (i.e. x_1). We can confidently infer with this large sample and the clear pattern in the histogram that the population is bimodal.



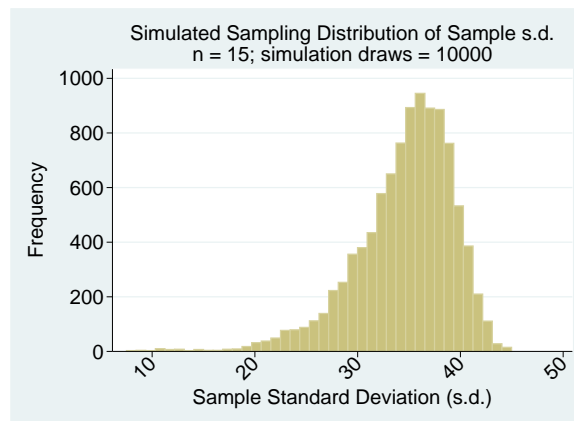
- (b) The mean of the 10,000 sample means is 37.22233 and the s.d. of the 10,000 sample means is 19.99509. The shape is unusual: four modes!



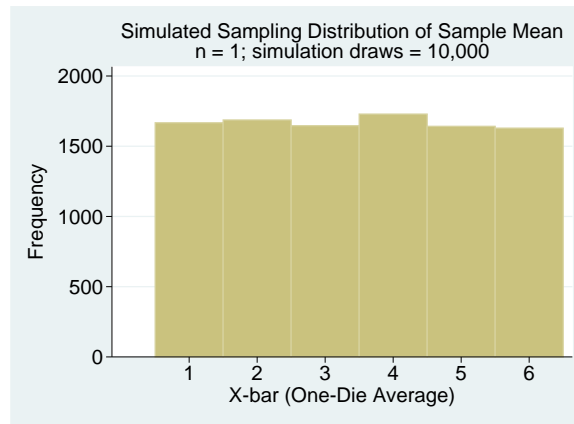
- (c) The mean of the 10,000 sample means is 37.56397 and the s.d. of the 10,000 sample means is 6.394322. The shape looks almost Normal (see histogram below): there are still some hints of modality but we've definitely gotten a lot closer to Normal with this larger sample size.



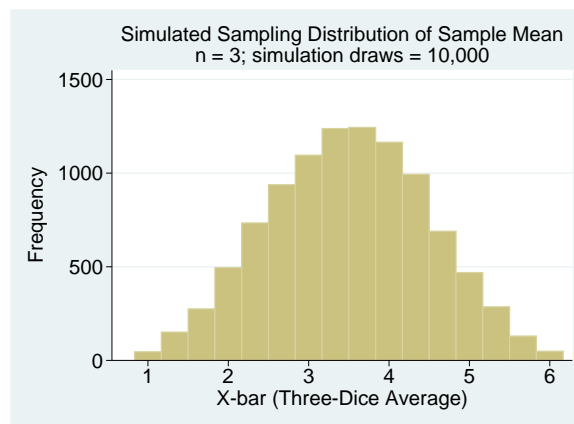
- (d) The mean of the 10,000 sample standard deviations is 34.58903 and the s.d. of the 10,000 sample standard deviations is 4.770962. The shape is negatively skewed (see histogram below).



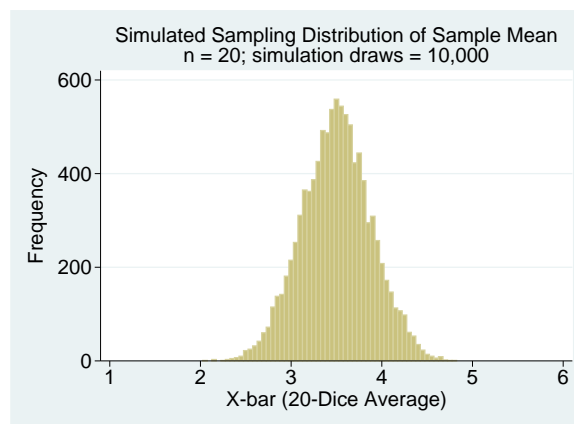
- A^e2.** (a) To generate tosses of a fair coin: `=ROUND(RAND(),0)`. To generate tosses of an unfair coin with a 40% chance of heads: `=ROUND(RAND()-0.1,0)`.
- (b) Six unique values of \bar{X} occur in the simulation. A value between 3 and 4 (including those endpoints) occurs 3,375 times.



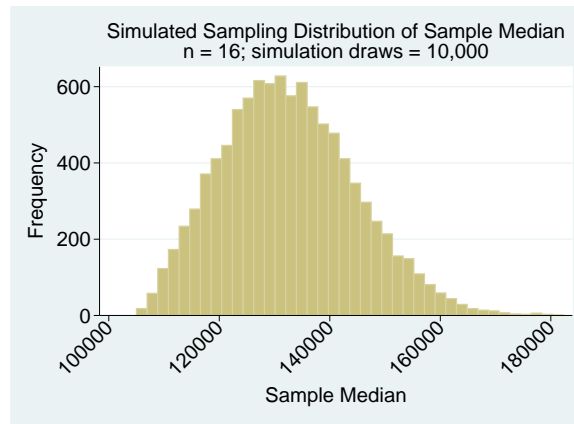
- (c) Sixteen unique values of \bar{X} occur in the simulation. A value between 3 and 4 (including those endpoints) occurs 4,743 times.



- (d) Fifty-four unique values of \bar{X} occur in the simulation. A value between 3 and 4 (including those endpoints) occurs 8,289 times.



- A^e3.** (a) The mean of the 10,000 sample medians is 132538.9 and the s.d. of the 10,000 sample medians is 12068.4. The shape is somewhat positively skewed.



- (b) The sample mean has a smaller s.d. than the sample median: by this metric, \bar{X} is less affected by sampling error. However, \bar{X} has a bigger range: by this metric, the sample median is less affected by sampling error. (The extreme right skew causes the large range for the sampling distribution of \bar{X} : including an outlier affects \bar{X} , but not the median.)

A^e4. Everyone will get the same results for the first row: for Draw #1, \$141,233.3. (Other rows depend on the draws from the random number generator.)

D Module D: Inference about μ & $(\mu_2 - \mu_1)$ & Using Dummies

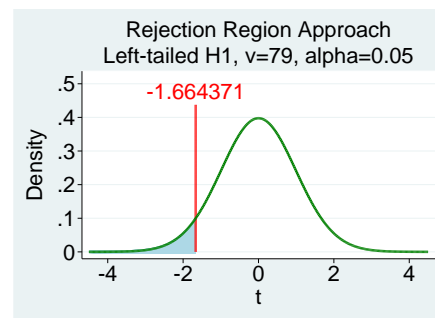
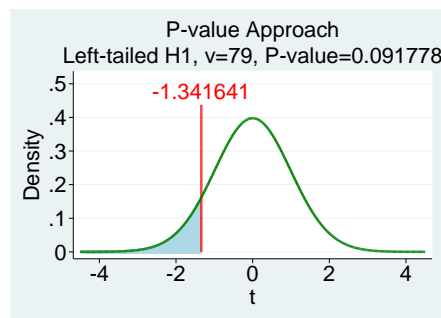
D.1 Module D.1: Inference about a Mean & Regression Refresher

Main concepts: Hypothesis testing for inference about a single mean. Refresher on simple regression analysis (in preparation for multiple regression).

Case studies: We work with “Sparton Resources of Toronto” (p. 430 of textbook) and revisit Currie and Schwandt (2016), “Mortality Inequality: The Good News from a County-Level Approach.”

Required readings: Sections 5.8 - 5.9 and Chapter 13 (including p. 430). The readings and module analysis of Currie and Schwandt (2016) in Module B.1. The refreshers and background below.

- To refresh Chapter 13 concepts, recall Karlan and List (2007) from Module C.3. Review again the second and third rows of results in Panel A, Columns (1) and (2) in Table 2A on page 88: “Dollars given, unconditional” and “Dollars given, conditional on giving.” Recall that unconditional means using all observations, including those who gave zero, and conditional means ignoring the zeros. Is the mean dollars given, conditional on giving, for the treatment group statistically significantly lower than \$50? If so, at which conventional significance levels?
 - Translating to formal hypotheses and standard notation yields $H_0 : \mu_{T|G} = 50$ versus $H_1 : \mu_{T|G} < 50$, with $T|G$ for those in the treatment group who gave something.
 - For inferences about *means*, use t test statistics and the Student t distribution. From our aid sheets: $t = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$ and $\nu = n - 1$, where s is the sample standard deviation and ν is the degrees of freedom.
 - Recall the two methods of hypothesis testing: the P-value approach and the rejection region approach.
 - For example, suppose of 1,000 people in the treatment group, 80 donate and for those 80, the average donation is \$48.50 with a s.d. of \$10.00. $t = \frac{48.50 - 50}{10/\sqrt{80}} = -1.341641$ and $\nu = 79$.

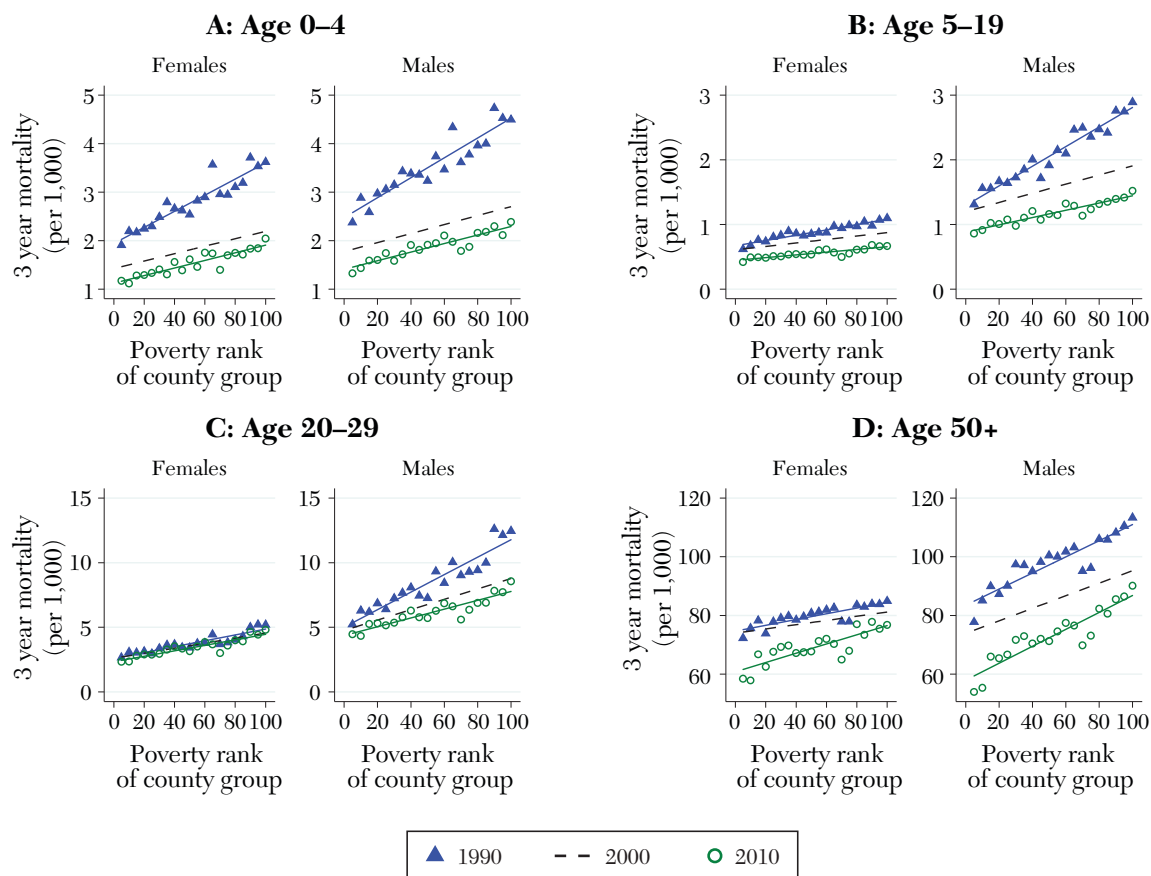


We obtain a P-value (left graph) of 0.091778, which is somewhat small and provides some support for the research hypothesis that the mean donation among those giving in the treatment group is below \$50. This P-value is less than 0.10, but bigger than 0.05, which means that the best conventional significance level we meet is $\alpha = 0.10$. For the rejection region approach, if we choose $\alpha = 0.05$, we obtain a critical value (right graph) of -1.664371. The t test statistic of -1.341641 is not in the rejection region: we have insufficient proof of H_1 at a 5% significance level.

- We revisit Currie and Schwandt (2016) to brush up on simple regression. Carefully review Figure 3 (reproduced below) noting the OLS regression lines. Our course includes extensive coverage of regression analysis, a workhorse in empirical research. Early on we use simple regression to describe the relationship between two interval variables (e.g. Modules B.1, B.2, B.3). After studying statistical inference, we return to simple regression (i.e. hypothesis testing and confidence interval estimation in a regression context). Finally we get to multiple regression: a new and powerful tool for both descriptive and inferential statistics. Multiple regression is the final frontier of ECO220Y and we spend weeks with this extremely important topic. In this module we review simple regression as a descriptive tool to get ready for our return to regression analysis.

Figure 3

Three-Year Mortality Rates across Groups of Counties Ranked by their Poverty Rate



Source: Authors using data from the Vital Statistics, the US Census, and the American Community Survey.
Note: Three-year mortality rates for four different age groups are plotted across county groups ranked by their poverty rate. Mortality rates in 2000 and 2010 are age-adjusted using the 1990 population, that is, they account for changes in the age structure within age, gender, and county groups since 1990. Table A3 provides magnitudes for individual mortality estimates and for the slopes of the fitted lines.

Figure 3: Currie and Schwandt (2016), p. 41. Panel C should say “Age 20-49,” not “Age 20-29.”

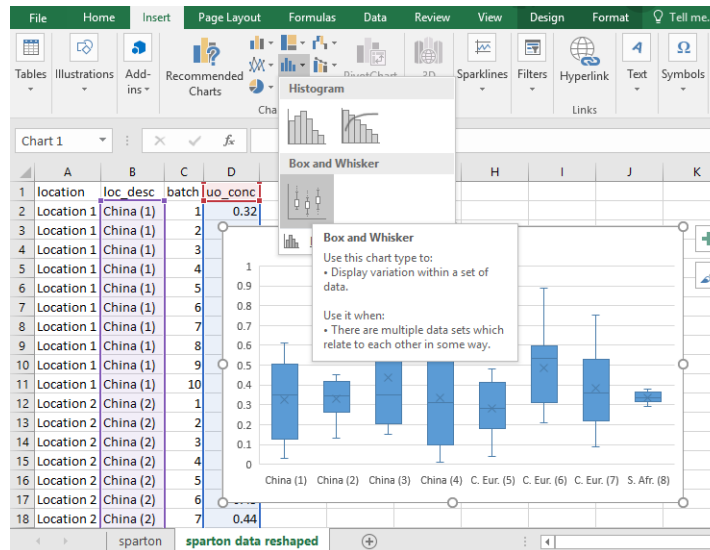
Datasets: For Currie and Schwandt (2016): [mort_in_figure_3_table_a3.xlsx](#), where the suffix says these data replicate Figure 3 and Table A3 in that paper. For Sparton Resources: [sparton.xlsx](#).

Interactive module materials for Module D.1:

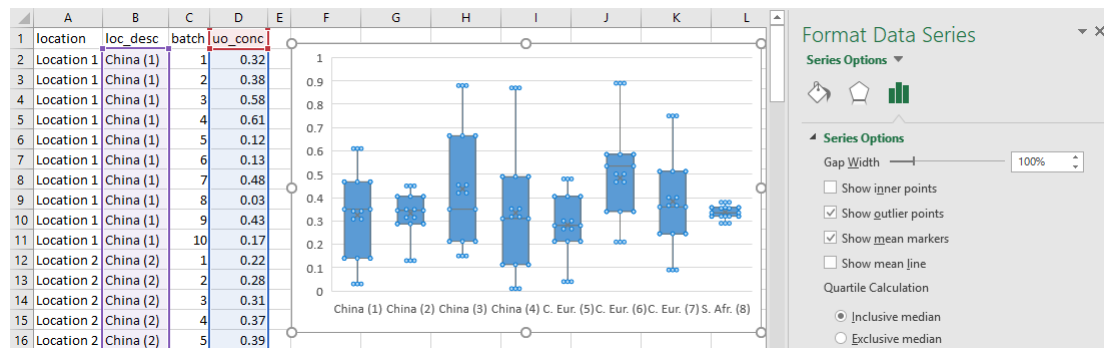
1. **Recall** the readings for Sparton Resources: a source of coal ash is economically profitable if the concentration of uranium oxide (UO) is **at least 0.32 pounds per tonne**. Use [sparton.xlsx](#).

- (a) **Create** one box plot that summarizes UO concentrations for each of the eight locations.

EXCEL TIPS: In the worksheet “sparton data reshaped,” select the variables loc_desc and uo_conc. Insert a Box and Whisker chart (see below). (Note: If you use the data as organized in the “sparton” worksheet, the horizontal axis will not be labeled properly.)



Excel offers two ways to compute percentiles that it calls “Inclusive median” and “Exclusive median.” You can choose. (Both return the same median.) Right click anywhere on the blue box plots. (With a mac, hold the control key and click.) Select Format Data Series... from the drop down menu. Under Quartile Calculation, select Inclusive median.



Note: Recall from Section 5.9 (textbook) that there are multiple ways to interpolate when finding percentiles. With many observations, it usually makes little difference. With a small sample (like $n = 10$), you can see the difference. Some worry about the technical reasons for the (typically tiny) difference in special cases: it’s not worth your attention.

Interpretation tips: Which location looks most promising? Of the eight, Location 6 in Central Europe has both the highest median and mean concentration of UO, both roughly 0.5 lbs/tonne. While two of the 10 batches fall short of the 0.32 lbs/tonne profitability threshold, some very far exceed it. However, 10 batches is a small sample: these promising

results may not hold up with further sampling. (Similarly, some other locations may appear worse than they really are because of sampling error.) Formal inference can help us figure out how much sampling error can affect our conclusions about these locations.

(b) To start, **focus on Location 1**.

- i. Describe the uranium oxide concentrations in the randomly selected batches of coal ash by **computing** standard summary statistics. Include the 25th and 75th percentiles.

Verify that you get $n_1 = 10$, $\bar{X}_1 = 0.325$ (sample mean), $s_1 = 0.204246583$ (sample standard deviation), 25th percentile ≈ 0.13 , and 75th percentile ≈ 0.48 .

Note: It says \approx because there is no exact correct answer. The values of 0.13 and 0.48 are what Stata returns. Excel offers two choices (and neither matches Stata).

EXCEL TIPS: Copy the data for Location 1 to a new worksheet named “Summary Statistics for Loc 1.” Use Descriptive Statistics under Data Analysis. (See page 19 if you forgot how to do this.) To add the requested quartiles, you can use the Excel functions: PERCENTILE.INC, QUARTILE.INC, PERCENTILE.EXC, or QUARTILE.EXC. QUARTILE.* is a special case of PERCENTILE.*: just enter 0.25 and 0.75 into the percentile function. The functions COUNT, AVERAGE, STDEV.S, MIN, and MAX can return specific values shown in Descriptive Statistics.

	A	B	C	D	E	F	G	H	I	J
1	location	loc_desc	batch	uo_conc					uo_conc	
2	Location 1	China (1)	1	0.32		0.14	=PERCENTILE.INC(\$D\$2:\$D\$11,0.25)			
3	Location 1	China (1)	2	0.38		0.1275	=PERCENTILE.EXC(\$D\$2:\$D\$11,0.25)		Mean	0.325
4	Location 1	China (1)	3	0.58		0.14	=QUARTILE.INC(\$D\$2:\$D\$11,1)		Standard Error	0.0645884
5	Location 1	China (1)	4	0.61		0.1275	=QUARTILE.EXC(\$D\$2:\$D\$11,1)		Median	0.35
6	Location 1	China (1)	5	0.12					Mode	#N/A
7	Location 1	China (1)	6	0.13					Standard Deviation	0.2042466
8	Location 1	China (1)	7	0.48		0.4675	=PERCENTILE.INC(\$D\$2:\$D\$11,0.75)		Sample Variance	0.0417167
9	Location 1	China (1)	8	0.03		0.505	=PERCENTILE.EXC(\$D\$2:\$D\$11,0.75)		Kurtosis	-1.467801
10	Location 1	China (1)	9	0.43		0.4675	=QUARTILE.INC(\$D\$2:\$D\$11,3)		Skewness	-0.010563
11	Location 1	China (1)	10	0.17		0.505	=QUARTILE.EXC(\$D\$2:\$D\$11,3)		Range	0.58
12									Minimum	0.03
13						10	=COUNT(D2:D11)		Maximum	0.61
14						0.2042	=STDEV.S(D2:D11)		Sum	3.25
15						0.325	=AVERAGE(D2:D11)		Count	10
16						0.03	=MIN(D2:D11)		Largest(1)	0.61
17						0.61	=MAX(D2:D11)		Smallest(1)	0.03

- ii. **Compute** the standard error of \bar{X}_1 . **Verify** that you get 0.064588441.

EXCEL TIPS: Go to the worksheet “sparton,” which we use for the rest of this module. Referencing the last Excel tip, add the functions for n, the mean, and the s.d. Program in $\frac{s}{\sqrt{n}}$ from our aid sheets referencing your cells with s and n .

- iii. **Enter** the value from the null hypothesis ($H_0 : \mu_1 = 0.32$). **Compute** the t test statistic using $t = \frac{\bar{X} - \mu_0}{\frac{s}{\sqrt{n}}}$ from our aid sheets. **Verify** that you get 0.077413233.

- iv. **Consider** the hypothesis test: $H_0 : \mu_1 = 0.32$ versus $H_1 : \mu_1 > 0.32$, where μ_1 is the mean uranium oxide concentration for *all* coal ash from Location 1. If Sparton Resources wants to source exclusively from places *proven profitable*, then we need a right-tailed test. **Compute** the P-value. **Verify** that you get 0.46999426.

EXCEL TIPS: Use the Excel function T.DIST. Do *not* use T.DIST.RT, which is simply (1 - T.DIST) and hence redundant. Do *not* use T.DIST.2T, which gives an error message if your test statistic is negative.

- v. **Consider:** $H_0 : \mu_1 = 0.32$ versus $H_1 : \mu_1 < 0.32$. If Sparton Resources wants to source from all locations except those *proven unprofitable*, then we need a left-tailed test. **Compute** the P-value. **Verify** that you get 0.53000574.
- vi. **Consider:** $H_0 : \mu_1 = 0.32$ versus $H_1 : \mu_1 \neq 0.32$. While a two-tailed tests does

not make business sense (we care about profitable or unprofitable, not *different* from break-even), many software packages automatically report two-tailed tests even when they do not make economic/business sense. You need to understand two-tailed tests well. **Compute** the P-value. **Verify** that you get 0.93998852.

EXCEL TIPS: Use the Excel functions T.DIST and ABS.

- (c) **Assess** the other seven locations. Verify that you can prove at a 5% significance level that Locations 6 and 8 are profitable, that you are unable to prove at a 10% significance level that any of the eight locations are unprofitable, and that you can prove that Location 6 is different from break-even at a 5% significance level and that Location 8 is different from break-even at a 10% significance level. Make sure you obtain a P-value of 0.44759458 for Location 5 for the two-tailed test (different from break-even).

Test/exam examples: The Sparton Resources case study has appeared on a test.

- Question (5), [February 2016 Test #3](#) (with [solutions](#))

2. Recall Currie and Schwandt (2016) and use [mort_in_figure_3_table_a3.xlsx](#):

- (a) **Run a regression** for females aged 0-4 in 1990 to replicate the blue line in Panel A of Figure 3 on page 108. **Verify** that you get $\hat{y} = 1.9404 + 0.0166x$, $n = 20$. (Hints: To identify the x-variable, see the variable descriptions worksheet “readme.” Also, make sure to run the regression on the correct subset of the data: females aged 0-4 in 1990.)

EXCEL TIPS: Add a variable for adjusted mortality to the original data. Use the Filter tool to select the subset. (For refreshers, see the Excel tips for part 2a in Module B.1 on page 43). After filtering the original data, copy the subset to a new worksheet. Put the regression output in that new worksheet. (Do not run regressions directly on the original data with the filter.) A new worksheet helps document which subset the regression results are for. Use the labels option (to include the variable names in the regression output).

- (b) Look at Figure 3 on page 108 for males in 1990 (the blue triangles). Compare the blue line for males aged 20-49 (Panel C) versus the blue line for males aged 50+ (Panel D).
- State** which age group of males in 1990 has a steeper OLS line and which has a larger value of s_e : the group aged 20-49 or the group aged 50+?
 - Check** your answers by running the two simple regressions that correspond to each. **Verify** that you obtain $\hat{y} = 5.018749 + 0.0675931x$, $n = 20$, $s_e = 0.73538$ for males in 1990 aged 20-49. **Verify** that you obtain $\hat{y} = 83.54337 + 0.2735487x$, $n = 20$, $s_e = 3.7934$ for males in 1990 aged 50+.

Interpretation tips: If the s_e measures scatter around the OLS line, why does Figure 3 show similar scatter (blue triangles) for males aged 20-49 and 50+ when the s_e is more than *five* times bigger for 50+? Notice the y-axis scale. Older males have much higher mortality rates. Figures zoom in to the range of the data. The 50+ graph starts at a mortality rate of 60 deaths per 1,000 (versus 0 for males aged 20-49) and tics off each extra 20 deaths per 1,000 (versus 5 for males aged 20-49). Hence, the s_e values of 0.7 and 3.8 deaths per 1,000 correctly reflect the scatter for males aged 20-49 and 50+, respectively.

Test/exam examples: For examples of Currie and Schwandt (2016), see page 45.

D.2 Module D.2: Inference about a Difference in Means

Concepts: Inference about how two means differ. Distinguishing independent samples from paired data. Review the foundation of statistical inference: using a random sample and its statistics to make an inference about a population and its parameters.

Case studies: We work with the *population* of *all* Ontario public sector employees making \$100K+ using the public disclosures of 2016 and 2015 salaries.

Required readings: Chapter 14. Also, review the background for the annual Ontario public sector salary disclosures in Module C.1 and the further background here:

- Browse Table D.1. The mean salary is (virtually) unchanged from 2015 to 2016: there is a tiny \$180 mean annual increase, just a 0.1% rise. Are the salaries of ON public sector employees not even keeping pace with inflation? The second row of results shows that among employees that had their 2015 salary disclosed, salaries increased by about 2.1% on average. The first row seems to suggest virtually no increase because each year there are a bunch of employees who, for the first time, cross the \$100,000 threshold and now have their salary disclosed. This influx of “low” salary employees holds the mean back.

Table D.1: Exploring Why the Mean Salary (CAN \$1,000's) is Virtually Unchanged

	2015 Salaries	2016 Salaries	Change
Unconditional	127.071 (37.445) [115,734]	127.250 (36.829) [124,267]	0.180
Conditional on employee having both her/his 2015 salary and 2016 salary disclosed: same-employee comparison	129.078 (38.311) [97,600]	131.831 (38.989) [97,602]	2.754

Note: Each cell gives the mean, (standard deviation), and [number of observations].

Note: For why $97,600 \neq 97,602$, see part 5 on page 118.

- This is why retail firms report revenue growth in **same-store sales** in annual reports to shareholders. Comparing total revenues would be misleading if the retailer were opening and/or closing retail outlets (which is commonplace). For example, all older stores could be doing worse, but if the retailer opens more outlets, the total sales may even go up compared to last year. Investors demand better information (same-store sales) to assess the firm's performance. For ON public sector salaries, the additional employees crossing the reporting threshold each year mask the rising salaries. Reporting **same-employee salaries** or same-store sales addresses composition effects (first explored in Module B.1).
- Are the numbers in Table D.1 statistics or parameters? Recall that statistics describe random samples whereas parameters describe populations. By law, *all* ON public sector employees, making \$100K or more, must have their salaries publicly disclosed. These data are certainly not a random sample. In fact, this is a rare instance where we have access to a population. Hence, we can hone understanding of statistical inference – making an inference about a population and its parameters using a random sample and its statistics – by exploiting this rare opportunity to check our inferences against the facts (the known population parameters).
- To enable random sampling where we all get the exact same results, each of the ON salary

datasets has four pre-generated variables: **random1**, **random2**, **random3**, and **random4**. While the names are the same across datasets, they are **NOT** the same variables. Each uses independent draws from a random number generator.⁹ Sorting by these puts the data in random order. For example, if we sort by random1 and take the first n observations, we all get the same random sample. (Alternatively, we could all use the last n or agree on something else.)

- To assess how salaries changed from 2015 to 2016 consider an independent samples versus a paired data approach. We can compare a random sample of employees in 2016 with an independent random sample of employees in 2015. With paired data, compare the *same employees* both years: compare each person with themselves. When it is possible, a paired data approach can really cut-down sampling error (lower standard errors). The paired data approach *holds constant* differences *across* people by comparing each person with themselves. While salaries vary wildly across people, they are quite predictable for a specific person given their current salary. In contrast, the independent samples approach must deal with all the noise caused by differences across people: you can get two very different samples of people when you sample independently. This module helps you appreciate the difference between these two approaches.

Datasets: For Ontario (2015) and (2016): [on_sal_2015.xlsx](#) and [on_sal_2016.xlsx](#). Also, [on_sal_16_15.xlsx](#) contains the merged data.

Interactive module materials for Module D.2:

1. Using [on_sal_2016.xlsx](#), **replicate** the results for the 2016 salaries of the same-employee subset in Table D.1. The table is reproduced below, with the numbers to replicate in boldface.

ON Public Sector Salaries (CAN \$1,000's)			
	2015 Salaries	2016 Salaries	Change
Unconditional	127.071 (37.445) [115,734]	127.250 (36.829) [124,267]	0.180
Conditional on employee having both her/his 2015 salary and 2016 salary disclosed: same-employee comparison	129.078 (38.311) [97,600]	131.831 (38.989) [97,602]	2.754

Note: Each cell gives the mean, (standard deviation), and [number of observations].

EXCEL TIPS: To separately describe subsets of data, use a PivotTable. Drag the variable disc2015 to ROWS and drag three copies of the variable salary to Σ VALUES (for the average, s.d., and count). One wrinkle: select StdDevp to compute the population s.d. σ (although with rounding, you cannot tell if the degrees of freedom correction has been done). Drag the field Σ Values (just Values in a mac) to ROWS to make the PivotTable look more like Table D.1.

Interpretation tips: What do these three numbers mean? For the 97,602 ON public sector employees who made at least \$100,000 CAN in *both* 2015 and 2016 (i.e. employees appearing in both annual disclosures), the mean of the 2016 salaries is just under \$132,000. The standard

⁹We use Stata to create the random1 to random4 variables separately (i.e. independently) for each dataset. For example, to create random1, we start with fresh random draws from the Uniform distribution (U[0,1]) for each observation in a dataset and then sort. Random1 records which observation is 1st, 2nd, ..., n^{th} . Converting from draws from U[0,1] to a rank integer keeps file sizes smaller (integers use less space than floats) and avoids machine precision issues.

deviation of the 2016 salaries is just under \$39,000. Given that everyone in the disclosure makes *at least* \$100,000, the huge standard deviation and modest mean imply that the distribution of these employee's salaries must be strongly positively skewed. This is not surprising as salary distributions are almost always positively skewed (a few top earners and a vast majority of more modest earners). These 97,602 employees represent about 79 percent of all ON public sector employees appearing in the disclosure of 2016 salaries.

2. **Browse** Table D.2, *including* the note. (Remember the required reading about **random3**).

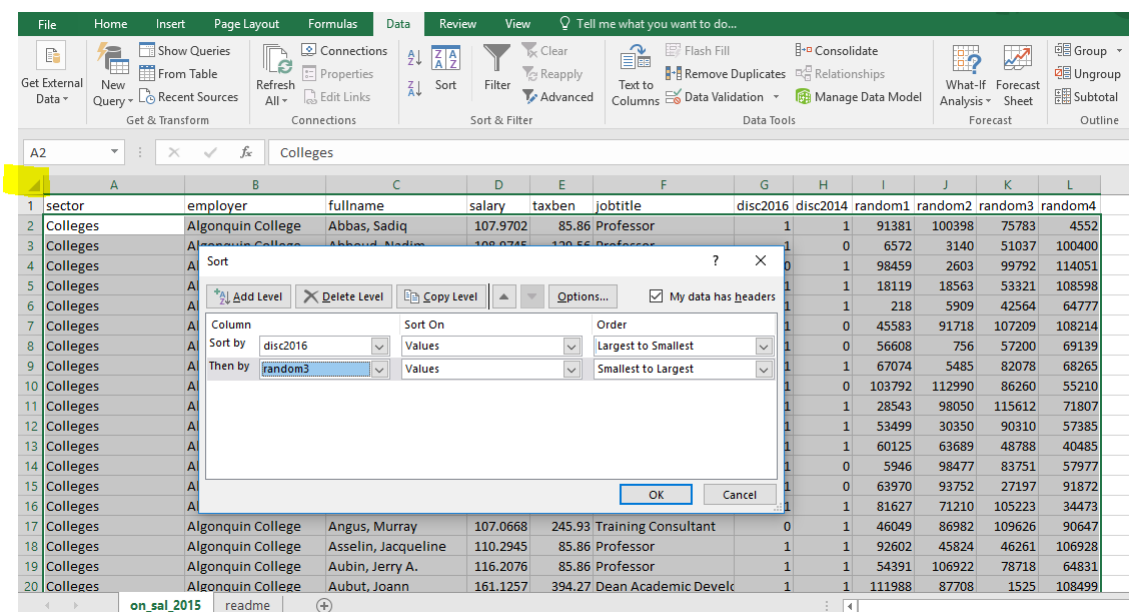
Table D.2: Independent Random Samples from the *Same-Employee Subset* of the Population

	2015 Salaries			2016 Salaries			Difference	
	mean	s.d.	s.e.	mean	s.d.	s.e.	mean	s.e.
$n_{15} = n_{16} = 100$	132.841	56.364	5.636	134.752	35.812	3.581	1.911	6.678
$n_{15} = n_{16} = 4,000$	129.377	37.893	0.599	132.195	39.156	0.619	2.818	0.862

Note: To replicate, sample from the subset with *both* the 2015 and 2016 salary disclosed. Sort by random3 in [on_sal_2016.xlsx](#) and [on_sal_2015.xlsx](#), using the first n observations.

- (a) Using [on_sal_2015.xlsx](#), **replicate** the first three numbers in boldface in Table D.2. Use the variable disc2016 to restrict to the same-employee subset.

EXCEL TIPS: Select the *entire* worksheet with the data: click the area right above row 1 and to the left of column A (highlighted yellow below). Sort by disc2016 in *descending* order and random3 in *ascending* order. Copy the first 100 observations (101 rows, with variable labels) to a new worksheet named “Random Sample of 100 for 2015.” Compute summary statistics using Descriptive Statistics under Data Analysis.



- (b) Using [on_sal_2016.xlsx](#), **replicate** the second three numbers in boldface. Use the variable disc2015 to restrict to the same-employee subset.
- (c) **Toggle** between your random sample of 100 employees from 2015 and from 2016. **Notice that** it is *not the same* 100 employees: these are two entirely independent random samples.
- (d) **Create** a dummy variable for 2016, named dum_2016, in *both* your 2015 and 2016 random samples of 100 employees. It will simply be a column of zeros for 2015 and a column of

ones for 2016. **Delete** the variables disc2014, disc2015, and/or disc2016 and the random1 to random4 variables. In other words, keep only the first six variables (sector, employer, fullname, salary, taxben, and jobtitle) and the newly created variable dum_2016.

EXCEL TIPS: The screenshot below shows the work for the 2016 sample.

- (e) **Combine** the 2015 and 2016 samples by stacking them (for a total of 200 rows of data). Do not repeat the row with the variable labels: your stacked data should have 201 rows. **Run** a regression where the y variable is salary and the x variable is dum_2016. **Verify** that the intercept exactly equals the sample mean 2015 salary and that the coefficient on the dummy variable exactly equals the difference between the sample means of the 2015 and 2016 salaries (last column of Table D.2).

EXCEL TIPS: Put the stacked data into a new worksheet (in either your 2015 or 2016 workbook) titled “Stacked 2015 and 2016 samples.” Use Regression under Data Analysis.

	A	B	C	D	E	F	G	H	I	J	K	L
1	sector	employer	fullname	salary	taxben	jobtitle	dum_2016					
2	Municipal	Sarnia	Lewis, Joh	112.5448	816.72	First Class	0		SUMMARY OUTPUT			
3	Ministries	Environm	Espie, Jon	118.5028	165.48	Chief Of S	0					
4	Ministries	Treasury	E Hammad,	106.7136	188.48	Senior Inf	0		Regression Statistics			
5	Municipal	Regional	I Culham, S	109.8464	537.69	Superviso	0		Multiple R	0.020332767		
6	Ministries	Education	Brisard, Br	120.3342	246.2	Education	0		R Square	0.000413421		
7	Municipal	Toronto	Ti Zuccarini,	112.2963	253.5	Senior Bu	0		Adjusted R Square	-0.004634996		
8	Ministries	Environm	Jun, Henn	100.5178	173.81	Senior Pol	0		Standard Error	47.2198437		
9	Universiti	University	Telmissan	122.0783	46.8	Professeu	0		Observations	200		
10	Universiti	University	O'Connor,	175.625	565.98	Professor	0					
11	Municipal	Toronto	Ti Biase, Car	101.7225	380.65	Engineeri	0		ANOVA			
12	Ontario	Pt Ontario	Pt O'Connor,	107.3857	1458.4	Mechanica	0			df	SS	
13	Hospitals	Childrens	Wolf, Bern	119.1948	601.5	Director F	0		Regression	1	182.594139	18
14	Other	Pvt Circle	Of F Scheinert,	400.0001	0	President	0		Residual	198	441483.3006	222
15	Municipal	Regional	I Oka, Kiyos	182.6927	1045.71	Director V	0		Total	199	441665.8947	
16	Universiti	Carleton	L Westerlur	129.2056	0	Faculty M	0					
17	Hospitals	Holland B	Hoffman,	192.7241	669.24	Physician	0			Coefficients	Standard Error	t
18	Ministries	Transport	Yeo, Debc	136.2565	221.72	Manager E	0		Intercept	132.840919	4.72198437	28.:
19	Municipal	Town Of	C Sweet, To	128.7971	559.16	Manager C	0		dum_2016	1.91099	6.677894338	0.2
20	Municipal	Toronto	P Taylor, Ke	115.7949	818.86	Police Cor	0					

Note: If you are surprised by the connection between differences in means (Chapter 14) and a simple regression with a dummy variable, review part 5 on page 49 on Module B.2.

- (f) **Replicate** the last two columns of results for random samples of size 100 in Table D.2. For the point estimate of the difference, it is simply $(\bar{X}_{2016} - \bar{X}_{2015})$. For the standard error of the difference in means, use $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ from your aid sheets.¹⁰

¹⁰While the regression in part 2e also replicates the standard error of the difference in this case (see the s.e. for the variable dum_2016), this is *not* generally true. When both independent samples have the same exact sample size (in this case, 100) the standard error of the difference in means is the same regardless of whether you pool or not. In other

Interpretation tips: What do the numbers 1.911 and 6.678 in the last column of Table D.2 mean? Focusing on the 97,602 ON public sector employees with both their 2015 and 2016 salaries disclosed, the 2016 salaries for a random sample of 100 employees is on average \$1,911 higher than the 2015 salaries for an independent random sample of 100 employees. This is a 1.4 percent annual increase, which is not surprising. The standard error of \$6,678 measures how much sampling error affects the point estimate of the mean salary change of \$1,911. The standard error is *huge*, which is not surprising because we have quite small sample sizes of 100 employees and high salary variability, which both increase the standard error. In other words, our estimate that salaries on average increased by \$1,911 is not precisely measured: it may be thousands of dollars too high or too low because of sampling error (which employees happened to end up in our samples). Formal inference via hypothesis testing or confidence interval estimation would require computing the degrees of freedom.¹¹ However, the point estimate and standard error make it obvious that we cannot even reject the null hypothesis that salaries are unchanged (or even that they went down by \$5,000!) and that our CI estimates will be very wide. This is why it is common for researchers to report point estimates and standard errors, the key building blocks for inference. In this rare instance, we have the entire population handy and we know $(\mu_{2016} - \mu_{2015})$ equals \$2,754 (see Table D.1 on page 112). Hence, we got pretty lucky with our point estimate of \$1,911: it is not that far off.

(g) **Replicate** the last row of Table D.2 for random samples of size 4,000.

3. Table D.2 shows a highly inefficient way to make an inference about how salaries have changed. We should use a paired data approach. **Browse** Table D.3, *including* the note below it.

Table D.3: Paired Data Random Sample from the Same-Employee Subset of Population

	2015 Salaries			2016 Salaries			Difference		
	mean	s.d.	s.e.	mean	s.d.	s.e.	mean	s.d.	s.e.
$n = 100$	127.695	35.273	3.527	130.623	35.336	3.534	2.929	9.295	0.930
$n = 4,000$	129.778	39.016	0.617	132.486	39.632	0.627	2.708	13.489	0.213

Note: To replicate, sort by random3 in [on_sal_16_15.xlsx](#), using the first n observations.

- (a) Using [on_sal_16_15.xlsx](#), **replicate** the last three columns for $n = 100$ in Table D.3, which are in boldface.

EXCEL TIPS: First, create a new variable named difference that is salary16 minus salary15. Next, recall the Excel tips given with part 2a on page 114.

Note: Recall that for paired data you make a new variable d , which is the difference, and find \bar{d} and s_d . For the standard error, use $\frac{s_d}{\sqrt{n}}$ from your aid sheets, where the numerator is just the regular s.d. of the difference variable. Descriptive Statistics in Data Analysis includes the standard error.

- (b) **Find** the correlation between salary15 and salary16 for $n = 100$. **Verify** that you get 0.965341.

words, the formulas for the s.e. in Sections 14.4 and 14.5 return the same value in this special case. Remember the homoscedasticity (aka equal variance or constant spread) assumption of regression analysis.

¹¹Recall $\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$.

EXCEL TIPS: In Data Analysis use Correlation or use the function CORREL.

Interpretation tips: What does the number 0.965341 mean? For our random sample of 100 ON public sector employees with both their 2016 and 2015 salaries disclosed, there is an extremely strong positive correlation between these two salaries. This is not surprising as we would generally expect people with a high salary in 2015 to have a high salary in 2016. A person's salary does not typically experience a dramatic change year-over-year.

- i. **Verify** the standard deviation of the 2015 to 2016 salary difference, which Table D.3 reports as 9.295, equals $\sqrt{sd_{2016}^2 + sd_{2015}^2 - 2 * r * sd_{2016} * sd_{2015}} = \sqrt{1248.640 + 1244.171 - 2 * 0.965341 * 35.3361 * 35.2728} = 9.295$.

- (c) For just this part, go back to your data for the independent samples analysis: [on_sal_2015.xlsx](#) and [on_sal_2016.xlsx](#). **Make a prediction** about the correlation between salary15 and salary16 for $n_{15} = n_{16} = 100$. **Compute** it to check. **Verify** that you get -0.033144.

EXCEL TIPS: It is a bit silly, but copy the 100 salaries from 2015 and paste them into a new worksheet. Next to them, paste the 100 salaries from 2016. (It is silly because these are not the same people so having them side-by-side has no practical meaning.)

	A	B	C	D	E	F	G	H	I	J
1	salary	salary								
2	112.5448	131.0002		=CORREL						
3	118.5028	117.3011								
4	106.7136	107.6391								

Interpretation tips: What does the number -0.033144 mean? As expected, there is virtually zero correlation between the 2015 and 2016 salaries in two *independent samples* each with 100 ON public sector employees (one sample for each year). It is not exactly zero, just near zero, because of sampling error: by chance, there will be a slight positive or negative correlation. In practice, we would refer to this as no correlation. This stands in stark contrast to the very strong positive correlation of 0.965341 between the 2015 and 2016 salaries in the *paired data* with a random sample of 100 ON public sector employees.

- i. **Create** a difference variable and **compute** its standard deviation. **Verify** that you get 67.773364.
- ii. **Verify** $67.773364 = \sqrt{sd_{2016}^2 + sd_{2015}^2 - 2 * r * sd_{2016} * sd_{2015}} = \sqrt{56.36434924^2 + 35.81183342^2 - 2 * (-0.033143657) * 56.36434924 * 35.81183342}$. Notice that $67.773364 \approx \sqrt{sd_{2016}^2 + sd_{2015}^2} = \sqrt{56.36434924^2 + 35.81183342^2} = 66.8$. Because the correlation in salaries across two independent samples is virtually zero, it makes little difference if we ignore it.

Interpretation tips: Why is the value 0.930 in the last column of Table D.3 (page 116) much smaller than the value 6.678 in the last column of Table D.2 (page 114)? Even though the sample sizes are the same in both tables ($n = 100$), the standard error of the difference (change) in salaries from 2015 to 2016 is much smaller with a paired data approach compared to independent samples. Generally, a paired data approach gives more precise estimates, which means smaller standard errors. But, why? The conceptual explanation is already in the required background readings for this module (page 113). We can illustrate these concepts numerically with results from parts 3a, 3b, and 3c. The

standard deviation of the difference is much bigger for independent samples: 67.773 versus 9.295. A bigger number in the numerator means a bigger *standard error*. But, why is 67.773 bigger than 9.295? Pulling two *independent* samples of 100 salaries breaks the positive correlation. The magic of the paired data approach is caused by the positive correlation that naturally occurs with paired data: people with high 2015 salaries tend to have high 2016 salaries. Remember $V[a + bX + cY] = b^2V[X] + c^2V[Y] + 2bc * SD[X] * SD[Y] * CORR[X, Y]$ from your aid sheets. For a difference, $a = 0$, $b = 1$, and $c = -1$: $V[X - Y] = V[X] + V[Y] - 2 * SD[X] * SD[Y] * CORR[X, Y]$. If X and Y are not correlated then $V[X - Y] = V[X] + V[Y]$, but if X and Y are positive correlated then the $V[X - Y]$ will be much smaller than $V[X] + V[Y]$ because we will subtract a lot with $2 * SD[X] * SD[Y] * CORR[X, Y]$. Hence, $V[Sal_{2016} - Sal_{2015}] < (V[Sal_{2016}] + V[Sal_{2015}])$.

4. **Replicate** the remaining six numbers for $n = 100$ in Table D.3 and the results for $n = 4,000$ in Table D.3.
5. **OPTIONAL: Issues in Merging Data.** Each year, Ontario publishes the required salary disclosure. However, a paired data approach requires merging the data for two different years together. The merge matches each employee's record for one year with that employee's record for another year. If the disclosure included a unique employee number (e.g. the employee's SIN), the merge would be easy: a simple piece of computer code can 100% accurately match people by employee number. Unfortunately employees are only identified by their names, employer name, and title. This creates a host of complexities. Some employees' names change: as examples, sometimes the name includes the middle initial and sometimes not or a person may change their name (e.g. to reflect marital status). Also, some names are common. For example, there are thirteen different people with the name "Brown, David" in the disclosure of 2016 salaries. Also, people enter and leave the disclosure each year because of retirements, changing jobs (leaving the public sector), or crossing the \$100K threshold. Employers' names are often typed in differently from one year to the next: for example, "Université d'Ottawa" for 2016 salaries and "University of Ottawa" for 2015 salaries. They also contain numerous typos. Merging one year's data with another year's data using computer code (i.e. not doing it by hand and going through over 100,000 people) is not 100% perfect (although still excellent). This explains why we get a very slightly different number of observations depending on how the merge is done.
6. **Recall** the more recent data discussed in step 10 on page 79. Note that the 2019 and 2020 salary data exclude the dummies recording the outcome of merges with other years, which are included with the 2014, 2015, and 2016 salary data.

- [on_sal.2020.xlsx](#).
- [on_sal.2019.xlsx](#).

Test/exam examples: In addition to the appearances of the ON public sector disclosure data already listed on page 80, here are some emphasizing the concepts in this module. An especially good example is Andreoni and Vesterlund's "Which is the Fair Sex: Gender Differences in Altruism" published in 2001 and replicated in three different years by students in ECO220Y.

- Question (4), [Summer 2019 Final Exam](#) (with [solutions](#))

- Question (1), [March 2016 Test #4](#) (with [solutions](#))
- Entire test (a one-question test) [February 2015 Test #4](#) (with [rubric](#))
- Questions (1) - (7), [November 2014 Test #2](#) (with [solutions](#))

D.3 Module D.3: Review & Dummy Variables in Regression Analysis

Main concepts: Within a fresh case study, review select topics (graphically summarizing data and analysis, descriptive statistics, and confidence interval estimation of μ). Using dummy variables to include categorical (nominal) variables in regression, including when there are *more than two* categories.

Case studies: We replicate parts of an academic journal article “How Much Energy Do Building Energy Codes Save? Evidence from California Houses,” abbreviated Levinson (2016).

Required readings: The refresher and primer on dummy variables (below). The background for Levinson (2016) (below). Also, a 2015 Freakonomics [podcast](#) gives a great (optional) introduction.

- For dummy variables, recall examples of a simple regression where the x-variable is a dummy from Module B.2 (part 5 on page 49) and Module D.2 (part 2e on page 115). What if there are *more than two categories*? Consider country of birth as an x variable in a regression. Country of birth is clearly a nominal (categorical) variable. Further, it is likely that there are more than two unique values in the data. It requires a *suite* of dummy variables. For example, suppose the data includes people born in Canada, China, the U.S., India and Pakistan. To include country of birth as an x variable requires 4 dummy variables because there are 5 unique countries. But why only 4 and not 5? Our regression already includes a constant term and that captures the average for the 1 category we omit. The coefficients on the other dummy variables tell how each other country on average *differs* from the one left out, which is the reference category (omitted category). We numerically illustrate these ideas today.
- For Levinson (2016), start with the abstract. Generally, an abstract distills an entire research project to a short paragraph that conveys the research question(s), main methods, and findings.

ABSTRACT: Regulations governing the energy efficiency of new buildings have become a cornerstone of US environmental policy. California enacted the first such codes in 1978 and has tightened them every few years since. I evaluate the resulting energy savings three ways: *comparing energy used by houses constructed under different standards, controlling for building and occupant characteristics*; examining how energy use varies with outdoor temperatures; and comparing energy used by houses of different vintages in California to that same difference in other states. All three approaches yield estimated energy savings significantly short of those projected when the regulations were enacted. [emphasis added]

– Modules D.3 and E.2 study the first way (*italicized* above) to answer the research question.

- Levinson (2016) uses the 2003 and 2009 Residential Appliance Saturation Study (RASS) surveys of households. The combined 2003 and 2009 surveys are repeated cross-sectional data: they are *not* the same houses in both years (they are not panel/longitudinal data). Table A.1 shows the subsets analyzed. All analysis is for single-family detached homes.
- Dropping observations with missing values for key variables leaves 19,512 for Figure A.2: “It plots the shares of homes with electric heat and hot water, by year of construction. Both peak for houses built between 1978 and 1982, at 7% for heat and 13% for hot water” (footnote 20).

Table A1. Sample Creation for the RASS Data

	2003	2009	Total
	(1)	(2)	(3)
RASS homes	21,920	25,721	47,641
Single-family detached	11,874	14,263	26,137
Data on income, bedrooms, area, residents, years at address, race, fridges, vintage	8,704	10,808	19,512
No electric heat or hot water	7,961	10,116	18,077
County identified	7,961	10,078	18,039
Part I			
One complete year of monthly bills			
Electricity	7,201	6,844	14,045
Natural gas	6,391	5,967	12,358

Figure of Table A.1: Levinson (2016), p. 5 of the Online Appendix.

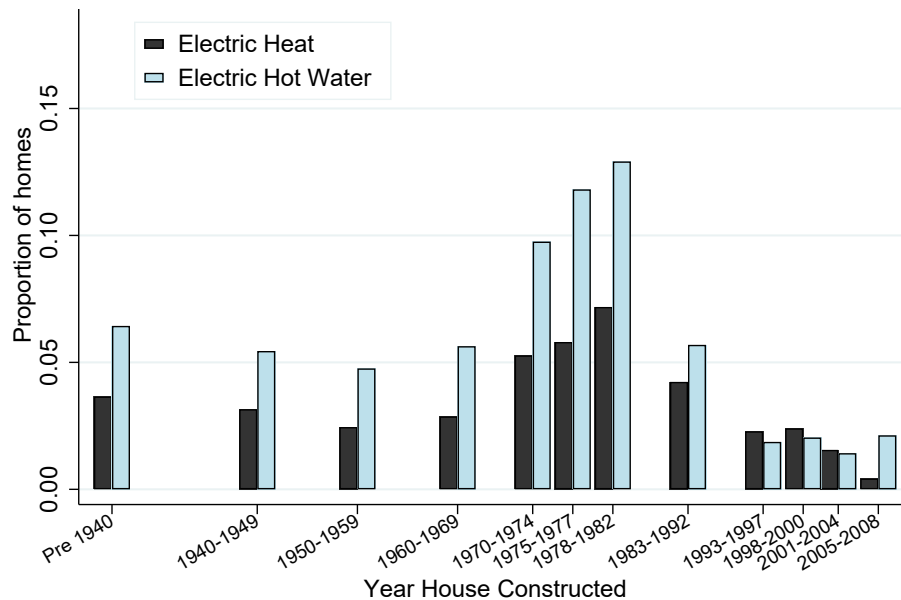


Figure A.2: Proportion of California Homes with Electric Heat or Hot Water. (Note: Combined 2003 and 2009 RASS data for single-family detached homes, $n = 19,512$.) (Levinson (2016), p. 3 of the Online Appendix).

- Keeping only those households that have *neither* electric heat *nor* electric hot water leaves 18,077 observations (see Table A.1) to create Figure 1.
- **Important excerpt, p. 2868 of Levinson (2016):** Figure 1 shows the current average annual household electricity and natural gas used by California houses according to when they were constructed, both measured in millions of British thermal units (MMBTU) of energy. Houses built recently are not using dramatically less energy than older houses built under less strict building energy codes. Newer houses use a third less natural gas but 50 percent *more* electricity. The comparison is not fair, of course, because houses built more recently are larger, have more occupants, and are in less temperate parts of the state, and because both patterns start before the first building codes in 1978.

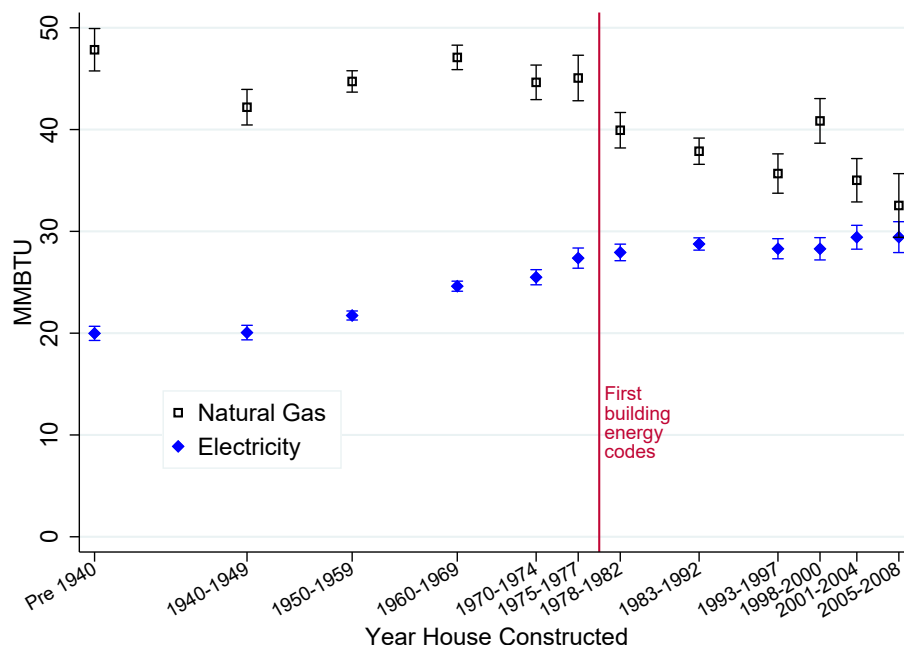


Figure 1: Average Annual Household Energy Use in California. (Note: Combined 2003 and 2009 RASS data for single-family detached homes without electric heat or hot water, $n = 18,077$.) (Levinson (2016), p. 2868).

- If Figure 1 is *not a fair comparison*, how can we measure the impact of building codes (rolled out over the years) on the energy efficiency of homes? Levinson (2016) uses multiple regression to control for differences between newer and older houses (so will you in Module E.2).
- Table 2 (reproduced on page 123) use subsets of 14,045 observations for electricity and 12,358 for natural gas (only homes with a natural gas connection), shown under “Part 1” Table A.1.
 - **Excerpt, p. 2874 of Levinson (2016):** Table 2 provides some descriptive statistics for the RASS and RECS data. For comparison, column 5 reports California values from the American Housing Survey for those variables it has in common with the other two surveys. The AHS is conducted by the Census Bureau and may therefore suffer less from nonresponse bias. In general, the values are similar, though the RASS seems to have lower incomes, smaller household sizes, fewer minorities, and more homeownership.
- Table 2 reports means. For the values in parentheses, a table usually says in the column titles or in the notes whether these are standard deviations (s.d.s) or standard errors (SEs). Table 2 does not. However, these must be s.d.s. The sample sizes (last row of results) are large. All standard error formulas in our course have something like \sqrt{n} in the denominator: as the sample size goes up, sampling error goes down. Hence, SEs would be much smaller (you will compute some today). In contrast, there is no relationship between sample size and the standard deviation of a sample (bigger samples are not systematically more or less variable than smaller samples).

Datasets: For Levinson (2016), to make things easier, we give *three* files: [calif_energy_fig_a2.xlsx](#) (subset to replicate Figure A.2 on page 121), [calif_energy_fig_1.xlsx](#) (subset to replicate Figure 1 on page 122), and [calif_energy_regressions.xlsx](#) (subset to replicate Table 2 on page 123, Table 3 on page 149, Figure 3 on page 151, and Figure 4 on page 151), where “calif_energy” abbreviates paper’s title.

TABLE 2—SELECTED CHARACTERISTICS OF CALIFORNIA SINGLE-FAMILY HOMES

Variables	RASS		RECS 1993–2009		AHS (CA)
	2003 (1)	2009 (2)	California (3)	US (4)	2011 (5)
Annual electricity (MBTU)	23.76 (12.54)	26.70 (13.63)	27.44 (18.01)	44.74 (26.81)	
Annual gas (MBTU)	53.14 (27.90)	43.09 (23.74)	48.31 (31.86)	53.75 (61.47)	
Square feet (1,000s)	1.84 (0.81)	1.93 (0.84)	2.19 (1.15)	2.57 (1.37)	
Bedrooms [Total rooms in RECS]	3.26 (0.86)	3.32 (0.86)	6.27 (1.64)	6.68 (1.79)	3.32 (0.91)
Electric cooking	0.28	0.23			0.28
Remodeled	0.16	0.15			
Years at address	16.3 (14.5)	18.7 (14.7)			13.8 (13.4)
Number of residents	2.90 (1.50)	2.84 (1.49)	3.23 (1.74)	2.89 (1.46)	3.04 (1.62)
Household income [Thousand \$2010]	97.00 (64.45)	92.04 (60.43)	68.25 (40.57)	62.22 (37.10)	102.5 (130.7)
Residents aged 0–5	0.25 (0.66)	0.22 (0.65)			
Residents aged 65+	0.44 (0.75)	0.54 (0.79)	0.34 (0.65)	0.32 (0.64)	0.40 (0.69)
Household head graduated college	0.56	0.60			0.48
Disabled resident	0.094	0.110			0.156
Household head black	0.032	0.030			0.048
Household head Latino [“Hispanic” in AHS]	0.122	0.134			0.216
Own home	0.92	0.93	0.80	0.88	0.78
Central AC	0.48	0.57	0.42	0.60	0.58
Room AC	0.10	0.11			0.13
Refrigerators	1.29 (0.50)	1.38 (0.55)	1.29 (0.49)	1.28 (0.50)	
Observations	7,201	6,844	1,904	15,868	14,692

Notes: RASS 2003 and 2009, single-family homes without electric heat or hot water. California homes in the Residential Energy Consumption Survey (RECS) 1993–2009. American Housing Survey (AHS) 2011, detached single-family homes in California. For gas, there were 6,391 and 5,967 observations in the 2003 and 2009 RASS, respectively.

Figure of Table 2: Levinson (2016), p. 2874.

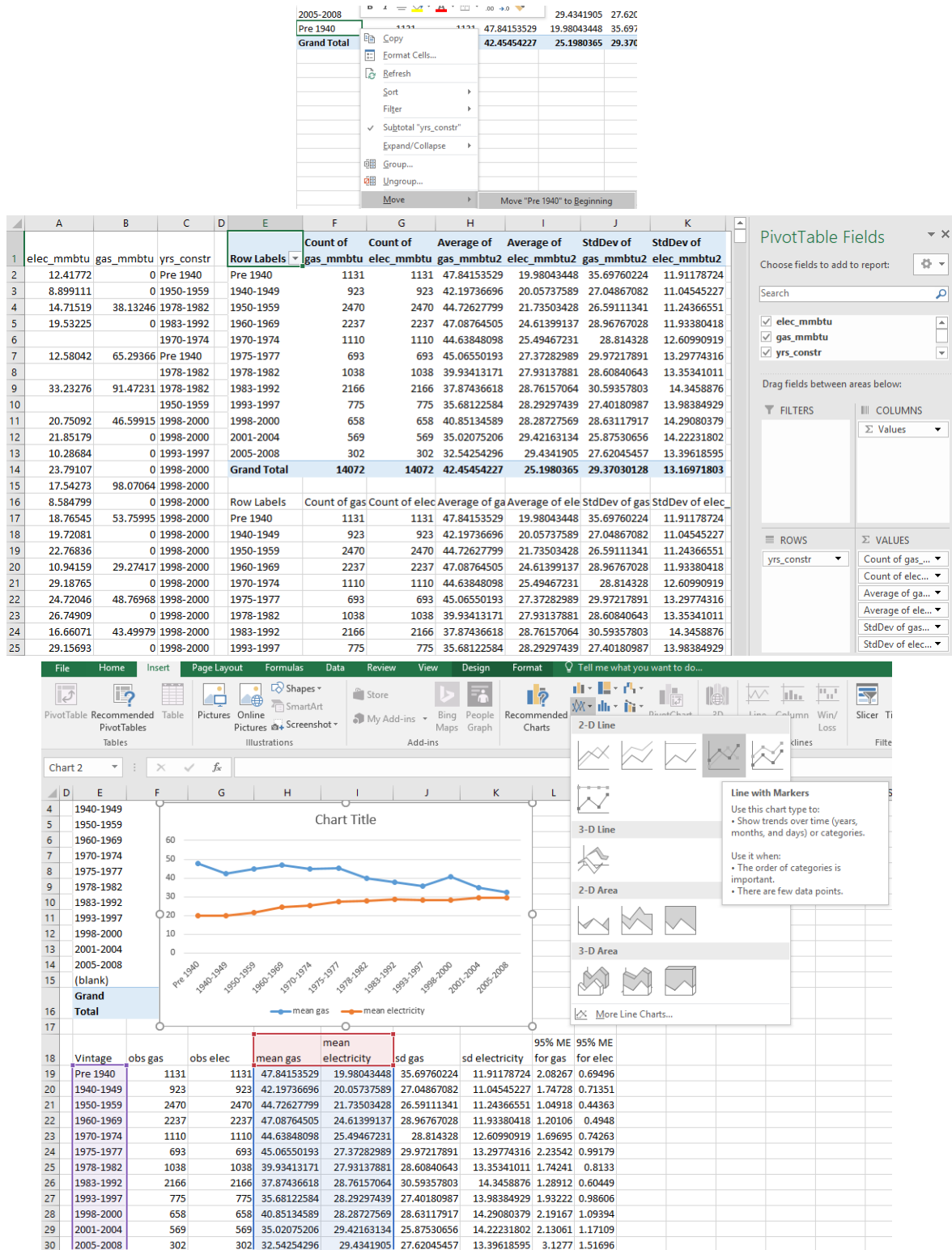
Interactive module materials for Module D.3:

1. For Levinson (2016), use [calif.energy_fig.1.xlsx](#). We start by replicating Figure 1 on page 122.

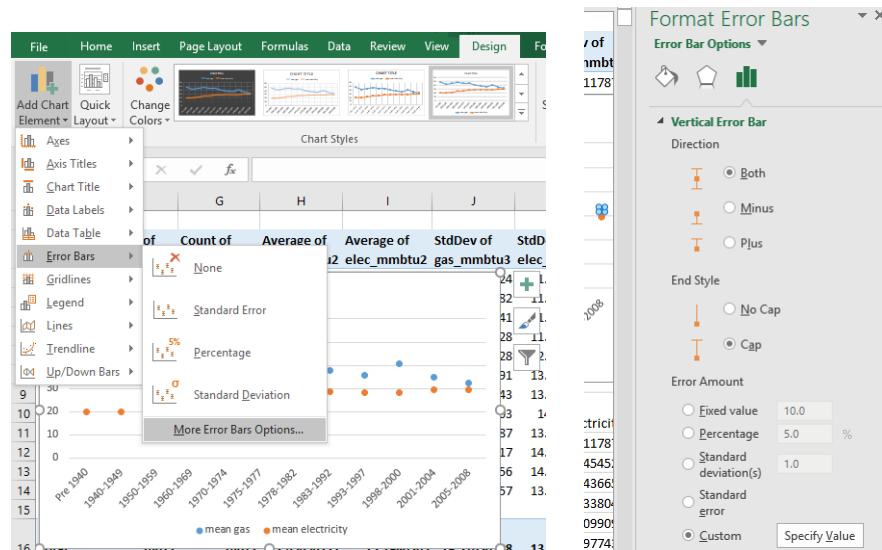
(a) To start, **replicate** the center points of the bands in Figure 1 for gas and electricity.

EXCEL TIPS: Copy yrs_constr, elec_mmbtu, and gas_mmbtu to a new worksheet named “Repl Fig 1.” Insert a PivotTable and drag yrs_constr to ROWS, and three copies each of gas_mmbtu and elec_mmbtu to Σ VALUES. Select Count, Average, and StdDev for each.

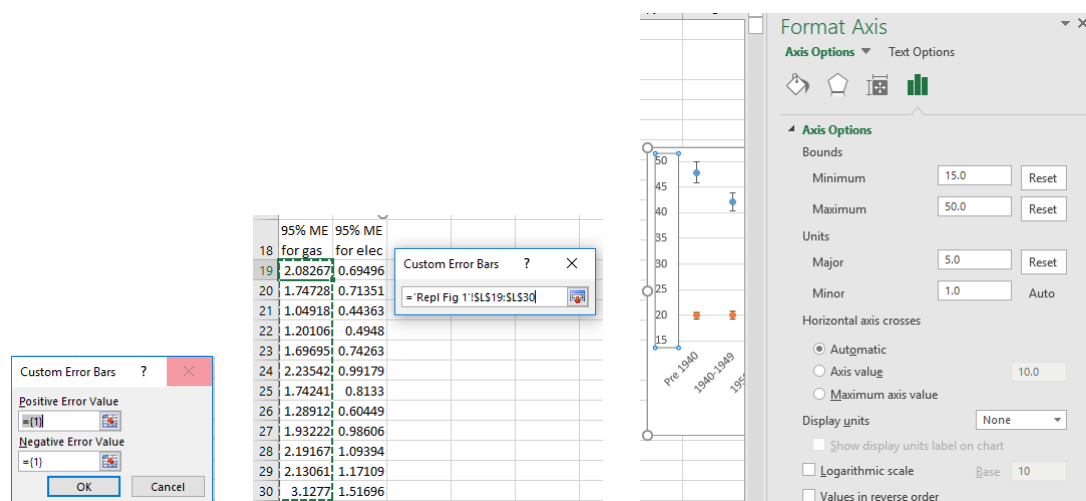
To move Pre 1940 first, right-click the cell with “Pre 1940” and select Move “Pre 1940” to Beginning (see screenshot below). Copy the PivotTable and Paste Special..., Paste Values. After renaming the columns, select vintage, mean gas, and mean electricity (with names) and insert a Line Chart with Markers. To remove the connecting line, right click on each line and select Format Data Series... Under the paint can icon, select No line. (You may think of just using a Scatter Chart in the first place, but that does *not* work.)



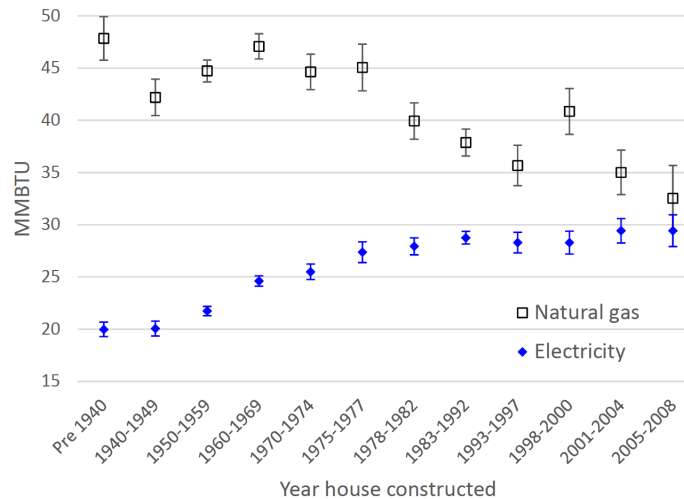
- (b) **Replicate** the error bands. Recall the margin of error (ME) for a mean is after \pm in the CI formula: $\bar{X} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$. Figure 1 shows bands for a 95% confidence level (i.e. $\alpha = 0.05$). **EXCEL TIPS:** In the *copy* of the PivotTable, add columns for the 95% ME. For example, in cell L19 (in the screenshot above) the formula is `=T.INV(0.975,F19-1)*J19/F19^0.5`. Next, click the Add Chart Element button (under the Design tab), select Error Bars, and More Error Bars Options... Select gas or electricity to do first. Under Format Error Bars, under the mini histogram icon, select Custom and Specify Value.



Next, under Specify Value (for Custom Error Bars), use the 95% margins of error created earlier. Select the correct margins of error for the series you are working with (gas or electricity). Enter the same thing for both the Positive Error Value and Negative Error Value. Next, repeat these steps for the other series (gas or electricity). Further, to make the bands more visible, click the values on the y-axis to open Format Axis and under the mini histogram icon, select Minimum and Maximum Bounds to focus on the used part of the plot area.



With some optional further fine-tuning, the replication in Excel is shown below. Note that it places the vintages equidistant on the x-axis (unlike the original Figure 1), which is not ideal, but we will skip the extra steps to fix this (using the variable `yrs.constr_mdpt`).



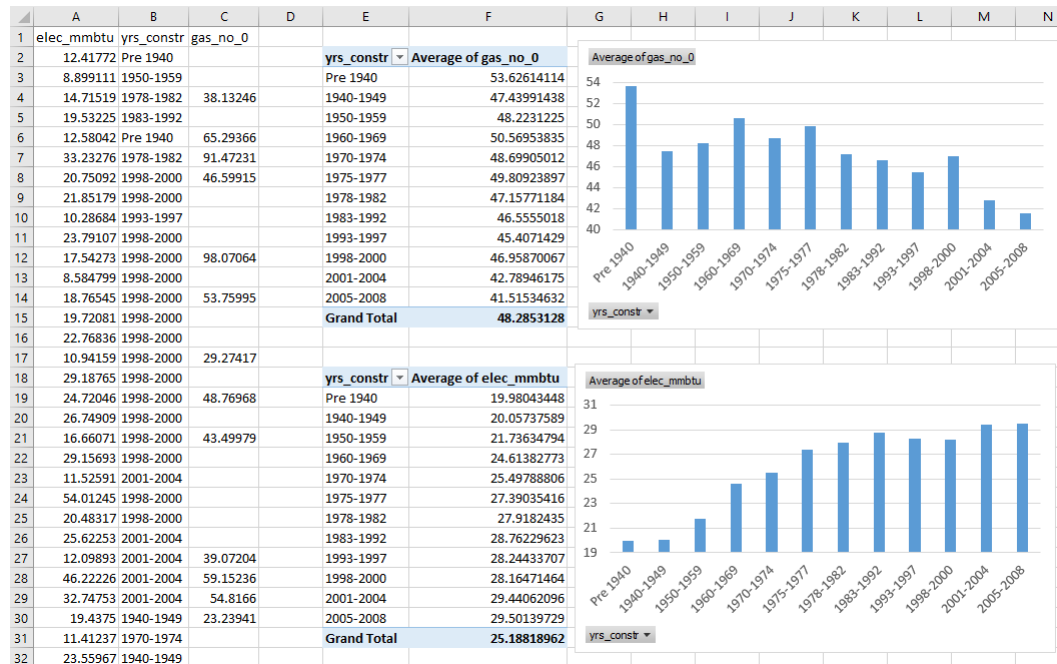
Interpretation tips: How to interpret the figure above (which replicates Figure 1 on page 122)? Levinson (2016) answers on pp. 2867-8. Notice how it gives the context, reminds us of the big research question, describes what the figure shows (including the units), and then cautions us about drawing an incorrect conclusion from the figure. It illustrates our interpretation requirements in Section 5 page 5.

Calculating the energy savings from building energy codes requires knowing how much energy would have been used in the absence of the codes, a far more difficult calculation than is sometimes suggested. Figure 1 illustrates that challenge. It shows the current average annual household electricity and natural gas used by California houses according to when they were constructed, both measured in millions of British thermal units (MMBTU) of energy. Houses built recently are not using dramatically less energy than older houses built under less strict building energy codes. Newer houses use a third less natural gas but 50 percent more electricity. The comparison is not fair, of course, because houses built more recently are larger, have more occupants, and are in less temperate parts of the state, and because both patterns start before the first building codes in 1978. Controlling for those home features, time trends, and the selection of people with high energy demand into recently built homes, is the objective [of this research paper].

2. How does vintage (x) affect energy use (y)? Use [calif_energy_regressions.xlsx](#).

- Brainstorm** how to include the variable yrs_constr on the right-hand-side of a regression. Notice that this variable takes values like “1970-1974,” which are categorical, not interval.
- You may have noticed yrs_constr_mdpt, which is an interval variable. However, using yrs_constr_mdpt forces a linear relationship between energy use and vintage. A line has a *constant* slope: it is either increasing at a constant rate, decreasing at a constant rate, or flat. **Inspect** Figure 1 on page 122 to see if a linear functional form fits. The trend does *not* appear linear. Recalling that the data used for Figure 1 are a somewhat larger subset of the analysis data *and* the figure averages in the zeros for gas (which the regressions do not), use [calif_energy_regressions.xlsx](#) and **check** if the pattern in Figure 1 still holds.

EXCEL TIPS: Copy the variables elec_mmbtu, yrs_constr, and gas_no_0 to a new worksheet. Insert a PivotChart and PivotTable – separately for gas (excluding zeros) and electricity – summarizing the mean for each vintage. (See screenshot below.)



Interpretation tips: What do the above figures show? For annual household gas use in California, houses built before 1940 use the most by far, nearly 54 MMBTUs. Houses constructed in the 1940s use substantially less gas, about 47.5 MMBTUs. However, it creeps back up for houses built in the 50s and 60s. It is fairly steady in the 70s and then generally declines for newer homes, with homes built between 2005 and 2008 (most recent in the data) having by far the lowest gas use, less than 42 MMBTUs. Overall, *the trend is NOT linear*. For electricity, we also see a nonlinear trend (although a different nonlinear trend than gas): quite steady increases until the early 1990s when it leveled off.

- (c) To explain energy use using vintage, we can flexibly include vintage with a group of dummies: one dummy variable for each vintage except one that serves as the omitted (reference) category. In the weeks we spend on multiple regression (coming soon), we will see a lot of dummies. How does it work when there are more than two categories? (With vintage there are 12 categories, ranging from Pre 1940 through 2005-2008.)

- i. **Run** a regression where the y-variable is elec_mmbtu and the x-variables are the eleven vintage dummies constr_40_49 to constr_05_08. Note that constr_pre40 is excluded to serve as the omitted (reference) category. Any category could serve as the reference: let's all pick the same one for now. **Verify** your output matches that given below.

EXCEL TIPS: In the original data, dummy variables for each vintage are already included. Copy the twelve dummy variables – constr_pre40 through constr_05_08 – and the variables elec_mmbtu and gas_no_0 to a new worksheet.

Note: If you needed to make dummies, use the IF function. For example, if the cell AH2 has the vintage of home, then =IF(\$AH2="Pre 1940",1,0) gives constr_pre40, =IF(\$AH2="1940-1949",1,0) gives constr_40_49, ..., and =IF(\$AH2="2005-2008",1,0) gives constr_05_08. You could autofill down to create the variables.

EXCEL TIPS: In the Data tab, click Data Analysis, select Regression. Set the Input Y Range to be elec_mmbtu (including the variable name in the first row). Set the Input X Range to be constr_40_49 through constr_05_08 (also including first row).

Click the Labels box. Choose where to put the output (either to the right of the data or a new worksheet). (If needed, review part 1e on page 41 in Module B.1.)

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	19.98043448	0.379125707	52.7013445	0	19.23729765	20.7235713
constr_40_49	0.076941418	0.5655651	0.136043434	0.891788896	-1.031641425	1.185524261
constr_50_59	1.755913466	0.457827187	3.835319346	0.00012595	0.858511267	2.653315666
constr_60_69	4.633393258	0.465266223	9.958585067	2.76374E-23	3.721409558	5.545376957
constr_70_74	5.517453587	0.539309404	10.23059035	1.76286E-24	4.4603354	6.574571773
constr_75_77	7.409919688	0.615351863	12.04176038	3.12522E-33	6.203748164	8.616091211
constr_78_82	7.937809028	0.548180387	14.48028643	3.53173E-47	6.863302534	9.012315522
constr_83_92	8.781861754	0.467823802	18.77172927	1.1472E-77	7.86486486	9.698858649
constr_93_97	8.26390259	0.59501369	13.88859237	1.44245E-43	7.097596593	9.430208588
constr_98_00	8.184280163	0.626344789	13.06673306	8.59027E-39	6.956561042	9.411999285
constr_01_04	9.46018648	0.656858664	14.40216441	1.07911E-46	8.172656104	10.74771686
constr_05_08	9.520962813	0.830216222	11.46805201	2.60517E-30	7.893628558	11.14829707

Interpretation tips: What do 19.98043448 and 9.520962813 mean? (These are in the first and last values in the first column of results titled “Coefficients.” In a regression to explain annual household electricity use in MMBTUs¹² using a full set of house vintage dummies for a sample of $n = 14,045$ California homes, the constant term says that the average electricity use is about 20 MMBTUs for homes constructed before 1940. Notice how 19.98043448 matches *exactly* with the plain-old mean for homes built before 1940 in part 2b. The other coefficients tell how the mean electricity usage *differs* for homes of other vintages *compared to* those built before 1940 (the reference category). Homes built between 2005 and 2008 use about 9.5 MMBTUs *more* electricity on average compared to homes built before 1940. Notice that 29.50139729 ($= 19.98043448 + 9.520962813$) matches *exactly* with the plain-old mean for homes built from 2005 to 2008 in part 2b.

- ii. Using the mean electricity use by vintage in part 2b, **find** the intercept for a regression like in part 2(c)i but that leaves out constr_05_08 instead of constr_pre40. To check your answer, **regress** elec_mmbtu on the vintage dummies constr_pre40 to constr_01_04 (leaving out constr_05_08). **Verify** your output matches that given below.

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	29.50139729	0.738595068	39.94258637	0	28.05365269	30.94914189
constr_pre40	-9.52096281	0.830216222	-11.468052	2.60517E-30	-11.14829707	-7.89362856
constr_40_49	-9.44402139	0.849500003	-11.1171529	1.36278E-28	-11.10915442	-7.77888836
constr_50_59	-7.76504935	0.781915664	-9.93080162	3.64614E-23	-9.297708081	-6.23239061
constr_60_69	-4.88756956	0.786294494	-6.21595292	5.24578E-10	-6.428811378	-3.34632773
constr_70_74	-4.00350923	0.832250567	-4.8104614	1.52147E-06	-5.634831067	-2.37218739
constr_75_77	-2.11104313	0.88342758	-2.38960519	0.016879612	-3.84267872	-0.37940753
constr_78_82	-1.58315379	0.838026318	-1.88914566	0.058892912	-3.225796867	0.059489296
constr_83_92	-0.73910106	0.787810563	-0.93817104	0.348172662	-2.283314578	0.805112461
constr_93_97	-1.25706022	0.869383496	-1.44592142	0.148221575	-2.961167546	0.4470471
constr_98_00	-1.33668265	0.891119615	-1.50000362	0.133635961	-3.083395657	0.410030358
constr_01_04	-0.06077633	0.912825108	-0.06658048	0.946916635	-1.850034994	1.728482329

Interpretation tips: Why do the regression coefficients above look so different from part 2(c)i? The coefficient on each dummy tells how that group *differs* from the reference group. In part 2(c)i the reference group is pre1940 and in part 2(c)ii it is 2005-08. While the choice of omitted category is arbitrary, it is crucial to the interpretation. Above, the intercept measures mean electricity usage for homes built from 2005 to 2008, which is about 29.5 MMBTUs annually. The negative coefficients

¹²Unfortunately, even with the labels option, Excel regression output does not identify the y-variable.

on all of the dummies mean that homes built earlier all use *less* electricity on average compared to the newest homes (built from 2005 to 2008) in the sample of $n = 14,045$ California homes. The flip side of that same coin is that homes built more recently use *more* electricity on average compared to the oldest homes (built before 1940).

- iii. Using the results in part 2b, **compute** the intercept and the coefficient on `constr_pre40` for a regression of gas use (excluding the zeros) on the vintage dummies, where 2005-08 serves as the reference category: figure out b_0 and b_1 in $\widehat{gas} = b_0 + b_1 \text{constr_pre40} + b_2 \text{constr_40_49} + \dots + b_{11} \text{constr_01_04}$. To check your answer, **regress** `gas_no_0` on the vintage dummies `constr_pre40` to `constr_01_04` (leaving out `constr_05_08`).

EXCEL TIPS: Gas is an extra challenge because of homes with no natural gas connection. Simply selecting all 14,046 rows for the regression returns an error because "" is not a numeric value. One option is to sort the data by the gas variable and only select the non-zero observations as regression inputs. If you sort by the variable with the 0's replaced with "", this means selecting rows 1 through 12,359.

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	41.51534632	1.718753059	24.15434033	5.4947E-126	38.14632194	44.8843707
constr_pre40	12.11079482	1.908438156	6.345919451	2.28834E-10	8.369958025	15.85163161
constr_40_49	5.924568057	1.949278374	3.039364791	0.002375738	2.103678061	9.745458054
constr_50_59	6.70777618	1.804862407	3.716502795	0.000202892	3.16996403	10.24558833
constr_60_69	9.054192027	1.813204221	4.993476146	6.01121E-07	5.500028618	12.60835544
constr_70_74	7.183703797	1.907372292	3.766282979	0.000166465	3.444956265	10.92245133
constr_75_77	8.293892654	2.015270495	4.115523289	3.88812E-05	4.343647795	12.24413751
constr_78_82	5.642365524	1.934913888	2.916080947	0.003550964	1.849632162	9.435098886
constr_83_92	5.040155477	1.829905237	2.754325948	0.005889926	1.453255469	8.627055486
constr_93_97	3.891796586	2.023373515	1.923419753	0.054450325	-0.074331458	7.857924629
constr_98_00	5.443354352	2.042033915	2.665653255	0.007693892	1.440649011	9.446059694
constr_01_04	1.274115434	2.110331589	0.603751297	0.546020096	-2.862464013	5.410694881

3. Use [calif_energy_regressions.xlsx](#). For Table 2 on page 123, **replicate** the first five rows of results (Annual electricity through Electric cooking) including values in parentheses. The note below the table – “For gas, there were 6,391 and 5,967 observations in the 2003 and 2009 RASS, respectively” – means they *exclude* the zeros when computing summary statistics for gas.

	2003	2009
Average of elec_mmbtu	23.76	26.70
StdDev of elec_mmbtu2	12.54	13.63
Average of gas_no_0	53.14	43.09
StdDev of gas_no_0_2	27.90	23.74
Average of sq_feet	1837.59	1911.92
StdDev of sq_feet2	787.64	800.00
Average of bedrooms	3.26	3.32
StdDev of bedrooms2	0.86	0.86
Average of elec_cook	0.28	0.23

EXCEL TIPS: Select all variables in the original data and Insert a PivotTable in a new worksheet. Drag the variable `rass_yr` to COLUMNS and drag two copies of `elec_mmbtu`, `gas_no_0`, `sq_feet`, `bedrooms`, and `elec_cook` to Σ VALUES. (Select `gas_no_0`, *not* `gas_mmbtu`.) To make it more readable, drag Σ Values (Values in a mac) from COLUMNS to ROWS.

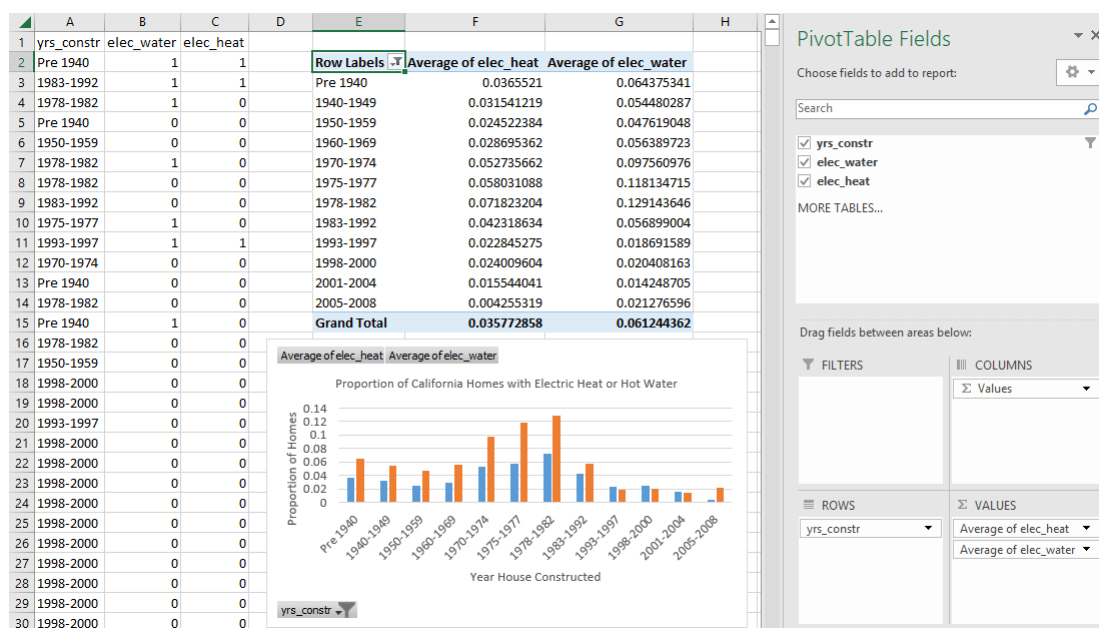
Note: There are some small typos in Table 2 for square feet. For 2003, the s.d. is 0.79 (not 0.81). For 2009, the mean is 1.91 (not 1.93) and the s.d. is 0.80 (not 0.84). Also, the units for annual electricity and gas are MMBTU (not MBTU).

Note: We could have created a new variable for square feet to measure it in 1,000s like in Table 2. However, dividing all values by 1,000 implies the mean and s.d. are each divided by 1,000 so it is easy to see that our replication works (noting the typos mentioned above).

Interpretation tips: What does 0.28 mean? (It is the average for elec_cook for 2003 in the output above.) In the subset of the 2003 RASS data that Levinson (2016) uses in the electricity regressions, 28 percent of homes had *both* an electric oven and an electric stove top.¹³

4. **Replicate** Figure A.2 on page 121 using [calif_energy_fig_a2.xlsx](#).

EXCEL TIPS: Under the Insert tab, Charts, insert a PivotChart (PivotChart & PivotTable).



Test/exam examples: Levinson (2016) and an earlier working paper version have appeared often. You are ready for the second through fourth now. You will be ready for the rest after completing Module E.

- Question (1), [March 2023 Test #4](#) (with [solutions](#))
- Question (9), [March 2022 Test #3](#) (with [solutions](#))
- Question (6), [November 2019 Test #2](#) (with [solutions](#))
- Question (6), [October 2019 Test #1](#) (with [solutions](#))
- Question (5), [April 2022 Final Exam](#) (with [solutions](#))

¹³You may wonder why the standard deviation is not reported for elec_cook. Typically, researchers do not bother when it is redundant. For dummy variables, there is a mechanical relationship: $s = \sqrt{\frac{n}{n-1} \bar{X}(1 - \bar{X})}$. More intuitively, once you know the fraction of 1's in a dummy variable, there is nothing else to summarize for that variable by itself.

- Question (5), [Summer 2019 Final Exam](#) (with [solutions](#))
- Question (10), [April 2019 Final Exam](#) (with [solutions](#))
- Question (3), [April 2019 Test #5](#) (with [solutions](#))
- Part 1, Question (3), [April 2015 Final Exam](#) (with [solutions](#))

D.0.0 Practice questions for Module D

Q1. Recall Carlin et al. (2017) and use [cred_card.xlsx](#). Complete this table for the variable age.

Select Percentiles for Age

5th	
10th	
25th	
50th	
75th	
90th	
95th	

Q2. You should have noticed that the key Excel commands necessary for confidence interval estimation and hypothesis testing for inference about a proportion, the difference between two proportions, a mean, and the difference between two means are: NORM.S.DIST, NORM.S.INV, T.DIST, and T.INV. **For each question below, answer with the appropriate Excel command, not the number.** For example, you would answer =NORM.S.INV(0.995) and not 2.575829304, which is the number that command returns.

- (a) In testing $H_0 : p = 0.5$ versus $H_1 : p < 0.5$, you obtain a test statistic of 0.43. What is the P-value?
- (b) In testing $H_0 : (p_1 - p_2) = 0$ versus $H_1 : (p_1 - p_2) > 0$, what is the critical value if you wish to use a 5% significance level?
- (c) In building a 99% confidence interval estimator of the difference in two means with 40 degrees of freedom, what value must you multiply the standard error by to obtain the margin of error?
- (d) In building a 90% confidence interval estimator of the difference in two proportions, what value must you multiply the standard error by to obtain the margin of error?
- (e) In testing $H_0 : (p_1 - p_2) = 0$ versus $H_1 : (p_1 - p_2) \neq 0$, you obtain a test statistic of -1.54. What is the P-value?
- (f) In testing $H_0 : \mu_d = 0$ versus $H_1 : \mu_d \neq 0$ with 24 degrees of freedom, you obtain a test statistic of -3.21. What is the P-value?
- (g) In testing $H_0 : \mu_3 = 0.32$ versus $H_1 : \mu_3 > 0.32$ with 9 degrees of freedom, you obtain a test statistic of 1.41. What is the P-value?
- (h) In testing $H_0 : (\mu_1 - \mu_2) = 0$ versus $H_1 : (\mu_1 - \mu_2) < 0$ with 18 degrees of freedom, what is the critical value if you wish to use a 0.1% significance level?

Q3. Recall Levinson (2016). The module materials already include a lot of practice with this case (including parts left for you to complete). A few extra practice questions are next.

- (a) Use [calif_energy_fig.1.xlsx](#). Compute the 99.9% CI estimate of the mean number of residents for homes constructed from 1983-1992.
- (b) Use [calif_energy_regressions.xlsx](#). What is the value of the test statistic in the test of whether homes in the 2009 RASS survey are larger (in square feet) on average compared to homes in the 2003 RASS survey? (Do not assume equal variances.)

- (c) Use [calif.energy_regressions.xlsx](#). For homes constructed from 1998-2000, what is the P-value for the test of whether a higher fraction have central air conditioning (AC) in the 2009 RASS survey compared to the 2003 RASS survey?
- (d) Use [calif.energy_fig_a2.xlsx](#). For homes constructed from 1960-1969, what percent have central air conditioning (AC)?

Q4. Recall the “Sparton Resources” case study and use [sparton.xlsx](#).

- (a) Using just the PivotTable output below, compute b_0, b_1, \dots, b_7 for this OLS regression: $uo_conc_i = b_0 + b_1 * d_loc_1_i + b_2 * d_loc_2_i + \dots + b_7 * d_loc_7_i$, where the prefix $d_$ abbreviates dummy (for each location), Location 8 is deliberately excluded, and $i = 1, 2, \dots, 80$.

location	Average of uo_conc
Location 1	0.325
Location 2	0.332
Location 3	0.437
Location 4	0.335
Location 5	0.283
Location 6	0.484
Location 7	0.383
Location 8	0.337

- (b) Continuing with the previous part, run a regression using the data in the worksheet “sparton data reshaped” to check your answer. (You need to first create the dummy variables for each location.)

Q5. Recall Karlan and List (2007) and use [char.give.xlsx](#). Review Table 1. The donor list was *randomly* divided among the control group and the treatment groups. Hence, there should be no systematic differences in the types of people in these groups. Researchers often present results to show that randomization worked, which what Table 1 does. (For any and all parts below that require comparing means, do not assume equal variances.)

- (a) Is there a statistically significant difference in the fraction female between the treatment group (i.e. all treatment groups combined) and the control group? Report the P-value. (**Hint:** Remember to limit your analysis to the non-missing values for the variable female.)
- (b) Is there a statistically significant difference in the highest previous contribution between the treatment group and the control group? Report the P-value.
- (c) Recalling that there are many treatment groups, is there a statistically significant difference in the number of prior donations between the treatment group that had a 3:1 match, a \$100,000 threshold for the matching grant, and a high example amount illustrating the match versus the treatment group that had a 1:1 match, an unstated threshold for the matching grant, and a medium example amount illustrating the match? Report the P-value.

Q6. Using [on_sal_2015.xlsx](#) and [on_sal_2014.xlsx](#) and considering only employees in the “Universities” sector, fill in all of blank lines in the table below. (**Note:** This is like Table D.1 except that it compares different years and focuses on one sector. **Note:** To answer, it is helpful to use the Filter button in the Data tab – filtering by sector and, for the conditional row, by either the disc2014 or disc2015 variable – and then to copy the filtered data to a new worksheet. Alternatively, you can use a PivotTable.)

Universities Sector: Comparing 2015 with 2014

	2014 Salaries (CAN \$1,000's)	2015 Salaries (CAN \$1,000's)	Change
Unconditional	_____ (_____) [_____]	_____ (_____) [_____]	_____
Conditional on employee having both her/his 2014 salary and 2015 salary disclosed: “same-employee” comparison	_____ (_____) [_____]	_____ (_____) [_____]	_____

Notes: Shows means with standard deviations in parentheses and number of observations in square brackets.

Q7. Consider one of the largest employers in the Ontario public sector salary disclosures: the “Toronto Police Service.” This employer listed 4,758 employees making at least \$100K in 2016 and 4,636 in 2015.

- (a) Use [on_sal_2015.xlsx](#) and [on_sal_2016.xlsx](#) to fill in the blanks. Follow the answer guides given in square brackets.

Among *all* ON public sector employees of the Toronto Police Service making at least \$100K in 2016 who also made at least \$100K in 2015, the mean salary in 2016 is _____ dollars [answer in dollars] and the standard deviation of salary is _____ dollars [answer in dollars]. Among *all* ON public sector employees of the Toronto Police Service making at least \$100K in 2015 who also made at least \$100K in 2016, the mean salary in 2015 is _____ dollars [answer in dollars] and the standard deviation of salary is _____ dollars [answer in dollars]. These four numbers are _____ [Answer with: parameters or statistics]. In comparing 2016 with 2015, this paragraph uses _____ [Answer with: an unconditional or a conditional (“same-police”)] approach.

- (b) Consider an independent samples (unequal variances) method of inference about the difference in mean salaries from 2015 to 2016 conditional on the employee having her/his salary disclosed in both 2015 and 2016. To make inference necessary, suppose that you did not have access to *all* relevant salaries in each year, only random samples from each. To use the same random sample as the answer key, use [on_sal_2016.xlsx](#) and sort *the relevant population subset* by random1 and take only the first 250 observations. Similarly, use [on_sal_2015.xlsx](#) and sort *the relevant population subset* by random1 and take only the first 250 observations. Use your random samples to conduct this hypothesis test $H_0 : \mu_{\text{samepolice16}} - \mu_{\text{samepolice15}} = 0$ versus $H_1 : \mu_{\text{samepolice16}} - \mu_{\text{samepolice15}} > 0$. What is the P-value?

Q8. Recall the Karlan and List (2007) data and tables of results explored in Modules C.2 and D.1. Consider the second and third rows of results in Panel A, Columns (1) and (2) in Table 2A: “Dollars given, unconditional” and “Dollars given, conditional on giving.” Use [char_give.xlsx](#). (For any and all parts below that require comparing means, do not assume equal variances.)

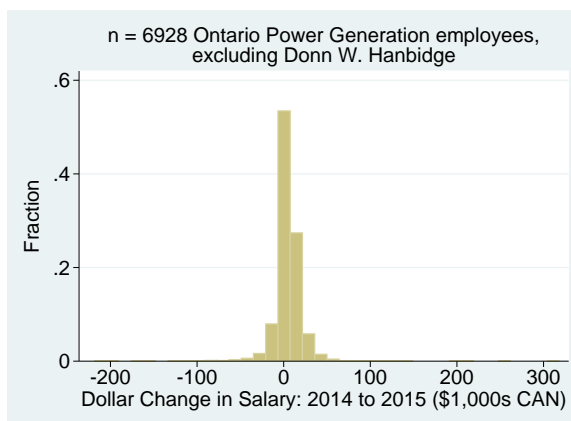
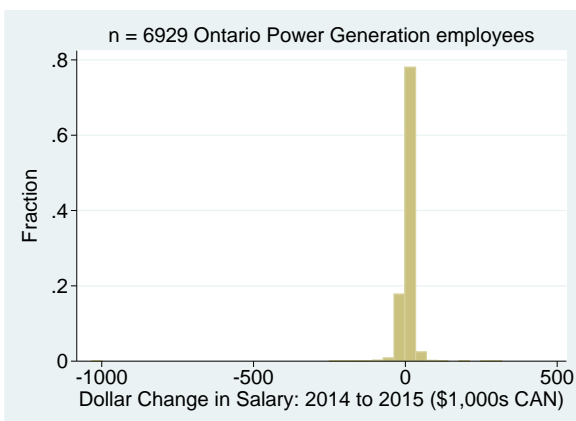
- (a) Is there a statistically significant difference in the dollars given, unconditional between the treatment and control group? Report the P-value. (If you find it convenient, you may use the worksheet “HT Diff. in Mean Giving” in [char_give.xlsx](#) as a template.)
- (b) Amongst females, is there a statistically significant difference in the dollars given, unconditional between the treatment and control group? Report the P-value.
- (c) What is the 99% confidence interval estimate of the difference in the dollars given, conditional on giving, between the treatment and control group? Report the point estimate and the margin of error. (If you find it convenient, you may use the worksheet “CI Est. of Diff. in Mean Giving” in [char_give.xlsx](#) as a template.)

For extra practice, additional questions, with an ^e superscript (*e* for extra), are next.

Q^e1. For all parts of this question, use [on_sal_15.14.xlsx](#) to analyze mean increases in salaries from 2014 to 2015 for “same-employees” in the “Ontario Power Generation” sector.

- (a) For these 6,929 employees, Donn W. Hanbidge is an outlier. His salary dropped by over a million dollars between 2014 and 2015 (<http://www.ontariosunshinelist.com/people/zkrdtf>). The second histogram below excludes him. Is the population, without Hanbidge, Normally distributed? To answer, fill in the blanks below.

Recall the Empirical Rule for samples drawn from a Normal populations. It says that about 68.3 percent of observations lie within one standard deviation of the mean, about 95.4 percent lie within two standard deviations of the mean, and about 99.7 lie within three standard deviations of the mean. Even without Hanbidge, this salary change distribution is *not* Normal. In that distribution ($N = 6,928$), _____ percent lie within one standard deviation of the mean, _____ percent lie within two standard deviations of the mean, and _____ percent lie within three standard deviations of the mean. [For all blanks, answer rounding to the nearest first decimal place to match the rounding used when stating the Empirical Rule.]



- (b) Use a paired data approach for inference about the mean *dollar change* in “same-employee” salaries between 2014 and 2015 for “Ontario Power Generation” employees. To make inference necessary, suppose that you did not have access to *all* relevant salaries, only a random

sample. To use the same random sample as the answer key, use [on_sal_15_14.xlsx](#) and sort *the relevant population subset* by random1 and take only the first 100 observations.

- i. Does the random sample include an outlier?
 - ii. Use your random sample to conduct this hypothesis test $H_0 : \mu_d = \$3,000$ versus $H_1 : \mu_d > \$3,000$. What is the P-value?
 - iii. Use your random sample to build a 90% confidence interval estimate of the mean difference. Report the point estimate and the margin of error.
 - iv. Now check your inferences in the previous two parts against the truth. In other words, see if your conclusions are consistent with the true population parameters.
 - v. What if you had used an independent samples approach instead? Defining X to be the 2015 salary and Y to be the 2014 salary and continuing to work with the random sample from the previous parts, appropriately plug into $V[a + bX + cY] = b^2V[X] + c^2V[Y] + 2bc * SD[X] * SD[Y] * CORR[X, Y]$ and verify that your answer matches the standard deviation of the difference variable. What would it be if there were no positive correlation between salaries (as would be expected with an independent samples approach)?
- (c) Using the same random sample as the previous part, make an inference about the mean *percent change* in “same-employee” salaries between 2014 and 2015 for “Ontario Power Generation” employees.
- i. Use your random sample to conduct this hypothesis test $H_0 : \mu_d = 4\%$ versus $H_1 : \mu_d > 4\%$. What is the P-value?
 - ii. Use your random sample to build a 95% confidence interval estimate of the mean difference. Report the point estimate and the margin of error.
 - iii. Now check your inferences in the previous two parts against the truth. In other words, see if your conclusions are consistent with the true population parameters.

Answers for Module D practice questions:

A1. See table below. (Be careful to select all observations: the age variable has some missing values.)

Note: With this large sample, you get the exact same answers whether you use the function PERCENTILE.INC or PERCENTILE.EXC or use other software like Stata. This is typical (it usually doesn't really matter), which is good because there is not a clear "correct" way. (There was a difference for the Sparton example because the sample sizes were tiny.)

Select Percentiles for Age

5th	20
10th	22
25th	25
50th	30
75th	39
90th	52
95th	57

A2. (a) =NORM.S.DIST(0.43,TRUE)

(b) =NORM.S.INV(0.95)

(c) =T.INV(0.995,40)

(d) =NORM.S.INV(0.95)

(e) =2*NORM.S.DIST(-1.54,TRUE)

(f) =2*T.DIST(-3.21,24,TRUE)

(g) =1-T.DIST(1.41,9,TRUE)

(h) =T.INV(0.001,18)

A3. (a) For the 2,615 homes constructed from 1983-1992, the point estimate is 2.9196941 residents with a margin of error of 0.0922048, which yields a 99.9% CI estimate of [2.82749, 3.01190]. (There are no missing values in these data for the variable measuring the number of residents.)

(b) Conduct a right-tailed hypothesis test for the difference in means, independent samples (not assuming equal variances). The value of the t test statistic is 5.545 with 13,981 degrees of freedom. (The P-value is basically zero, which means we have overwhelming evidence to support the conclusion that homes are in fact larger on average in the 2009 RASS survey compared to the 2003 RASS survey.) (There are no missing values in these data for the variable measuring square feet.)

(c) Conduct a right-tailed hypothesis test for the difference between proportions. The value of the z test statistic is 0.215141434 and the P-value is 0.414828525. (The P-value is large, which means we lack support for the conclusion that a higher fraction of homes of this vintage (1998-2000) have central AC in the 2009 survey compared to 2003. This lack of difference is not surprising because in the vast majority of cases central AC is installed during the construction of a home (not later) and we are comparing homes constructed during the same period.) (There are no missing values in these data for the variable measuring whether or not a home has central AC.)

- A4.** (a) If Location 8 is excluded that means it is the reference category (i.e. omitted category) for this suite of dummies. The intercept will simply be the mean UO concentration for Location 8: $b_0 = 0.337$, which we obtain from the PivotTable provided with the question (which you could also construct yourself). The other coefficients tell how each other location *differs* from Location 8. For example, the mean UO concentration at Location 1 is 0.012 *lower* than Location 8 ($-0.012 = 0.325 - 0.337$) so $b_1 = -0.012$. Similarly, we find the other coefficients: $b_2 = -0.005$, $b_3 = 0.1$, $b_4 = -0.002$, $b_5 = -0.054$, $b_6 = 0.147$, and $b_7 = 0.046$.
- (b) See the screenshot below showing the data and the regression output from Regression in Data Analysis Tools.

E2					=IF(\$A2="Location 1",1,0)																
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O						
1	location	loc_desc	batch	uo_conc	d_loc_1	d_loc_2	d_loc_3	d_loc_4	d_loc_5	d_loc_6	d_loc_7										
2	Location 1	China (1)	1	0.32	1	0	0	0	0	0	0		SUMMARY OUTPUT								
3	Location 1	China (1)	2	0.38	1	0	0	0	0	0	0										
4	Location 1	China (1)	3	0.58	1	0	0	0	0	0	0										
5	Location 1	China (1)	4	0.61	1	0	0	0	0	0	0										
6	Location 1	China (1)	5	0.12	1	0	0	0	0	0	0										
7	Location 1	China (1)	6	0.13	1	0	0	0	0	0	0										
8	Location 1	China (1)	7	0.48	1	0	0	0	0	0	0										
9	Location 1	China (1)	8	0.03	1	0	0	0	0	0	0										
10	Location 1	China (1)	9	0.43	1	0	0	0	0	0	0										
11	Location 1	China (1)	10	0.17	1	0	0	0	0	0	0										
12	Location 2	China (2)	1	0.22	0	1	0	0	0	0	0										
13	Location 2	China (2)	2	0.28	0	1	0	0	0	0	0										
14	Location 2	China (2)	3	0.31	0	1	0	0	0	0	0										
15	Location 2	China (2)	4	0.37	0	1	0	0	0	0	0										
16	Location 2	China (2)	5	0.39	0	1	0	0	0	0	0										
17	Location 2	China (2)	6	0.45	0	1	0	0	0	0	0										
18	Location 2	China (2)	7	0.44	0	1	0	0	0	0	0										
19	Location 2	China (2)	8	0.13	0	1	0	0	0	0	0										
20	Location 2	China (2)	9	0.32	0	1	0	0	0	0	0										
21	Location 2	China (2)	10	0.41	0	1	0	0	0	0	0										
22	Location 3	China (3)	1	0.71	0	0	1	0	0	0	0										
23	Location 3	China (3)	2	0.22	0	0	1	0	0	0	0										
24	Location 3	China (3)	3	0.78	0	0	1	0	0	0	0										
25	Location 3	China (3)	4	0.15	0	0	1	0	0	0	0										
26	Location 3	China (3)	5	0.19	0	0	1	0	0	0	0										

- A6.** See complete table:

Universities Sector: Comparing 2015 with 2014

	2014 Salaries (CAN \$1,000's)	2015 Salaries (CAN \$1,000's)	Change
Unconditional	145.493 (39.612) [16,373]	147.332 (41.485) [17,063]	1.839
Conditional on employee having both her/his 2014 salary and 2015 salary disclosed: "same-employee" comparison	146.849 (39.537) [15,201]	151.121 (41.446) [15,202]	4.271

Notes: Shows means with standard deviations in parentheses and number of observations in square brackets. For why 15,201 \neq 15,202, see part 5 on page 118.

- A7.** (a) (**Note:** To answer, it is helpful to use the Filter button in the Data tab – filtering by both employer and either the disc2015 or disc2016 variable – and then to copy the filtered data to a new worksheet.) Among *all* ON public sector employees of the Toronto Police Service making at least \$100K in 2016 who also made at least \$100K in 2015, the mean salary in 2016 is 122,012 dollars and the standard deviation of salary is 18,056 dollars. Among *all* ON public sector employees of the Toronto Police Service making at least \$100K in 2015 who also made at least \$100K in 2016, the mean salary in 2015 is 122,141 dollars and the standard deviation of salary is 18,578 dollars. These four numbers are parameters. In comparing 2016 with 2015, this paragraph uses a conditional (“same-police”) approach. (**Note:** To compute σ , use the Excel function STDEV.P. The function STDEV.S, returns s , which is the sample standard deviation. Usually you use STDEV.S: this is a special case because the data we are working with are the entire population, not a random sample.)
- (b) The relevant subset for 2016 are observations with employer=“Toronto Police Services” and disc2015=1. The relevant subset for 2015 are observations with employer=“Toronto Police Services” and disc2016=1. Following the random sampling method given in the question yields: $(\bar{X}_{samepolice16} - \bar{X}_{samepolice15}) = -1.8737295$, $SE(\bar{X}_{samepolice16} - \bar{X}_{samepolice15}) = 1.651477$, $t = -1.134578$, $\nu = 494.49326$, and P-value = 0.87144907. (There is no way we can prove same-police salaries have gone up from 2015 to 2016 when in our random samples salaries actually dropped on average by nearly \$2,000.)
- A8.** (a) There is not a statistically significant difference at a 5% significance level, but there is at a 10% significance level. The difference in the amount donated is 15 cents. The standard error of this differences is 8.01 cents. The t test statistic is -1.9182626. The degrees of freedom are 36216.06. The P-value is 0.05508557.
- (b) There is not a statistically significant difference at any conventional significance level. Among females, the difference in the amount given between the control and treatment groups is 5 cents. The standard error of this differences is 13.68 cents. The t test statistic is -0.37831525. The degrees of freedom are 9870.9201. The P-value is 0.70520455.
- (c) The point estimate is \$1.6683935 and the margin of error is \$7.3763073.

Answers to the additional questions for extra practice.

- A^e1.** (a) In that distribution, 85.3 percent lie within one standard deviation of the mean, 96.2 percent lie within two standard deviations of the mean, and 98.4 percent lie within three standard deviations of the mean.

- (b) i. No.
- ii. The t test statistic is 2.092257 and the P-value is 0.01948680.
- iii. The point estimate is \$5,841 and the margin of error is \$2,254.
- iv. In the population ($N = 6,929$), the mean change in salaries is $\mu_d = \$5,006$. We did well with our inferences: we inferred at a 5% significance level that μ_d is greater than \$3,000, and it is ($\$5,006 > \$3,000$). Also, our 90% CI estimate of $\$5,841 \pm \$2,254$ does include the true parameter.
- v. $V[a + bX + cY] = b^2V[X] + c^2V[Y] + 2bc * SD[X] * SD[Y] * CORR[X, Y] = V[X - Y] = V[X] + V[Y] - 2 * SD[X] * SD[Y] * CORR[X, Y] = 39.28423^2 + 34.22107^2 - 2 * 39.28423 * 34.22107 * 0.9410 = 184.3$, which matches the standard deviation of the variable measuring the dollar difference in salaries. If there were no positive correlation, $V[X - Y] = V[X] + V[Y] = 39.28423^2 + 34.22107^2 = 2,714.3$.
- (c) i. The t test statistic is 0.20763288 and the P-value is 0.41797107.
- ii. The point estimate is 4.2% and the margin of error is 1.8%.
- iii. In the population ($N = 6,929$), the mean percent change in salaries is $\mu_d = 4.004\%$. We did OK with our inferences. We were not able to prove at any conventional significance level that salaries rose by more than 4% even though they did. However, our 95% CI estimate of $4.2\% \pm 1.8\%$ does include the true parameter.

E Module E: Multiple Regression

E.1 Module E.1: The Big Idea of Multiple Regression & Applied Research

Concepts: The *big* difference between simple and multiple regression. How correlation does *not* relate to a slope coefficient in multiple regression, like it does for simple regression. Using logarithms.

Case studies: We replicate parts of an academic journal article “Toward an Understanding of Learning by Doing: Evidence from an Automobile Assembly Plant,” abbreviated Levitt et al. (2013).

Required readings: Sections 20.1 - 20.3 and “Logarithms in Regression Analysis with Asiaphoria” (Quercus). Also, this background reading for Levitt et al. (2013), who use extensive data at the daily level for an automobile manufacturer that started production of redesigned vehicles.

- “Figure 1 plots the average number of defects per car by week. When production begins in mid-August, average defect rates were around 75 per car. Eight weeks later, they had fallen by two-thirds, to roughly 25 defects per car. These strong initial learning effects are consistent with findings in the broader literature on learning by doing.” (pp. 653-4)

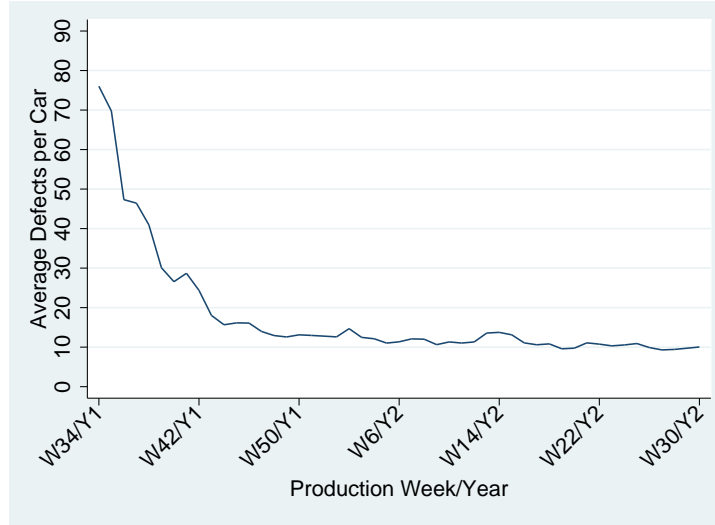


Figure 1: Levitt et al. (2013), p. 654.

Notes: Average defect rates per car. The figure plots the average number of production defects per car by week over the production year. Weeks are labeled on the horizontal axis; for example, W34/Y1 indicates the thirty-fourth calendar week of calendar year 1 (the production spanned two calendar years, from August of year 1 to June of year 2).

- The researchers start with Equation 1, a simple empirical model of the learning process, where t indexes either a day or a week, D_t is the average defects per car in a time period, and E_t is the production experience to date (cumulative production). They use natural log transformations.

$$\ln(D_t) = \alpha + \beta \ln(E_t) + \varepsilon_t \quad (1)$$

They also try Equation 2, an alternative model with a time trend, where t is a variable measuring the number of time periods since the start of production.

$$\ln(D_t) = \alpha + \beta \ln(E_t) + \gamma * t + \varepsilon_t \quad (2)$$

- “Table 1 shows the results of estimating these specifications with our sample. Panel A shows results from specifications using weekly data (average defect rates over the week and production experience at the week’s outset); Panel B shows results using daily observations.” (p. 655)

Table 1: Estimates of Learning By Doing		
	(1)	(2)
Panel A. Weekly Data		
Estimated learning rate, $\hat{\beta}$	-0.289* (0.007)	-0.335* (0.017)
Time trend		0.007* (0.002)
Observations	47	47
R^2	0.961	0.969
Panel B. Daily Data		
Estimated learning rate, $\hat{\beta}$	-0.306* (0.006)	-0.369* (0.014)
Time trend		0.001* (0.0002)
Observations	224	224
R^2	0.931	0.943

Notes: Column (1) in both panels shows estimation results for $\ln(D_t) = \alpha + \beta \ln(E_t) + \varepsilon_t$, where D_t is the average defects per car in time period t and E_t is production experience up to that point: cumulative number of cars produced before the current period. Column (2) in both panels shows estimation results for $\ln(D_t) = \alpha + \beta \ln(E_t) + \gamma * t + \varepsilon_t$. Heteroskedasticity-robust standard errors are in parentheses. * Significant at the 5 percent level.

Figure of Table 1: A clarified version of Table 1 on p. 655 of Levitt et al. (2013).

- “The simple empirical model fits the data very well, with the R^2 of the weekly and daily specifications at 0.961 and 0.931, respectively. This fit can also be seen in Figure 2, which plots the logged average defect rate against cumulative production in the daily data.” (p. 656)

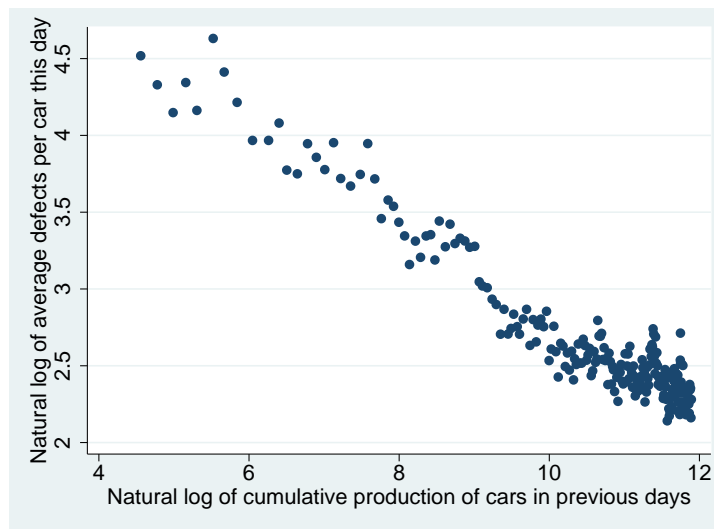


Figure 2: Levitt et al. (2013), p. 656.

Notes: Log defects per car versus log production experience (cumulative output), daily data. The figure plots daily data on the (logged) average number of production defects per car versus (logged) cumulative production. Cumulative production is the cumulative number of cars produced before the day of observation. [This plots Panel B, Column (1) in Table 1.]

- The coefficient on the time trend in Table 1, Panel A, Column (2) is *positive*. But, Figure 1 shows a strong and clear *negative* association between average defects per car and the time trend. Why

is there a big discrepancy? This is *not* about the natural log: the simple correlation between $\ln(\text{average defects per car})$ and the time trend is also strongly negative. In general, a *multiple* regression coefficient can have the opposite sign as the correlation. This dramatically illustrates the big idea of multiple regression: each coefficient measures the change in y associated with a change in that x variable *after controlling for the other included x variables*. Column (2) gives the *estimates* of Equation 2. There are two common notations for parameter estimates: (1) a (for α), b (for β), and g (for γ), or (2) $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\gamma}$. While g would certainly be negative in $\ln(\widehat{D_t}) = a + g * t$, it is *not* surprising that in $\ln(\widehat{D_t}) = a + b\ln(E_t) + g * t$ the value of g is not negative ($g = 0.007$). Production experience (E_t) drives down defects, not time passing. Once we control for experience, the simple passage of time becomes virtually irrelevant.

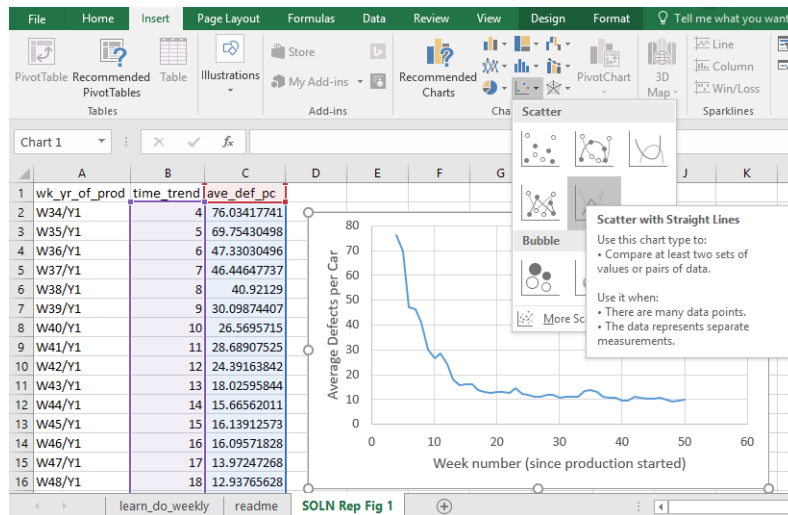
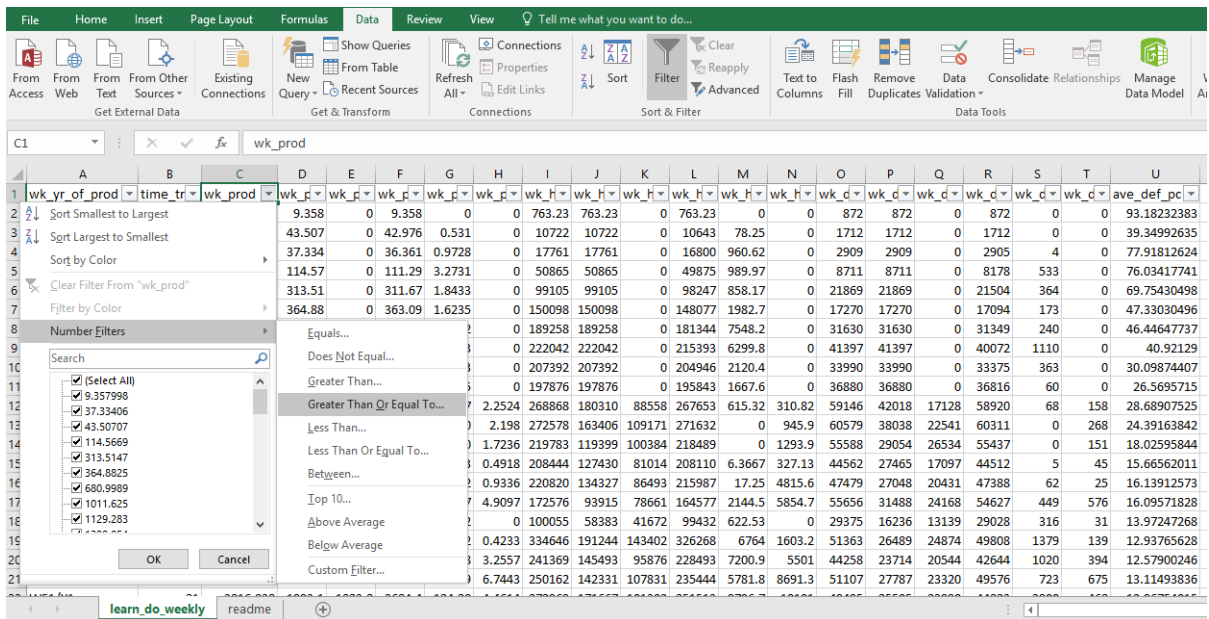
- Next is some background about automobile production needed to understand the data.
 - Major automobile manufacturers (e.g. Toyota, Ford, etc.) each offer many models of vehicles and introduce new versions (with minor to major revisions) each year. For example, Toyota offers these related models: Yaris, Corolla, and Camry. As consumers, we may view these very differently, but from a production standpoint they have much in common. In the data, the models are simply referred to as model 1, model 2, and model 3. The authors do not disclose which manufacturer provided the data: if they used model names (e.g. “Camry,” “Civic,” or “Focus”) that would give away the identity of the manufacturer (e.g. Toyota, Honda, or Ford).
 - Shift work is common in industries ranging from automobile manufacturing to nursing. The first shift is sometimes called the “day shift” and the second shift, the “night shift.” For example, consider this [job posting](#) in Ontario for joining Toyota’s “Production Team” (retrieved February 21, 2017): “Shift start and end times may vary based on business condition. Our core hours of work are Monday to Friday with a day shift [first shift] starting as early as 6:15 a.m. and ending at 3:45 p.m. and an afternoon shift [second shift] starting at 5:45 p.m. and ending as late as 4:15 a.m.” Production of a new or revised model may start with only a first shift. Adding a second shift for a model can ramp production up to full capacity.

Datasets: For Levitt et al. (2013): [learn_do_weekly.xlsx](#), where “learn_do” abbreviates “Learning by Doing” from the title and “weekly” refers to the fact that these are the weekly (not daily) data.

Interactive module materials for Module E.1:

1. For all parts, use [learn_do_weekly.xlsx](#) for Levitt et al. (2013). **Browse** the data.
2. **Replicate** Figure 1 on page 141. **Create** the y-axis variable and name it ave_def_pc. Like the authors, display only weeks with production of at least 100 cars in the figure. (For the x-axis tick values in Figure 1, just use the week number.)

EXCEL TIPS: Add ave_def_pc to the original data (i.e. in Column U). To restrict to weeks with production of at least 100 cars, use the Filter tool, which allows you to specify a range. (See screenshot below.) Copy the variables wk_yr_of_prod, time_trend, and ave_def_pc from the filtered data to a new worksheet. Insert a Scatter Chart with Straight Lines (see screenshot below).



- Replicate** the simple regression in Table 1, Panel A, Column (1). First, **create** a variable named *cum_prod* for the *cumulative* production in *previous* weeks. Next, create the x variable named *ln_cum_prod*. Create the y variable, *ln_ave_def_pc*. **Run** the regression: use only weeks when at least 100 cars are produced (your number of observations should match the table). (Note: Cumulative production includes *all* weeks, including those with production below 100 cars.) To visualize, see Figure E2 on page 145. Excel cannot compute robust standard errors: you will obtain (0.009) instead of (0.007).

EXCEL TIPS: Go back to the original data (removing the Filter) and create the cumulative *previous* production variable in Column V. Use the function `=SUM(C$2:C2)` in row 3 of the new variable. The sum is up to but not including row 3. Autofill down. Check that row 51 has `=SUM(C$2:C50)` and evaluates to 143074.0461. In Columns W and X, create *ln_cum_prod* and *ln_ave_def_pc*. (Note: When creating *ln_cum_prod*, do not apply the LN function to the empty cell in row 2.) Again, use the Filter tool to select weeks with production of at least 100 car. Copy and paste the filtered data to a new worksheet. Use Regression under Data Analysis.

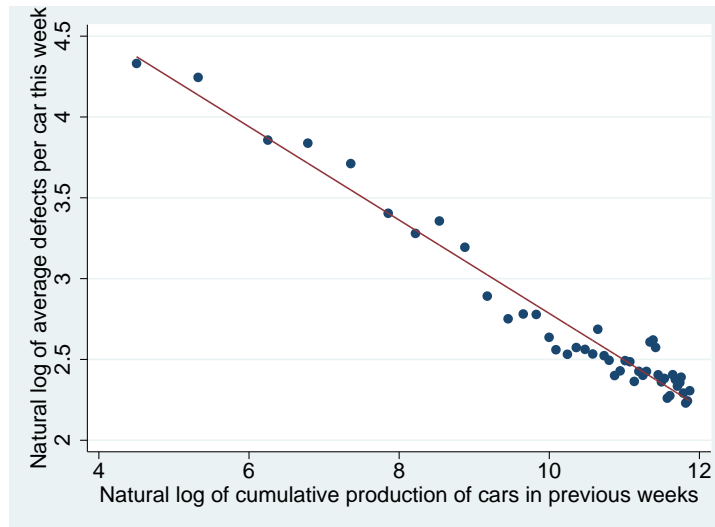


Figure E2: A figure like Figure 2 on p. 656 of Levitt et al. (2013), *except that* it shows the weekly data (instead of the daily data). This figure corresponds exactly to Table 1, Panel A, Column (1).

4. **Replicate** the multiple regression in Table 1, Panel A, Column (2). With regular (not robust) standard errors you get (0.016), not (0.017). But, rounded, it's (0.02) either way.

EXCEL TIPS: From the worksheet for part 3, copy the variables `time_trend`, `ln_cum_prod`, and `ln_ave_def_pc` to a new worksheet. (This step puts the x variables in adjacent columns, as required by the multiple regression tools in Excel.) Use Regression under Data Analysis.

	A	B	C	D	E	F	G	H	I	J	K
1	time_trend	ln_cum_prod	ln_ave_def_pc		SUMMARY OUTPUT						
2	4	4.50201976	4.331182942								
3	5	5.321868	4.244979139								
4	6	6.250517041	3.857150787		Regression Statistics						
5	7	6.78351004	3.838300625		Multiple R	0.984217189					
6	8	7.355105578	3.711650466		R Square	0.968683476					
7	9	7.853910447	3.404483446		Adjusted R Square	0.967259997					
8	10	8.217457465	3.279766632		Standard Error	0.096358562					
9	11	8.535646699	3.356516397		Observations	47					
10	12	8.87553107	3.194240386		ANOVA						
11	13	9.173504394	2.891812855			df	SS	MS	F	Significance F	
12	14	9.451097264	2.751468509		Regression	2	12.63693147	6.318465733	680.5045227	8.07269E-34	
13	15	9.652888662	2.781246493		Residual	44	0.408538787	0.009284972			
14	16	9.825987911	2.778553291		Total	46	13.04547025				
15	17	9.997267523	2.637089157								
16	18	10.0886679	2.560142151								
17	19	10.24134269	2.532028953			Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
18	20	10.3595555	2.573751913		Intercept	5.960340695	0.118540795	50.28092383	1.62439E-40	5.72143742	6.19924397
19	21	10.47599151	2.562449323		time_trend	0.007043661	0.002104069	3.347637544	0.001677496	0.002803188	0.011284134
20	22	10.5333722	2.532028953		ln_cum_prod	-0.335390465	0.015966081	-21.00643697	1.45132E-24	-0.367567986	-0.303212944

Interpretation tips: What does -0.335390465 mean? (It is the coefficient on `ln_cum_prod` in the output above.) Using the weekly data from Levitt et al. (2013) for an undisclosed automobile manufacturer, after controlling for a time trend, as cumulative production experience rises by 10%, the average defects per car are 3.4% lower, on average. This is consistent with learning by doing.

Interpretation tips: Continuing, what does 0.968683476 mean? (It is the R^2 in the output above.) About 97 percent of the variation across weeks in the natural log of average production defects per car is explained by the combination of the natural log of cumulative production experience and a time trend. This is consistent with the claim in Levitt et al. (2013) that the simple model in Equation 2 fits these data well.

5. **Construct a correlation matrix** for the variables: $\ln(\text{average defects per car})$, $\ln(\text{cumulative production})$, and the time trend. Use the data from part 4. **Verify** it matches:

	time_trend	ln_cum_prod	ln_ave_def_pc
time_trend	1		
ln_cum_prod	0.870334115	1	
ln_ave_def_pc	-0.809082222	-0.980156745	1

Interpretation tips: What does -0.809082222 mean? As expected given Figure 1, there is a strong negative correlation between the time trend and natural log of defects per car: defects are dropping over time. This does *not* contradict the positive coefficient on the time trend in the multiple regression in part 4. (The background reading at the start of this module elaborates.)

6. Table 1, Panel A, Column (1) shows a *simple regression*, which means an uncomplicated relationship between the correlation and slope coefficient: they *will* have the *same* sign.

- (a) To illustrate, **standardize** each variable in the regression: $\ln(\text{average defects per car})$ and $\ln(\text{cumulative production})$. Recall that standardizing a variable means transforming it by subtracting its mean and dividing by its standard deviation: $z_X = \frac{X - \bar{X}}{s_X}$.

EXCEL TIPS: Use the functions AVERAGE and STDEV.S. Use the \$ to anchor to the cells with the sample mean and sample s.d. Name the new variables s_*.

	A	B	C	D	E
1	mean:	10.25768227	2.710190716		
2	s.d.:	1.806927037	0.53253843		
4		ln_cum_prod	ln_ave_def_pc	s_ln_cum_prod	s_ln_ave_def_pc
5		4.50201976	4.331182942	-3.185331997	3.043897181
6		5.321868	4.244979139	-2.731606847	2.882023787
7		6.250517041	3.857150787	-2.217668529	2.153760192

- (b) **Run** the simple regression on the standardized data. **Verify** that the slope coefficient in the standardized regression equals the correlation in part 5 (-0.9802).

7. Table 1, Panel A, Column (2) shows a *multiple regression*, which means the correlation and the slope coefficient can differ wildly and can even have opposite signs.

- (a) What if you standardized $\ln(\text{average defects per car})$, $\ln(\text{cumulative production})$, and the time trend? **Write a sentence** predicting the sign (+ or -) of each regression coefficient.
- (b) **Standardize** those three variables. **Run** the multiple regression using them.

	A	B	C	D	E	F	G	H	I	J	K	L
1	mean:	27	10.25768227	2.710190716					SUMMARY OUTPUT			
2	s.d.:	13.7113092	1.806927037	0.53253843								
4		time_trend	ln_cum_prod	ln_ave_def_pc	s_time_trend	s_ln_cum_prod	s_ln_ave_def_pc		Regression Statistics			
5		4	4.50201976	4.331182942	-1.677447402	-3.185331997	3.043897181		Multiple R	0.984217189		
6		5	5.321868	4.244979139	-1.604514906	-2.731606847	2.882023787		R Square	0.968683476		
7		6	6.250517041	3.857150787	-1.531582411	-2.217668529	2.153760192		Adjusted R Square	0.967259997		
8		7	6.78351004	3.838300625	-1.458649915	-1.922696465	2.118363381		Standard Error	0.180941987		
9		8	7.355105578	3.711650466	-1.385717419	-1.60636076	1.880539869		Observations	47		
10		9	7.853910447	3.404483446	-1.312784923	-1.330309287	1.303742022		ANOVA			
11		10	8.217457465	3.279766632	-1.239852428	-1.129112996	1.069548946			df	SS	MS
12		11	8.535646699	3.356516397	-1.166919932	-0.953018872	1.213669558		Regression	2	44.55943988	22.27971994
13		12	8.87553107	3.194240386	-1.093987436	-0.7649181	0.908947866		Residual	44	1.440560121	0.032740003
14		13	9.173504394	2.891812855	-1.02105494	-0.600011981	0.341049825		Total	46	46	
15		14	9.451097264	2.751468509	-0.948122445	-0.446384932	0.077511389					
16		15	9.652888662	2.781246493	-0.875189949	-0.33470837	0.13342845		Coefficients			
17		16	9.825987911	2.778553291	-0.802257453	-0.238910784	0.128371158		Intercept	-2.09074E-15	0.026393101	-7.92152E-14
18		17	9.997267523	2.637089157	-0.729324957	-0.144120232	-0.137270017		s_time_trend	0.181353695	0.054173635	3.347637544
19		18	10.0886679	2.560142151	-0.656392462	-0.093536907	-0.28176101		s_ln_cum_prod	-1.137995052	0.054173635	-21.00643697

Interpretation tips: What does -2.09074E-15 mean? (It is the intercept coefficient in the output above.) Standardizing all variables causes a zero (to machine precision) for the intercept. A regression equation always passes through the mean. In other words, if you plug the mean value of each x into $\hat{y} = b_0 + b_1x_1 + \dots + b_kx_k$ then \hat{y} equals \bar{y} . All of the variables' means are zero after standardization, including y , so b_0 must be zero. Given the limits of machine precision, Excel cannot return an exact value of zero: in this instance, it says -0.00000000000000209074. Pulling this together, the natural log of average defects per car is predicted to be average – zero standard deviations away from the mean – for a week around the middle of the observed period (average value of the time trend) and for an average value of the natural log of cumulative production.

8. Using the data from part 7, **construct a correlation matrix** using the standardized versions of the variables: $\ln(\text{average defects per car})$, $\ln(\text{cumulative production})$, and the time trend. **Verify** the correlations are identical to those found in part 5.
9. **Review** these questions and answers.

- (a) Why are many of the values in the multiple regression output for part 7 (standardized) *identical* to part 4 (not standardized)? All unit-free statistics are not affected by removing the units of measurement with standardization. (Remember $z_X = \frac{X - \bar{X}}{s_X}$ removes the original units of measurement of X : they cancel out because both the numerator and denominator are measured in the original units of X .) Hence, t ratios, the F statistic, P-values, and the measures of R^2 will *not change* with the standardization of the variables. However, statistics that depend on units – including the s_e , SST , SSR , SSE , regression coefficients, and standard errors of regression coefficients – will change.
- (b) Why does standardization change the regression coefficients on the x variables? (We already discussed how and why it changes the intercept.) Again, $z_X = \frac{X - \bar{X}}{s_X}$ removes the original units of measurement and records how many standard deviations above or below the variable's mean a particular value is. Hence, standardization does affect the *magnitude* of the regression coefficients and changes their interpretation (because of the change in units). For example, in $\widehat{s_y} = b_0 + b_1s_{x_1} + \dots + b_k s_{x_k}$, where the s_{\cdot} prefix signifies the variable has been standardized, suppose b_2 is negative. What does b_2 say? A one standard deviation increase in x_2 is associated with a b_2 standard deviation decrease in y on average, after controlling for x_1, x_3, \dots, x_k . However, standardization will *not* change the *sign* of the regression coefficient. (If the relationship was negative, it will still be negative after standardization, which just affects the units of measurement.)

Test/exam examples: Levitt et al. (2013) has appeared.

- Question (1), [April 2022 Test #4](#) (with [solutions](#))
- Question (3), [April 2017 Final Exam](#) (with [solutions](#))

E.2 Module E.2: More on Multiple Regression, Including Inference

Concepts: Reinforce the big idea of multiple regression. Using natural logs. Using dummy variables. Inference with simple and multiple regression. Assessing a multiple regression model overall.

Case studies: We revisit “How Much Energy Do Building Energy Codes Save? Evidence from California Houses,” abbreviated Levinson (2016) and “Fitting Percentage of Body Fat to Simple Body Measurements,” abbreviated Johnson (1996) (“Just Checking” on p. 695 of the textbook).

Required readings: Sections 20.1 - 20.6 and the refresher below. Module D.3 coverage of Levinson (2016), including dummy variables with multiple categories, and the tables and figures below.

- First, a quick refresher on statistical inference with simple and multiple regression:
 - For an *individual regression coefficient*, there are two methods of statistical inference: confidence interval (CI) estimation and hypothesis testing (via P-value or rejection region (aka critical value) approach). We use the Student t distribution. To fix ideas, consider an inference about β_j in $y = \beta_0 + \beta_1 x_1 + \dots + \beta_j x_j + \dots + \beta_k x_k + \varepsilon$.
 - * The CI estimate of the coefficient on x_j is: $b_j \pm t_{\alpha/2} s_{b_j}$ with $\nu = n - k - 1$ where k is the number of x variables and s_{b_j} is the standard error of the slope coefficient on x_j .
 - * In hypothesis testing for a regression coefficient, there are three cases:
 - Two-tailed test: $H_0 : \beta_j = \beta_{j0}$ versus $H_1 : \beta_j \neq \beta_{j0}$
 - Right-tailed test: $H_0 : \beta_j = \beta_{j0}$ versus $H_1 : \beta_j > \beta_{j0}$
 - Left-tailed test: $H_0 : \beta_j = \beta_{j0}$ versus $H_1 : \beta_j < \beta_{j0}$
 - * The most common test – automatically run by Excel – is the classic test of statistical significance of a coefficient: $H_0 : \beta_j = 0$ versus $H_1 : \beta_j \neq 0$, which is two-tailed.
 - * For hypothesis tests, use a t test statistic given by $t = \frac{b_j - \beta_{j0}}{s_{b_j}}$ with $\nu = n - k - 1$.
 - For inferences about the *overall statistical significance of a regression*, use an F test.
 - * Only one case: $H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$ versus $H_1 : \text{Not all the slopes are zero}$.
 - * Use a F test statistic: $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$, $F = \frac{(SST-SSE)/k}{SSE/(n-k-1)}$, $F = \frac{SSR/k}{SSE/(n-k-1)}$, or $F = \frac{MSR}{MSE}$, where $\nu_1 = k$ and $\nu_2 = n - k - 1$.
 - For *simple* regression, $k = 1$ and the F and t tests give the *same* P-values and conclusions because they are the same test: $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. However, for *multiple* regression, $k > 1$ and you must use the F test to assess *overall* statistical significance.
 - * Testing whether a *correlation* is statistically significant is the same as testing if a *simple* regression is statistically significant, which can be done via an F test (or a t test). Recalling that for a simple regression the R^2 is the coefficient of correlation squared, it is convenient to use this test statistic formula: $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$.
- For Levinson (2016), Table 3 and Figures 3 and 4 show key multiple regression results. Table 3 (p. 149) reports FOUR separate multiple regressions. In Column (1) the y-variable is the natural log of annual electricity use (MMBTU). Column (2) is identical except that it has one more x-variable: “Age of central AC.” Column (3) is like (1) except that the y-variable is the natural log of annual natural gas use (MMBTU) *excluding* homes with zero gas use. Column

(4) is like (3) except that it controls for the age of the home heater. Actually, Column (4) also controls for the age of the water heater, but those results are not shown. More on that next.

TABLE 3—ANNUAL ENERGY USE BY CALIFORNIA HOUSEHOLDS IN THE RASS SURVEY

Dependent variable: ln(annual MBTUs)	Electricity		Natural gas	
	Full controls (1)	Age of AC (2)	Full controls (3)	Age of heater (4)
Cooling degree-days (100s)	0.021 (0.003)	0.021 (0.003)	−0.016 (0.006)	−0.016 (0.006)
Heating degree-days (100s)	−0.001 (0.004)	−0.001 (0.004)	0.029 (0.007)	0.028 (0.007)
ln (square feet)	0.266 (0.016)	0.266 (0.016)	0.377 (0.021)	0.387 (0.025)
Bedrooms	0.025 (0.007)	0.025 (0.007)	0.016 (0.008)	0.009 (0.007)
Electric cooking	0.036 (0.007)	0.036 (0.007)	−0.037 (0.007)	−0.032 (0.008)
Remodeled	0.027 (0.010)	0.026 (0.010)	−0.010 (0.013)	−0.011 (0.013)
ln (years at address)	0.026 (0.005)	0.027 (0.005)	0.011 (0.004)	0.010 (0.004)
ln (number of residents)	0.243 (0.009)	0.243 (0.010)	0.131 (0.013)	0.138 (0.013)
ln (household income)	0.095 (0.008)	0.095 (0.008)	0.064 (0.008)	0.058 (0.007)
Household head graduated college	−0.051 (0.007)	−0.051 (0.007)	−0.034 (0.011)	−0.038 (0.012)
Disabled resident	0.143 (0.009)	0.142 (0.009)	0.080 (0.017)	0.083 (0.016)
Own home	−0.055 (0.017)	−0.055 (0.017)	−0.078 (0.025)	−0.087 (0.024)
Refrigerators	0.191 (0.007)	0.191 (0.007)		
Room AC	0.058 (0.011)	0.058 (0.011)		
Central AC	0.141 (0.029)	0.153 (0.031)		
Central AC × sq. feet (1,000s)	0.030 (0.009)	0.030 (0.009)		
Age of central AC		−0.0012 (0.0005)		
Home heater age				0.0017 (0.0005)
Year home built	Coefficients displayed in Figures 3 and 4			
Observations	14,045	14,045	12,358	11,644
R^2	0.437	0.437	0.266	0.267

Notes: RASS 2003 and 2009, standard errors clustered by county. Regressions also include 13 climate zone indicators, kids, seniors, black, Latino, and an indicator for the 2009 survey year. Excluded construction category is homes “Built pre-1940.” Full set of coefficients in online Appendix Tables A2 and A3.

Figure of Table 3: Levinson (2016), p. 2879.

- Table 3 does not report all coefficients. What is k (number of x-variables) for Column (1)? Review the table (and *Notes* below it) and come up with a value of k . Column (1) reports coefficients for 16 x-variables: “Cooling degree-days (100s)” to “Central AC × sq. feet (1,000s).”

Add the 11 dummies for “Year home built” (Figure 3 displays those coefficients). (Recall from Module D.3 the 12 unique vintages from pre 1940 through 2005-2008.) The *Notes* say “13 climate zone indicators,” but checking the data there are only 12 unique zones (no home in zone 6), so that means 11 dummies (one is omitted). “Kids, seniors, black, Latino, and an indicator for the 2009 survey year” mean 5 more variables. Hence, $k = 43 (= 16 + 11 + 11 + 5)$.

- To clarify $k = 43$ for Column (1) of Table 3, see the Stata output (below) using the data [calif_energy_regressions.xlsx](#). Some variables had a natural log transformation (`ln_*`).

Linear regression		Number of obs		=	14,045	
		F(41, 51)		=	.	
		Prob > F		=	.	
		R-squared		=	0.4370	
		Root MSE		=	.3844	
		(Std. Err. adjusted for 52 clusters in county_id)				
ln_elec_mmbtu	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
cool_deg_days	.0208966	.0032435	6.44	0.000	.014385	.0274083
heat_deg_days	-.0006305	.0037184	-0.17	0.866	-.0080956	.0068346
ln_sq_feet	.2658211	.0161376	16.47	0.000	.2334234	.2982188
bedrooms	.0251613	.0067003	3.76	0.000	.0117099	.0386126
elec_cook	.0360624	.0067736	5.32	0.000	.0224639	.0496661
remodeled	.027179	.0101623	2.67	0.010	.0067774	.0475806
ln_yrs_at_address	.0264959	.0046375	5.71	0.000	.0171858	.035806
ln_num_res	.2434402	.009495	25.64	0.000	.2243781	.2625023
ln_hhinc1000	.0950838	.0083527	11.38	0.000	.078315	.1118526
college	-.0509267	.0072453	-7.03	0.000	-.0654722	-.0363812
anydisabled	.1430721	.0086389	16.56	0.000	.1257288	.1604154
own	-.054588	.0170567	-3.20	0.002	-.0888309	-.0203452
refrigerators	.19082	.0072572	26.29	0.000	.1762505	.2053895
room_ac	.0582558	.0113664	5.13	0.000	.0354369	.0810748
central_ac	.1410409	.0291489	4.84	0.000	.082522	.1995598
central_acXsq_feet	.0296008	.0088588	3.34	0.002	.0118161	.0473856
num_res_0_5	-.0494288	.00498	-9.93	0.000	-.0594265	-.039431
num_res_6_5_up	-.0306708	.0065793	-4.66	0.000	-.0438792	-.0174624
black	.0490492	.0206276	2.38	0.021	.0076376	.0904607
latino	-.0848364	.0129473	-6.55	0.000	-.1108291	-.0588436
yr_2009	.0520844	.0081459	6.39	0.000	.0357309	.0684379
constr_40_49	.0120263	.0166578	0.72	0.474	-.0214156	.0454683
constr_50_59	.0331037	.026129	1.27	0.211	-.0193525	.0855599
constr_60_69	.0624592	.0221947	2.81	0.007	.0179014	.1070169
constr_70_74	.0566884	.025483	2.22	0.031	.0055291	.1078476
constr_75_77	.0700479	.0268705	2.61	0.012	.0161032	.1239926
constr_78_82	.0646985	.0274141	2.36	0.022	.0096624	.1197347
constr_83_92	.038191	.0238407	1.60	0.115	-.0096713	.0860533
constr_93_97	-.0048374	.0255182	-0.19	0.850	-.0560672	.0463925
constr_98_00	-.0162601	.0216121	-0.75	0.455	-.0596483	.0271281
constr_01_04	-.0254243	.0295766	-0.86	0.394	-.0848019	.0339532
constr_05_08	-.0871186	.0284832	-3.06	0.004	-.1443009	-.0299362
climate_zone_2	-.007353	.0269794	-0.27	0.786	-.0615165	.0468105
climate_zone_3	.0521531	.0233689	2.23	0.030	.0052381	.0990681
climate_zone_4	-.0710166	.0286263	-2.48	0.016	-.1284864	-.0135468
climate_zone_5	-.1202979	.0362195	-3.32	0.002	-.1930117	-.0475841
climate_zone_7	.018367	.0463466	0.40	0.694	-.0746777	.1114116
climate_zone_8	-.1552594	.0447684	-3.47	0.001	-.2451358	-.0653831
climate_zone_9	-.1426846	.0359443	-3.97	0.000	-.2148458	-.0705235
climate_zone_10	-.0935142	.028398	-3.29	0.002	-.1505256	-.0365028
climate_zone_11	-.2141044	.0484761	-4.42	0.000	-.3114241	-.1167846
climate_zone_12	-.0591326	.0400839	-1.48	0.146	-.1396045	.0213392
climate_zone_13	-.1562692	.0475679	-3.29	0.002	-.2517657	-.0607726
_cons	1.816048	.1130848	16.06	0.000	1.58902	2.043075

The notes for Table 3 say “standard errors clustered by county” and the Stata output says “robust” standard errors. We use the “regular” standard error formulas. However, real research usually uses robust standard errors. For example, there are alternative formulas for computing standard errors robust to heteroskedasticity that are correct even with a violation of the equal variance assumption (pp. 182, 606, and 697 of the textbook). Regardless of the formula to compute standard errors, the coefficient estimates are the same.

- “Full controls” in Figures 3 and 4 show the 95% CI estimates for the vintage dummy coefficients for Columns (1) and (3), respectively, of Table 3. Together, we will replicate the point estimates (center dot) for “No controls” in Figure 3. These are the coefficients on the vintage dummy variables. We cannot exactly replicate the intervals using $b_j \pm t_{\alpha/2} s_{b_j}$ because Excel lacks robust s.e.’s (s_{b_j}). However, robust and “regular” s.e.’s are often similar (including in this case).

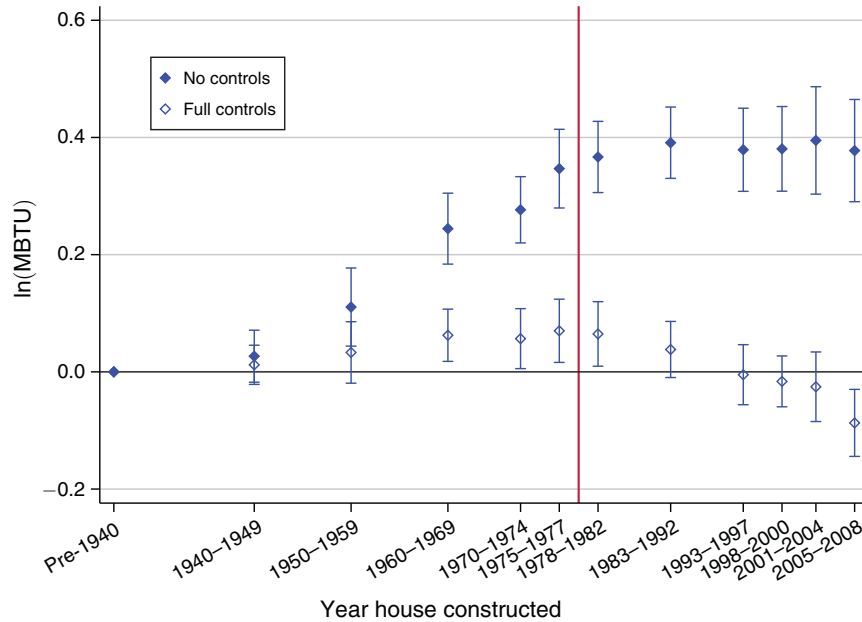


Figure 3: Residential ELECTRICITY Use in California, Controlling for Characteristics

Note: RASS 2003 and 2009, single-family detached California homes without electric heat or hot water.

Source: Levinson (2016), p. 2880 (including typo: ln(MBTU) should say ln(MMBTU)).

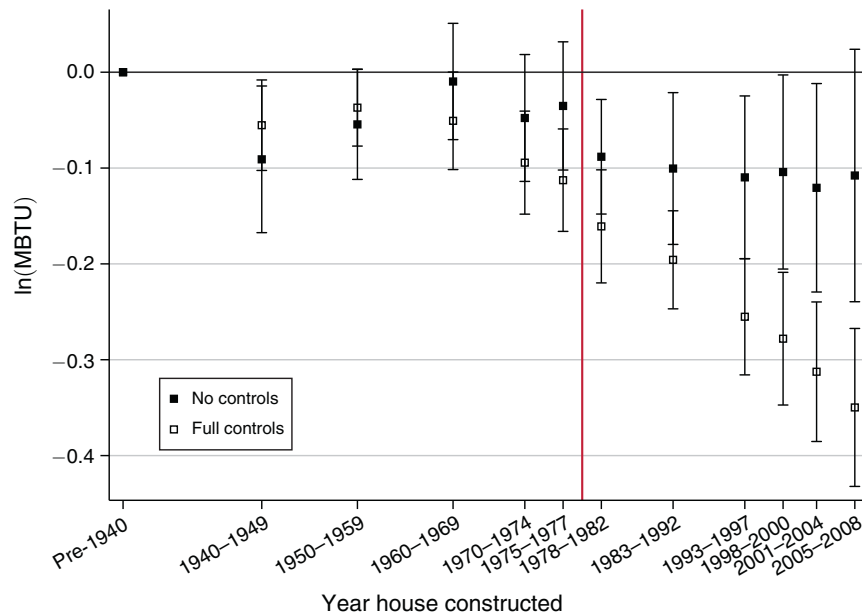


Figure 4: Residential NATURAL GAS Use in California, Controlling for Characteristics

Note: RASS 2003 and 2009, single-family detached California homes without electric heat or hot water.

Source: Levinson (2016), p. 2881 (including typo: ln(MBTU) should say ln(MMBTU)).

- From the textbook, carefully review the “Just Checking” box on p. 695, reproduced below.

Just Checking

Body fat percentage is an important health indicator, but it's difficult to measure accurately. One way to do so is to take a magnetic resonance image (MRI), but this is expensive. Insurance companies want to know if body fat percentage can be estimated from easier-to-measure characteristics such as *Height* and *Weight*. A scatterplot of *Percent Body Fat* against *Height* shows no pattern, and the correlation is -0.03 and is not statistically significant. A multiple regression using *Height* (centimetres), *Age* (years), and *Weight* (kilograms) finds the following model:

$s = 5.382$ on 246 degrees of freedom
Multiple R-squared: 0.584,
F-statistic: 115.1 on 3 and 246 DF, P-value: <0.0001

	Coeff	SE(Coeff)	t-ratio	P-value
Intercept	57.27217	10.39897	5.507	<0.0001
Height	-0.50164	0.05909	-8.064	<0.0001
Weight	0.55805	0.03263	17.110	<0.0001
Age	0.13732	0.02806	4.895	<0.0001

Datasets: For Levinson (2016): [calif.energy_regressions.xlsx](#). For Johnson (1996): [pct_body_fat.xlsx](#).

Interactive module materials for Module E.2:

1. Consider Levinson (2016) and use [calif.energy_regressions.xlsx](#). **Review** the multiple regression coefficient estimates in Column (1) of Table 3 on page 149. We *cannot* replicate these results because Excel: (1) allows at most 16 explanatory variables and (2) does not include an option for robust standard errors (in parentheses). But, there is still *a lot* that we *can do* and learn.
 - (a) **Review** the “No controls” results in Figure 3 on page 151. Notice the natural log transformation of the y-variable (in both Table 3 and Figure 3). Also, despite the name, the “No controls” case does control for the year of the RASS survey. (Compared to Column (1) in Table 3, which controls for *many* other factors, this has “no controls.”) Noting that the researcher set pre 1940 as the reference category, **replicate** the center point of the intervals for “No controls” in Figure 3. **Verify** that your output matches the below.
EXCEL TIPS: Remember to create the y-variable using the LN function. Also, remember to arrange the x-variables (see output below for the variable names) to be adjacent to each other in the spreadsheet to use the Regression tool in Data Analysis.

Regression Statistics	
Multiple R	0.293323517
R Square	0.086038686
Adjusted R Square	0.085257077
Standard Error	0.489231658
Observations	14045

ANOVA					
	df	SS	MS	F	Significance F
Regression	12	316.1656134	26.34713445	110.0789512	4.9946E-263
Residual	14032	3358.525736	0.239347615		
Total	14044	3674.69135			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2.80091723	0.014998686	186.7441713	0	2.77151781	2.830316651
yr_2009	0.099643802	0.008411839	11.84566212	3.22698E-32	0.083155478	0.116132125
constr_40_49	0.026738028	0.021701186	1.232099856	0.217932431	-0.015799184	0.06927524
constr_50_59	0.110574793	0.01757221	6.292594734	3.21478E-10	0.076130924	0.145018662
constr_60_69	0.244376666	0.017853879	13.68759493	2.25322E-42	0.209380687	0.279372644
constr_70_74	0.276537056	0.020694214	13.36301306	1.75802E-40	0.235973642	0.317100469
constr_75_77	0.346758564	0.023614619	14.68406364	1.8657E-48	0.300470769	0.393046359
constr_78_82	0.366688048	0.021044291	17.42458582	2.73986E-67	0.325438439	0.407937658
constr_83_92	0.391032426	0.017953747	21.77998972	1.8423E-103	0.355840693	0.426224159
constr_93_97	0.378969248	0.022834935	16.59602942	2.85445E-61	0.334209738	0.423728758
constr_98_00	0.380513957	0.024034536	15.83196574	5.71125E-56	0.333403067	0.427624846
constr_01_04	0.394942885	0.025326069	15.59432225	2.27406E-54	0.345300419	0.44458535
constr_05_08	0.377592192	0.032209664	11.72294738	1.36488E-31	0.314456965	0.440727418

Interpretation tips: What do the *extremely* tiny P-values mean? For example, in the test of statistical significance for the dummy for homes constructed from 1983 to 1992 – $H_0 : \beta_{83.92} = 0$ versus $H_1 : \beta_{83.92} \neq 0$ – the P-value is 0.00...18423 where there are 102 zeros after the decimal point! (This is beyond machine precision: you can think of this as zero.) We have *overwhelming* evidence that mean electricity use differs for homes constructed from 1983 to 1992 compared to homes constructed before 1940, after controlling for the year of the RASS survey. Remember, the coefficients on the vintage dummies say how each vintage compares with the reference category (homes constructed before 1940). The only non-tiny P-value is for homes constructed in the 1940s. We have insufficient evidence to conclude that mean energy use differs between homes constructed in the 1940s compared to those constructed earlier, after controlling for the year of the RASS survey.

Interpretation tips: What does 0.377592192 mean? (It is the point estimate of the coefficient for homes constructed from 2005 to 2008.) Compared with homes built before 1940, homes built from 2005 to 2008 use approximately 37.8 percent more electricity on average, after controlling for the year of the RASS survey. (Remember the y-variable has been logged.)

Interpretation tips: What do the values under *Lower 95%* and *Upper 95%* mean? As already discussed in the previous interpretation tips, the coefficients on the vintage dummies compare each vintage with the reference category (homes constructed before 1940). For example, we are 95% confident that compared with homes constructed before 1940, homes constructed from 2005 to 2008 use between 31.4 and 44.1 percent more electricity on average, after controlling for the year of the RASS survey. Looking again at Figure 3 on page 151, we can see that this interval matches up very well, even though our CI estimate uses regular standard errors, not robust standard errors. (The center of the interval matches perfectly.)

- (b) Figure 3 on page 151 shows 95% confidence intervals. What if we wanted 99% intervals? **Recall** that the formula for a CI estimate of a regression coefficient is $b_j \pm t_{\alpha/2} s_{b_j}$. The Stata output on page 150 gives the exact regression results for the “Full controls” intervals in Figure 3. For houses constructed in 2005-2008, **compute** the 99% confidence interval. For convenience (so you don’t have to flip back-and-forth), the coefficient on `constr_05_08` is -0.0871186 with a robust standard error of 0.0284832 and degrees of freedom $\nu = 51$.¹⁴ Hence, you need to find the correct value of $t_{\alpha/2}$ and then plug that and the values above (from Stata) into the CI estimate formula. **Verify** that you obtain [-0.16333, -0.01091].
- EXCEL TIPS:** Recall the function `T.INV` that returns the value $t_{\alpha/2}$ given the cumulative area *to the left* of that value and the degrees of freedom.
- Interpretation tips:** What does [-0.16333, -0.01091] mean? We are 99% confident that once we control for differences in climate across California, the size of the house, the home’s appliances (including air conditioning), key characteristics of the home’s residents (including number of residents), and year of the RASS survey, homes built from 2005-2008 on average use between 1.1 and 16.3 percent *less* electricity than homes built before 1940, which is consistent with newer building codes leading to more energy efficient homes. These estimates stand in sharp contrast to the unfair comparison in part 1a, where a simple comparison that failed to control for key lurking/unobserved/confounding/omitted variables estimated that homes built from 2005-2008 use 37.8 percent *more* electricity. As shown in Figure 3, it makes a big difference whether we control for these other factors.
- (c) Page 122 explains why it is *not fair* to simply compare energy use across homes built in different time periods (under different building codes): “houses built more recently are larger, have more occupants, and are in less temperate parts of the state” (p. 2868).
- To make the comparison in part 1a *more fair*, **pick** FOUR variables from those in Column (1) of Table 3 (p. 149). Pick those that you think may best improve fairness. (Why four? We already have 12 x-variables and Excel can handle at most 16.)
 - Run** a regression that adds the four variables you picked in part 1(c)i to your regression from part 1a. (Remember to apply the natural log as needed.)
 - Compare** your results in part 1(c)ii (“some controls”) with those in part 1a (“no controls”). If you picked four important lurking variables to control for, the coefficients on the vintage dummies should be very different: turning what appears to be increasing energy usage by homes built under more recent building codes into fairly flat to decreasing energy usage (as in Figure 3).
 - Compare** your results in part 1(c)ii (“some controls”) with the sample guess (“some controls”) shown below.

Interpretation tips: Overall, how do the key results with “some controls” (next page) compare with the key results with “full controls” (shown in Figure 3 and the Stata output on page 150)? The key results – the coefficients on the vintage dummies, which capture the changing building codes over time – are remarkably similar. Our simple Excel approach – just controlling for a handful of the most important lurking variables and using plain-vanilla standard errors – still leads to the same overall answer to the original research question. The impact of building codes on energy efficiency

¹⁴Using this type of robust standard errors affects the degrees of freedom. In this case, it is *not* the usual $\nu = n - k - 1$.

seems disappointing. Levinson (2016) explains on pp. 2879-80: “Only the very newest houses built after 2005 have coefficients statistically significantly lower than houses built before 1978. And as we have seen [in other parts of this paper and in other published research], very new homes use less electricity for reasons likely unrelated to building codes: new appliances, well-sealed windows, etc.”

Sample educated guess to select four important variables to control for:

Regression Statistics	
Multiple R	0.590986523
R Square	0.34926507
Adjusted R Square	0.348522858
Standard Error	0.412871224
Observations	14045

ANOVA					
	df	SS	MS	F	Significance F
Regression	16	1283.441333	80.21508331	470.5727886	0
Residual	14028	2391.250017	0.170462647		
Total	14044	3674.69135			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	2.299491392	0.014919424	154.1273549	0	2.270247335	2.32873545
cool_deg_days	0.023994715	0.001136583	21.11128134	2.03848E-97	0.021766861	0.026222568
ln_sq_feet	0.43683163	0.009618505	45.41575184	0	0.41797808	0.45568518
ln_num_res	0.256371747	0.006796996	37.71838984	1.2487E-296	0.24304873	0.269694764
central_ac	0.211006685	0.008275383	25.4981183	3.1576E-140	0.194785834	0.227227537
yr_2009	0.051487662	0.007339257	7.015378481	2.39835E-12	0.037101743	0.065873582
constr_40_49	0.029368085	0.018332682	1.601952457	0.109188639	-0.006566412	0.065302581
constr_50_59	0.037917498	0.014877628	2.548625141	0.010825406	0.008755366	0.06707963
constr_60_69	0.078540997	0.015253059	5.149196341	2.65146E-07	0.04864297	0.108439023
constr_70_74	0.077211207	0.017697734	4.362773726	1.29341E-05	0.042521293	0.11190112
constr_75_77	0.093859498	0.020283551	4.627370234	3.73654E-06	0.054101039	0.133617957
constr_78_82	0.088566174	0.018226966	4.859073978	1.19211E-06	0.052838896	0.124293453
constr_83_92	0.059345294	0.015944254	3.722048946	0.000198385	0.028092434	0.090598154
constr_93_97	0.000500876	0.020048108	0.02498371	0.980068313	-0.038796084	0.039797836
constr_98_00	-0.030807821	0.021152757	-1.456444687	0.145292111	-0.072270041	0.010654399
constr_01_04	-0.052840257	0.022329139	-2.366426071	0.017974268	-0.096608342	-0.009072172
constr_05_08	-0.131700471	0.028135997	-4.680853193	2.88357E-06	-0.18685077	-0.076550172

Note: If we left square feet measured in feet (as in the original data) before taking the natural log, the output would be *identical* to the above (which defines \ln_sq_feet as the natural log of square feet measured in 1,000s) except for the row of results for the intercept. (The intercept would be -0.718034608, with a standard error of 0.071087154.) It would make absolutely no difference for the row of results for square feet (or any of the other results). This is part of the appeal of logarithms: we do not have to worry about the units of measurement for variables that have been logged.

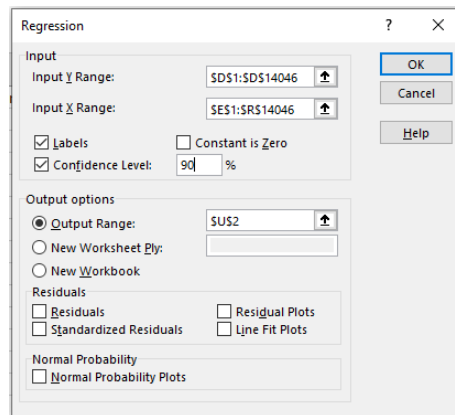
- (d) Suppose a California real estate agent claims that for every 10 percent increase in the size of a home, on average electricity use is at least 5 percent higher.
- Run** an appropriate regression to check if the RASS data support this claim. (Hint: Estimate a log-log *simple* regression model: $\ln(elec)_i = \alpha + \beta * \ln(size)_i + \varepsilon_i$.)
 - Conduct** an appropriate hypothesis test and assess the strength of the evidence to support the claim. (Hint: Test $H_0 : \beta = 0.5$ versus $H_1 : \beta > 0.5$ and use a P-value approach.) Verify that you get a t test statistic of 4.3332886 ($= \frac{0.5419978 - 0.5}{0.0096919}$), degrees of freedom of 14,043 ($\nu = 14,045 - 2$), and a P-value of 7.4E-06 (i.e. 0.0000074).

Interpretation tips: What does 0.0000074 mean? This tiny P-value means that the RASS data overwhelmingly support the claim that for every 10 percent increase

in the size of a home, on average electricity use is at least 5 percent higher. (It is technically correct to say: for every 1 percent increase in the size of a home, on average electricity use is at least 0.5 percent higher. However, an interpretation should pick a change that people can relate to: 1 percent seems too small to visualize how the home would appear bigger (it is barely an extra broom cupboard).)

- (e) **Run** a multiple regression where the y variable is the natural log of annual household electricity usage and the fourteen x variables are: the natural log of house size (in square feet), the natural log of the number of residents, a dummy indicating if the observation is from the 2009 RASS data, and a full set of dummies for the climate zones (with any one serving as the omitted category). For the slope coefficient for the natural log of house size, what is the 90% confidence interval estimate?

EXCEL TIPS: The easiest way to include an alternative to the automatically reported 95% confidence interval estimates of the OLS coefficient estimates is by typing the desired confidence level in the “Confidence Level:” box in the Regression pop-up window in the Analysis ToolPak. The screenshot below shows entering the requested 90% confidence level.



Verify that the first FOUR rows of the output for the coefficients match the Excel regression output below. This includes the requested 90% confidence interval estimate for the coefficient for the natural log of house size, which is from the lower confidence limit of 0.483326869 to the upper confidence limit of 0.513295679.

Interpretation tips: What does 0.483326869 to 0.513295679 mean? We are 90% confident that once we control for differences across climate zones in California, the number of residents, and year of the RASS survey, homes that are 10 percent larger on average use between 4.8 percent and 5.1 percent *more* electricity. Note that this does *not* control for the home vintage and we should definitely stick to a descriptive, not causal, interpretation. You will notice that with a more full set of controls, as in Table 3 on page 149, that the coefficient estimate on the size of the home is quite a bit smaller.

Note that whether your remaining rows of coefficients – i.e. the climate zone coefficients – match the Excel regression output below depends on which climate zone you specified as the omitted category. The output below (arbitrarily) sets Climate Zone 13 to be the reference category. If you happened to make that same choice, then all rows will perfectly match. However, the coefficients on the variables `ln_sq_feet`, `ln_num_res`, and `yr_2009` will

be exactly the same regardless of which climate zone you pick to serve as the omitted category. Try re-running your regression picking a different reference category for the climate zone (you will still have 14 x variables) and verify that this is true.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.552628824
R Square	0.305398617
Adjusted R Square	0.304705501
Standard Error	0.42652984
Observations	14045

ANOVA					
	df	SS	MS	F	Significance F
Regression	14	1122.245658	80.16040411	440.6168065	0
Residual	14030	2552.445692	0.181927704		
Total	14044	3674.69135			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 90.0%	Upper 90.0%
Intercept	-0.997727648	0.068341367	-14.5991761	6.38272E-48	-1.131685822	-0.863769474	-1.110146616	-0.88530868
ln_sq_feet	0.498311274	0.00910927	54.70375669	0	0.480455893	0.516166655	0.483326869	0.513295679
ln_num_res	0.254707046	0.007014679	36.31057702	5.3261E-276	0.240957341	0.268456751	0.243168164	0.266245928
yr_2009	0.098781124	0.007339822	13.45824509	4.94789E-41	0.084394096	0.113168152	0.086707394	0.110854854
climate_zone_1	0.119751722	0.023199156	5.16189994	2.47796E-07	0.074278289	0.165225155	0.081589986	0.157913457
climate_zone_2	0.245044905	0.019806209	12.37212553	5.62919E-35	0.2062221	0.283867711	0.212464439	0.277625371
climate_zone_3	0.351070901	0.016322382	21.50855768	5.4937E-101	0.31907686	0.383064943	0.324221199	0.377920604
climate_zone_4	0.094813333	0.012748135	7.437427651	1.08614E-13	0.069825292	0.119801375	0.073843132	0.115783534
climate_zone_5	-0.075068516	0.014406819	-5.210624032	1.90874E-07	-0.103307798	-0.046829234	-0.098767188	-0.051369843
climate_zone_7	0.254883067	0.021591752	11.80464965	5.234E-32	0.21256036	0.297205774	0.21936545	0.290400684
climate_zone_8	0.042355975	0.01301171	3.255219605	0.001135733	0.016851291	0.067860658	0.020952202	0.063759747
climate_zone_9	0.098571494	0.013690267	7.200114803	6.32212E-13	0.071736749	0.12540624	0.076051522	0.121091467
climate_zone_10	0.208906422	0.013773337	15.16745184	1.48851E-51	0.181908849	0.235903995	0.186249803	0.231563041
climate_zone_11	-0.146143622	0.055667912	-2.625275778	0.008667272	-0.255260138	-0.037027105	-0.237715235	-0.054572008
climate_zone_12	0.197976786	0.087645086	2.25884639	0.023908227	0.026180753	0.369772819	0.053803929	0.342149643

Test/exam examples: For examples of Levinson (2016), see page 130.

2. Consider Johnson (1996) and use [pct.body.fat.xlsx](#):

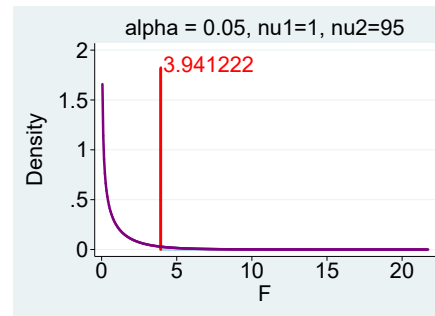
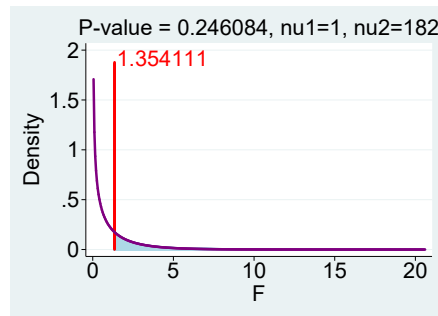
- (a) **Compute** the coefficient of correlation between percent body fat and height. Recalling that for a simple regression, the R^2 is equal to the coefficient of correlation squared, **compute** the R^2 and then use it to compute the F statistic recalling that $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$.

EXCEL TIPS: Use parentheses to preserve the proper order of operations.

F.DIST		X	✓	f _x	=(B3/1)/((1-B3)/(B1-1-1))
	A	B	C	D	
1	n=	250.0000000			
2	r=	-0.02938959			
3	R-squared=	0.000863748003681			
4	F=	=(B3/1)/((1-B3)/(B1-1-1))			

- (b) **Verify** the value of the F statistic by **running a simple regression** where the y-variable is percent body fat and the x-variable is height. **Identify** the P-value to test if there is a statistically significant correlation between percent body fat and height. **Note** that the P-value for the test of the coefficient is identical to the P-value for the overall test statistical significance because this is a simple regression.

- **Explain** how percent body fat and height can be unrelated while the multiple regression results show a large and statistically significant negative coefficient on height. (To check your answer, review the “Just Checking” questions and answers.)
- (c) For another sample of 184 males, the correlation between body fat and height is 0.08593732. **Assess** whether this correlation is statistically significant. If so, at which significance levels? **Verify** that you obtain an F test statistic of 1.354111 and a P-value of 0.246084. **EXCEL TIPS:** The F.DIST function returns the cumulative area to the left, given the F test statistic and the numerator and denominator degrees of freedom. Set the logical value to TRUE (for the area, not the height of the F density function). Remember the P-value is the area *to the right*. For example, `=1-F.DIST(5.84893192,2,20,TRUE)` returns 0.01, which matches the F table in the aid sheets for $\alpha = 0.01$, $\nu_1 = 2$ and $\nu_2 = 20$.



Note: The figure on the left visually illustrates the answer to part 2c. The figure on the right visually illustrates the answer to part 2d.

Interpretation tips: What does 0.246084 mean? The P-value of 0.25 is higher than any conventional significance level, including $\alpha = 0.10$, which means we have insufficient evidence to support the research hypothesis that the correlation differs from zero. In other words, we fail to rule out the null hypothesis, which says that height and body fat are completely unrelated. Remember that we cannot prove the null (we can only prove the research hypothesis, which occurs when we reject the null). However, common sense suggests that a male’s height and percent body would be unrelated: you can be tall and skinny, short and skinny, tall and fat, or short and fat. There is no good reason to expect that male height would be related to percent body fat so this big P-value is not surprising.

- (d) For another sample of 97 males, you wish to assess the correlation between body fat and height. **Compute** the critical value (i.e. edge of the rejection region) for the F test to assess whether this correlation is statistically significant at $\alpha = 0.05$. **Verify** that you obtain a critical value of 3.941222.

EXCEL TIPS: The F.INV function returns the critical value given a cumulative area to the left and the degrees of freedom. For example, `=F.INV(0.99,2,20)` returns 5.84893192, which matches the F table in the Aid Sheets for $\alpha = 0.01$, $\nu_1 = 2$ and $\nu_2 = 20$.

Interpretation tips: What does 3.941222 mean? This critical value means the computed F test statistic from the sample must be at least 3.94 for the correlation to be statistically significant at a 5% significance level. Given that the key ingredient in the F test statistic formula is the correlation (which squared is the R^2), this means our sample correlation between height and body fat must be sufficiently strong (far from zero) to

yield a big enough F test statistic to support the conclusion that height and body fat are related. Of course, as discussed above, we are unlikely to meet this burden of proof, because there is no reason to think that height and percent body fat are related, so the correlation will be near zero and the F test statistic is unlikely to be at least 3.94.

- (e) ***Replicate*** the results in the “Just Checking” box on page 152. To convert height to centimeters use 1 inch = 2.54 cm. To convert weight to kilograms use 1 pound = 0.45359237 kg. Also, the coefficient on weight is 0.5592256, which doesn’t perfectly match 0.55805 because the textbook did not precisely convert pounds to kilograms.

Test/exam examples: Johnson (1996) has appeared on a term test.

- Questions (29) - (34), [April 2016 Test #5](#) (with [solutions](#))

E.3 Module E.3: Interaction Terms & Quadratic Terms

Concepts: Dummies and interactions when relationships differ across groups. Quadratic terms (and higher order polynomials) for non-monotonic nonlinearities.

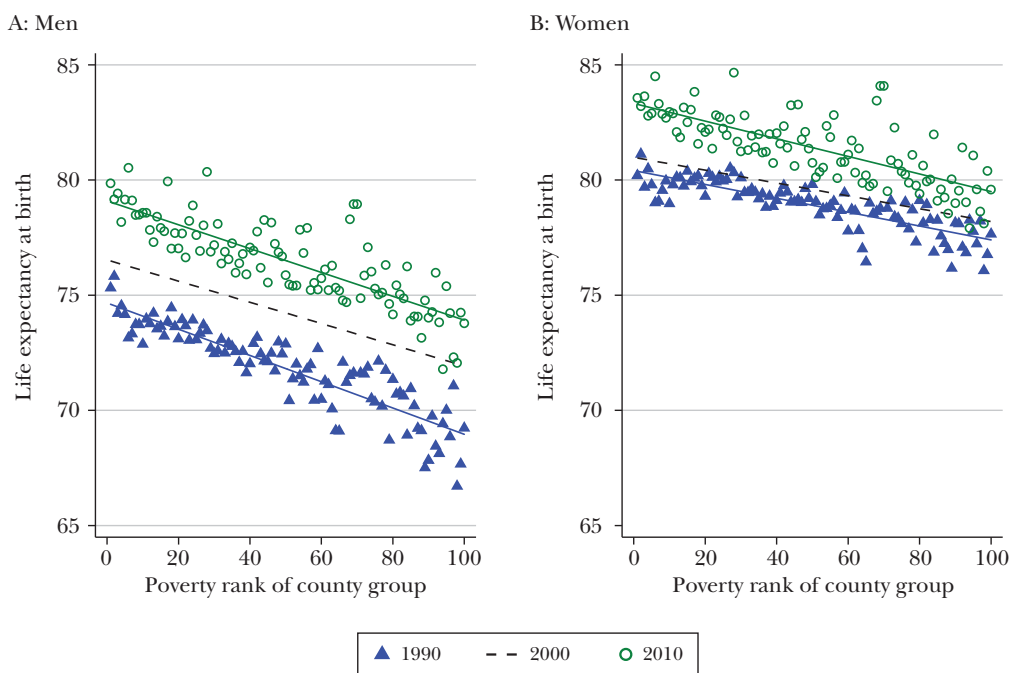
Case studies: We continue with Currie and Schwandt (2016). We study women’s downhill Olympic skiing from the textbook’s (online) Section 21.6 “Quadratic Terms.” It considers the 2002 Salt Lake City Winter Olympics. Data from five Olympic games are in DACM (2002, 2010, 2014, 2018, and 2022). The module case is the 2010 Vancouver Winter Olympics.

Required readings: Sections 21.1 - 21.2, 21.6 (online “Quadratic Terms”) of the textbook.

- Also, recall *everything* about Currie and Schwandt (2016) in Modules B.1 and D.1. Review Figure 2, including the note and the units of measurement of the x and y variables, and Table A2 on page 161 (closely linked to Figure 2). Note the substantial negative slopes: both men and woman living in the poorest counties (i.e. a high poverty rank) have substantially shorter life expectancies compared to those living in the richest counties (i.e. a low poverty rank). Life expectancies are increasing for everyone over time: the green lines (2010) are everywhere higher than the blue lines (1990). However, there is still a steep slope: inequality in life expectancy between those living in rich versus poor counties has not gone away.

Figure 2

Life Expectancy at Birth across Poverty Percentiles



Source: Authors using data from the Vital Statistics, the US Census, and the American Community Survey.
Note: Counties are ranked by their poverty rate in 1990, 2000, and 2010, and divided into groups each representing about 1 percent of the overall population. Each marker represents the life expectancy at birth in a given county group. Lines are fitted using OLS regression. For 2000, markers are omitted and only the regression line is shown. Table A2 provides magnitudes for individual life expectancy estimates and for the slopes of the fitted lines.

Figure 2: Currie and Schwandt (2016), p. 39.

- Table A2, which goes with Figure 2, reports slopes and tests if the 1990 and 2010 slopes differ.

Table A2: Life expectancy for selected county groups and slope of regression lines, 1990 vs. 2010

Life expectancy at birth across gender, years, and county groups								
	Males				Females			
	1990		2010		1990		2010	
	value	std. err.	value	std. err.	value	std. err.	value	std. err.
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<u>(a) LE at birth by poverty ranking of county group</u>								
1	75.32	0.13	79.86	0.11	80.20	0.13	83.57	0.10
25	73.07	0.15	77.60	0.12	79.96	0.14	82.23	0.11
50	72.88	0.17	75.87	0.12	79.81	0.16	80.74	0.12
75	70.36	0.19	75.29	0.14	78.11	0.18	80.37	0.13
100	69.23	0.16	73.78	0.13	77.65	0.14	79.59	0.12
<u>(b) Slope of fitted regression line</u>								
	Slope 1990		Slope 2010		Slope 1990		Slope 2010	
	-0.0570		-0.0518		-0.0301		-0.0383	
<u>(c) p-value of test Slope1990=Slope2010</u>								
	0.2749				0.0445			

Notes: Panel (a) shows life expectancy along with standard errors for the counties in the 1st, 25th, 50th, 75th and 100th poverty percentile, as plotted in Fig. 2. Panel (b) reports the slopes of the fitted regression lines plotted in Figure 2. Panel (c) reports the p-value of the difference between the two slopes.

Figure of Table A2: Currie and Schwandt (2016), p. 14 of appendix. Panel (b) of this table gives the slopes of the four lines in Figure 2 on page 160. Panel (c) tests if those slopes differ between 1990 and 2010.

- Note that Currie and Schwandt (2016) interpret the key figures and tables in sections “Life Expectancy at Birth” on pp. 38-40 and “Age-specific Mortality” on pp. 40-41.

Datasets: For Vancouver: [skiing_2010.xlsx](#). For Currie and Schwandt (2016), the suffix says the figure/table that the data replicate: [mort_in_figure_2_table_a2.xlsx](#)¹⁵ and [mort_in_figure_3_table_a3.xlsx](#).

Interactive module materials for Module E.3:

1. Consider Currie and Schwandt (2016) and use [mort_in_figure_2_table_a2.xlsx](#).

- (a) **Run** the simple regression shown in Panel A (males) in Figure 2 for the year **1990**. **Verify** your slope matches what Panel (b) of Table A2 reports for males in 1990.

EXCEL TIPS: Recall the Filter tool Excel tip given in Module B.1, part 2a on page 43).

- (b) **Run** the simple regression shown in Panel A (males) in Figure 2 for the year **2010**. **Verify** your slope matches what Panel (b) of Table A2 reports for males in 2010.

¹⁵These data have some slight anomalies: there are only 99 observations for 1990 (quantile 70 is missing) and there are only 99 observations for 2000 (quantile 87 is missing). These (and other) issues are present in the original replication files (i.e. are not an error in producing the data for you to use). Hence, we will work with the data as is.

(c) Next, run *one multiple regression* that captures the simple regressions in parts 1a and 1b.

- i. First, **create** the necessary dummy variable, either for the year 1990 or 2010. It does not matter which year you make the reference (omitted) category.

EXCEL TIPS: Use the IF function to create the dummy variable. For example, if you choose to create a dummy named yr2010, use =IF(A2=2010,1,0).

- ii. Next, **create** the interaction term variable: the product of the year dummy and the x variable (quantile).
- iii. Finally, **run** the multiple regression to test if the slopes differ (i.e. Panel (c) in Table A2). Do *not* include the year 2000 data. Your multiple regression should have $k = 3$ and $n = 199$. **Check** your output against that shown below.

EXCEL TIPS: To exclude the year 2000 data, filter the original data and then copy the data to a new worksheet.

Regression Statistics						
Multiple R	0.946833247					
R Square	0.896493197					
Adjusted R Square	0.894900784					
Standard Error	0.968412983					
Observations	199					

ANOVA						
	df	SS	MS	F	Significance F	
Regression	3	1583.92247	527.9741568	562.9780454	9.652E-96	
Residual	195	182.8756226	0.937823706			
Total	198	1766.798093				

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	74.660796	0.195152433	382.5768127	2.2311E-282	74.27591558	75.04567642
quantile	-0.056971543	0.003362608	-16.94266698	3.49941E-40	-0.063603292	-0.050339794
yr2010	4.432336033	0.275981518	16.06026395	1.55642E-37	3.888044164	4.976627901
yr2010xquantile	0.005202275	0.004749962	1.095224508	0.274769173	-0.00416562	0.01457017

Note: There is a tiny typo in Panel (c), Table A2 on page 161: the P-value should be 0.2748, not 0.2749. It should match the P-value for the interaction term above.

Interpretation tips: What does -0.056971543 mean? (It is the coefficient on quantile above.) In 1990, males living in a county ranked 10 percentiles poorer – for example, comparing the 70th to 80th percentile, which means a non-trivial increase in poverty – are expected to live about 7 months less on average. (To make the interpretation clear, notice how we converted from a fraction of a year to months and picked a non-trivial change in the quantile variable.)

Note: Alternatively, you may make the year 2010 the reference (omitted) category:

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	79.09313204	0.195144373	405.3057278	2.9317E-287	78.70826751	79.47799656
quantile	-0.051769268	0.003354849	-15.43117806	1.24003E-35	-0.058385714	-0.045152822
year1990	-4.432336033	0.275981518	-16.06026395	1.55642E-37	-4.976627901	-3.888044164
year1990xquantile	-0.005202275	0.004749962	-1.095224508	0.274769173	-0.01457017	0.00416562

Interpretation tips: What does -0.005202275 mean? (It is the coefficient on the interaction term above.) First, this coefficient is *not* statistically different from zero: the P-value is very large (0.27). Hence, these data do not rule out the possibility that mortality inequality, as measured by average life expectancy, is unchanged for males

between 1990 and 2010. A steeper negative slope in 1990 would mean that mortality inequality was more severe in 1990 compared to 2010, a less negative (flatter) slope would mean mortality inequality was less severe in 1990 compared to 2010, but we cannot reject that the slope is the same. The (slight) negative point estimate means that the slope was a bit steeper in 1990, but we cannot prove that mortality inequality lessened by 2010.

2. Consider Currie and Schwandt (2016) and use [mort.in.figure.3.table.a3.xlsx](#). Recall Figure 3 on page 108 and see Table A3 below. Figure 3 and Table A3 go together.

Table A3: Age-specific mortality in the richest and poorest county groups and slope of regression lines, 1990 vs. 2010

	3-year mortality (per 1,000) in 5% of the population living in								Slope of fitted regression line		
	counties with <i>lowest</i> poverty rate				counties with <i>highest</i> poverty rate						
	1990		2010		1990		2010				
	rate	std. err.	rate	std. err.	rate	std. err.	rate	std. err.	1990	2010	p-value of difference
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Males											
Age 0-4	2.38	0.07	1.33	0.05	4.49	0.09	2.39	0.07	0.020	0.009	<0.001
Age 5-19	1.31	0.03	0.86	0.03	2.89	0.04	1.52	0.03	0.015	0.006	<0.001
Age 20-49	5.23	0.04	4.46	0.04	12.45	0.07	8.56	0.06	0.068	0.034	<0.001
Age 50+	77.74	0.23	53.93	0.19	113.33	0.27	90.09	0.25	0.274	0.286	0.773
Age 65+	154.96	0.50	108.26	0.43	185.84	0.49	147.17	0.45	0.247	0.324	0.098
Females											
Age 0-4	1.91	0.07	1.17	0.05	3.62	0.09	2.04	0.07	0.017	0.008	<0.001
Age 5-19	0.62	0.02	0.42	0.02	1.10	0.03	0.67	0.02	0.004	0.002	<0.001
Age 20-49	2.66	0.03	2.34	0.03	5.19	0.04	4.80	0.04	0.023	0.021	0.705
Age 50+	72.27	0.20	58.43	0.18	84.91	0.21	76.78	0.20	0.098	0.158	0.032
Age 65+	132.35	0.39	109.46	0.35	136.36	0.35	124.08	0.34	0.052	0.155	0.007

Notes: Columns (1) to (8) show mortality rates for the bottom and top ventile of county groups, as plotted in Fig. 3 (age group 65+ is added), along with standard errors. Columns (9) and (10) report the slope of the fitted regression lines for 1990 and 2010 in Fig. 3, and (11) reports the p-value of the difference between the two slopes.

Figure of Table A3: Currie and Schwandt (2016), p. 15 of the appendix.

- (a) **Run** the simple regression shown in Panel A (aged 0-4 years) in Figure 3 for females for the year **1990**. **Verify** your slope matches what Table A3 reports in Column (9).
- (b) **Run** the simple regression shown in Panel A (aged 0-4 years) in Figure 3 for females for the year **2010**. **Verify** your slope matches what Table A3 reports in Column (10).
- (c) To check if the slopes in parts 2a (1990) and 2b (2010) differ in a statistically significant way requires a multiple regression.
 - i. **Create** a dummy variable either for year 1990 or for year 2010.
 - ii. **Create** an interaction term between that year dummy and the x variable (quantile).
 - iii. **Run** the multiple regression to test if the slopes differ (i.e. Column (11) in Table A3). Do *not* include the year 2000 data. Your multiple regression should have $k = 3$ and $n = 40$. **Verify** the P-value on the interaction term matches Column (11) of Table A3. (More exactly, the absolute value of your t test statistic should be 5.22.)

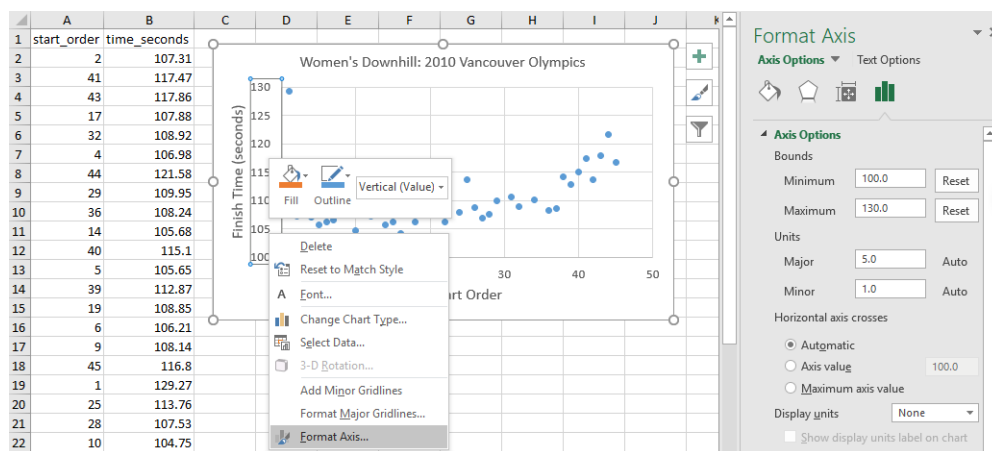
- (d) **Review** Table A3 and especially Columns (9), (10), and (11). Notice how your simple regression coefficients in parts 2a and 2b match up to your multiple regression coefficients in part 2c. (It was *not* necessary to run the simple regressions. These help you connect the original two lines (1990 and 2010) to the multiple regression coefficients.)

Test/exam examples: For examples of Currie and Schwandt (2016), see page 45.

3. Use [skiing_2010.xlsx](#) for women's downhill skiing at the 2010 Vancouver Winter Olympics.

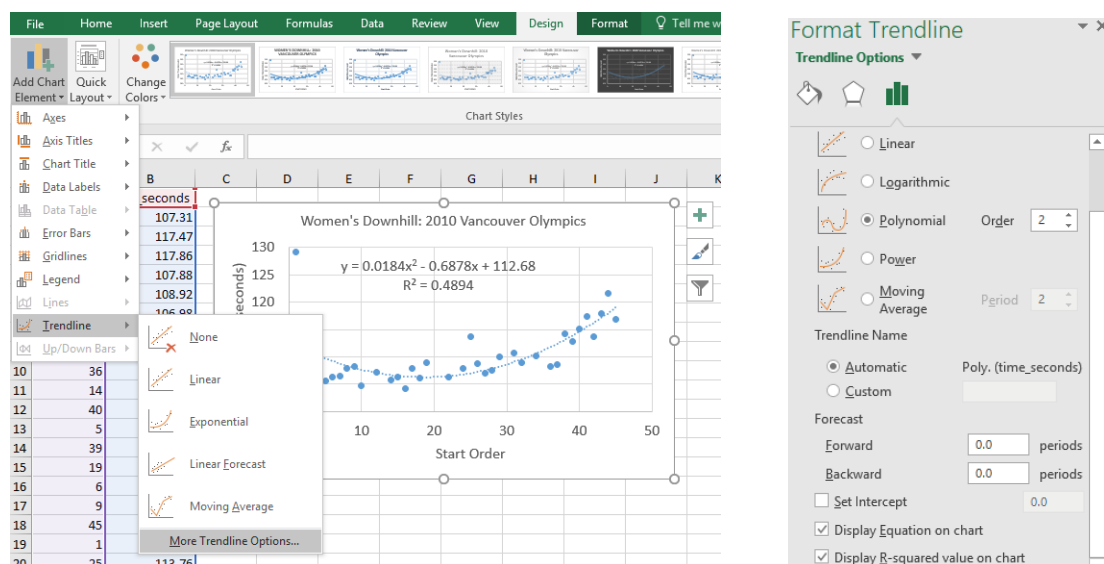
- (a) **Construct** a scatter plot to explain finish time (seconds) using starting order.

EXCEL TIPS: Copy the variables `start_order` and `time_seconds` to a new worksheet. The default scale of the vertical axis is not ideal. Hence, format the axis to range from 100 to 135: set the “Minimum” under “Bounds” to 100 (not 0.0) and the “Maximum” under “Bounds” to 130 (not 140.0). Under “Add Chart Element” you can add “Axis Titles.”



- (b) **Add** a fitted quadratic equation to explain finish time using starting order.

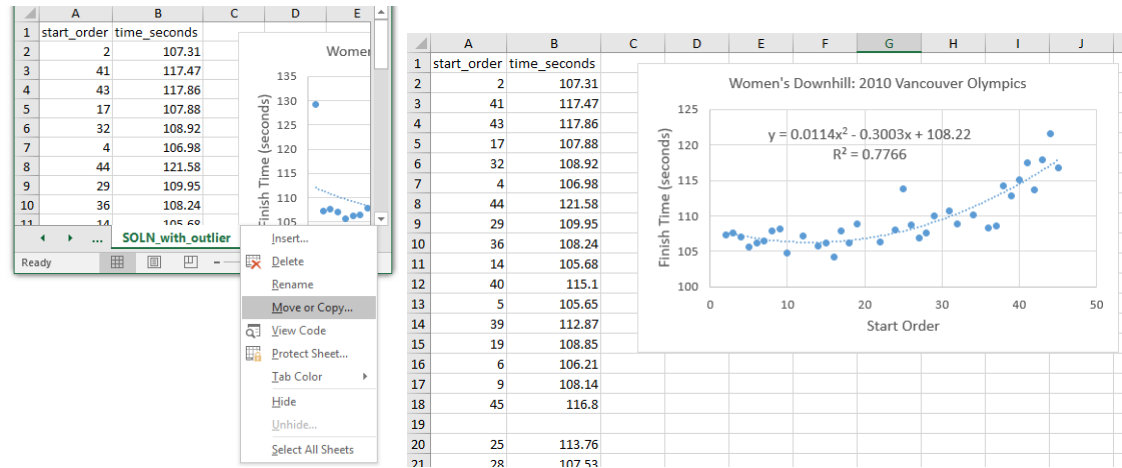
EXCEL TIPS: The Excel tip on page 40 shows how to add a trendline and display the equation. To find the quadratic functional form, select More Trendline Options...



Next, click “Polynomial” (Order 2) and check the boxes for “Display Equation on chart” and “Display R-squared value on chart.”

- (c) **Redo** part 3b but removing the outlier. (Klara Krizova crashed and lost a ski pole but got up to finish in last place with a time above 129 seconds.)

EXCEL TIPS: Create a duplicate copy of the worksheet from part 3b. Clear the cell with the outlier and it automatically updates. Name it (e.g. “Quad without outlier”).



Interpretation tips: What does -0.3003 mean? (It is the coefficient on x in the Excel output above.) We cannot interpret the coefficient on start order (-0.3003) by itself. The regression to explain finish time includes both start order and start order squared. We *cannot* consider a change in start order holding start order squared constant: obviously they would both change. The coefficients on start order (-0.3003) and start order squared (0.0114) only have meaning together. The best way to interpret them is to graph the OLS equation (as above). It shows a U-shaped parabolic relationship where the best starting positions (fastest times) are neither first nor last: the best position is around thirteenth (obtained by taking the derivative and setting it equal to zero: $0.0228x - 0.3003 = 0$).

- (d) Do the results in part 3c mean that the starting order is a very important determinant of finish time? To answer, we must remember (from the required readings) that start order is *not random* but is assigned based on prior performance (skill). What if we control for a measure of skill? **Run** a multiple regression that includes start order, start order squared, and FIS points (continuing to exclude the outlier from the previous part). **Verify** that your results match those below.

Regression Statistics	
Multiple R	0.897236627
R Square	0.805033565
Adjusted R Square	0.786755462
Standard Error	1.928968142
Observations	36

ANOVA					
	df	SS	MS	F	Significance F
Regression	3	491.6479433	163.8826478	44.04360529	1.81892E-11
Residual	32	119.0693789	3.720918092		
Total	35	610.7173222			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	105.9527426	1.467545848	72.19722829	5.18128E-37	102.9634496	108.9420357
start_order	0.001132798	0.174008971	0.006509997	0.994846227	-0.353311877	0.355577473
start_order_sq	0.001518216	0.005071081	0.29938713	0.766580644	-0.008811237	0.01184767
fis_points	0.096981974	0.044902079	2.159854842	0.038381859	0.005519431	0.188444516

- (e) Rerun the regression in part 3d but find the 90% CI estimates of the regression coefficients.

EXCEL TIPS: Use Regression under Data Analysis and see the screenshot below.

	A	B	C	D	E	F	G	H	I	J	K
1	time_seconds	start_order	start_order_sq	fis_points							
2	107.31	2	4	22.09							
3	117.47	41	1681	63.03							
4	117.86	43	1849	84.58							
5	107.88	17	289	4.51							
6	108.92	32	1024	18.74							
7	106.98	4	16	5.62							
8	121.58	44	1936	90.89							
9	109.95	29	841	5.92							
10	108.24	36	1296	31.46							
11	105.68	14	196	4.86							
12	115.1	40	1600	61.7							
13	105.65	5	25	5.45							
14	112.87	39	1521	53.49							
15	108.85	19	361	4.46							
16	106.21	6	36	5.56							
17	108.14	9	81	4.85							
18	116.8	45	2025	105.22							
19	113.76	25	625	14.01							
20	107.53	28	784	28.21							
21	104.75	10	100	5.31							

	Coefficients	Standard Error	t Stat	P-value	Lower 90.0%	Upper 90.0%
Intercept	105.9527426	1.467545848	72.19722829	5.18128E-37	103.4668832	108.438602
start_order	0.001132798	0.174008971	0.006509997	0.994846227	-0.29361904	0.295884636
start_order_sq	0.001518216	0.005071081	0.29938713	0.766580644	-0.00707163	0.010108063
fis_points	0.096981974	0.044902079	2.159854842	0.038381859	0.020922847	0.173041101

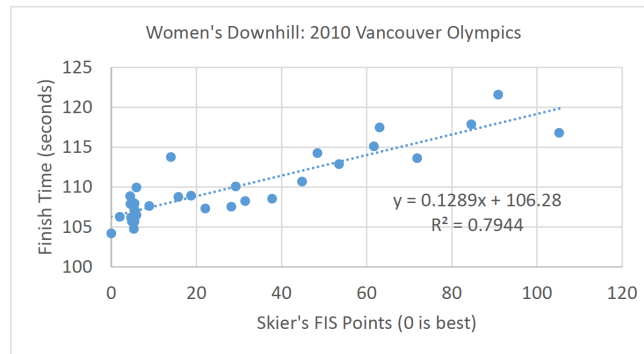
Interpretation tips: In the bottom right of the regression output above, what do 0.020922847 and 0.173041101 mean? We are 90% confident that after flexibly controlling for start order, women skiers in the 2010 Vancouver Olympics who had 10 more FIS points, which means a considerably *lower* skill level, on average take somewhere between 0.2 to 1.7 additional seconds to finish the downhill race. We can say “flexibly” rather than a second-order polynomial for start order: the point is that we controlled for start order. The interpretation for FIS points *must* be clear that higher points imply a *lower* the ranking (skill) of the skier (which is not obvious to those outside of alpine skiing).

4. **Note** the unusual result that the multiple regression in part 3d is *highly* statistically significant overall (huge F test statistic with a tiny P -value), while none of the coefficients are statistically significant at a 1% level. As explained in the required readings, this is an example of multicollinearity: very strong correlations among the x -variables.

- (a) Construct a correlation matrix. Verify it matches the below.

	time_seconds	start_order	start_order_sq	fis_points
time_seconds	1			
start_order	0.777476899	1		
start_order_sq	0.851918583	0.97333479	1	
fis_points	0.891280684	0.805469367	0.907093627	1

- (b) Note from the correlation matrix that the single best predictor of finish time is skill (as measured by FIS points). Construct a scatter plot with linear trend line to illustrate the relationship between finish time and skill. Verify it matches the below.



Interpretation tips: Referencing the results in parts 3c, 3d, 4a, and 4b, do the results in part 3c mean that the starting order is a very important determinant of finish time? No. While part 3c shows that a second-order polynomial of start order predicts finish time well – the R^2 is 0.78 – skiing skill is an important lurking variable. The correlation matrix in part 4a shows that a skier’s skill (as measured by FIS points) is *both* strongly correlated with finish time *and* starting order: 0.89 and 0.81, respectively. (FIS points are also strongly correlated with starting order squared: 0.91.) Once we control for a skier’s skill by including it as an explanatory variable in the regression to explain finish time (part 3d), starting order no longer has predictive power. While parts 4a and 3d show that multicollinearity affects the statistical tests for start order and start order squared, the results in part 4b indicate that – as expected – skill is a very important determinant of performance. Given that skill also directly affects starting order, the regression in part 3c *cannot* be interpreted to mean that starting order is a very important determinant of finish time. A good chunk of the apparently strong relationship between finish time and starting order simply reflects the (obvious) fact that the best skiers get the best positions and the best skiers (regardless of position) typically ski the best. This does not mean that skiing position is irrelevant, just that it is far less important than suggested by part 3c.

Test/exam examples: Women’s downhill Olympic skiing has appeared on a term test and a final exam. Also, questions directly derived from the skiing case but applied to the Zheng and Kahn (2017) data on Chinese air pollution have appeared.

- Question (4), [April 2022 Final Exam](#) (with [solutions](#))
- Question (2), [April 2019 Test #5](#) (with [solutions](#))
- Questions (10) - (17), [April 2016 Test #5](#) (with [solutions](#))

E.0.0 Practice questions for Module E

Q1. Before trying multistage questions with data analysis, practice using software to do hypothesis testing. The four parts below, (a) to (d), are entirely unrelated to each other.

- (a) Two variables have a correlation of 0.37142139 in data with 25 observations. Is that correlation statistically significant? If so, at which significance levels? Fully assess the strength of the evidence (in favor of the conclusion that the correlation is not zero) by computing the P-value.
- (b) A simple regression has an R^2 of 0.00831922 in data with 1,052 observations. Is the slope coefficient statistically significant? (In other words, test $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$.) If so, at which significance levels? Fully assess the strength of the evidence (in favor of the conclusion that the slope is not zero) by computing the P-value.
- (c) Suppose the multiple regression coefficient estimate for X_3 is 1.42013098 with a standard error of 0.18321045. There are 93 observations and five x variables.
 - i. What is the t test statistic for the hypothesis test $H_0 : \beta_3 = 0$ versus $H_1 : \beta_3 \neq 0$?
 - ii. What is the P-value for the hypothesis test $H_0 : \beta_3 = 1$ versus $H_1 : \beta_3 > 1$. Conclusion?
 - iii. What is the 99% confidence interval estimate of β_3 ?
 - iv. What is the critical t value for the hypothesis test $H_0 : \beta_3 = 2$ versus $H_1 : \beta_3 < 2$ given $\alpha = 0.01$?
- (d) For a multiple regression with 34 x variables and 1,420 observations, what is the critical value for the overall test of statistical significance for $\alpha = 0.001$? (Recall the formal hypotheses are $H_0 : \beta_1 = \beta_2 = \dots = \beta_{34} = 0$ versus $H_1 : \text{Not all } \beta\text{'s are zero.}$)

Q2. Recalling Johnson (1996), use [pct_body_fat.xlsx](#).

- (a) Run a multiple regression where the y-variable is percent body fat and the x-variables are age, weight, height, neck circumference, chest circumference, and abdominal circumference. Is the coefficient on weight positive or negative? What is the P-value for the test of the statistical significance of the coefficient on weight?
- (b) Run a simple regression where the y-variable is percent body fat and the x-variable is weight. Is the coefficient on weight positive or negative? What is the P-value for the test of the statistical significance of the coefficient on weight?
- (c) Run a multiple regression where the y-variable measures percent body fat and the three x-variables measure weight, height, and age where you first standardize all four variables. The only difference between this regression and the one in “Just Checking,” which you replicated in Module B.2, is that this regression uses standardized data. For each of these, indicate whether the value is higher, lower or the same in the regression with standardization: the R-squared, the s_e , the SST, and the F statistic.

Q3. Recalling Levitt et al. (2013), use [learn_do_daily.xlsx](#) for all subparts.

- (a) Replicate the simple regression results in Table 1, Panel B, Column (1), page 142.

Hint 1: Remember the steps for the *weekly* data: Step 2 on page 143 and Step 3 on page 144. **Hint 2:** Like the authors, *use only days where at least 20 cars are produced*, which is what they did for the *daily* data (in contrast to at least 100 in the weekly data).

- Note: Excel does not compute robust standard errors. However, for (0.006), the regular and robust standard errors differ only a little: rounded they are equal.
- (b) Replicate the multiple regression results in Table 1, Panel B, Column (2), page 142. In addition to **Hint 1** and **Hint 2** above, **Hint 3**: Remember Step 4 on page 145, where we did the same replication as this question but for the weekly data.
- For Table 1, Panel B, Column (2) you will obtain (0.011) instead of (0.014), which makes little difference because the coefficient is precisely estimated and highly statistically significant either way. For (0.0002), the regular and robust standard errors differ only a little: rounded they are equal.
- Q4.** Levitt et al. (2013) check if the results in Table 1 hold up to scrutiny by exploring many variations on the analysis, one variation at a time. In addition to re-running everything for weekly and daily data, they consider: an alternative measure of learning (hours per car instead of defects per car: with learning, production should get faster as well as involving fewer mistakes), first shift versus second shift considered separately, and each model of car considered separately (recall there are three models in the data). For example, see Table 2, which looks at ONE variation: it is JUST LIKE Table 1 except that it considers the first shift and second shift separately. Use [learn_do_weekly.xlsx](#) for all subparts.

TABLE 2
SHIFT SPECIFIC LEARNING BY DOING AND RAMP UP SPILLOVERS

	FIRST SHIFT		SECOND SHIFT	
	Weekly (1)	Daily (2)	Weekly (3)	Daily (4)
A. Shift Specific Learning by Doing				
Estimated learning rate (β)	-.323* (.010)	-.346* (.008)	-.154* (.011)	-.088* (.020)
Observations	47	224	39	190
R^2	.946	.907	.782	.158

Figure of Table 2: Levitt et al. (2013), p. 663.

- (a) Replicate the simple regression results in Table 2, Column (1). **Hint 1**: Remember the steps for *overall production* (both shifts combined) with the weekly data: Step 2 on page 143 and Step 3 on page 144. **Hint 2**: You will need to use variables `wk_prod_s1` and `wk_defs_s1` (instead of `wk_prod` and `wk_defs`). **Hint 3**: Like the authors, use only weeks where at least 100 cars are produced during the FIRST SHIFT.
- Note: Excel does not compute robust standard errors. You will obtain (0.011) instead of (0.010).
- (b) Replicate the simple regression results in Table 2, Column (3). **Hint 1**: Remember the steps for *overall production* (both shifts combined) with the weekly data: Step 2 on page 143 and Step 3 on page 144. **Hint 2**: You will need to use variables `wk_prod_s2` and `wk_defs_s2` (instead of `wk_prod` and `wk_defs`). **Hint 3**: Like the authors, use only weeks where at least 100 cars are produced during the SECOND SHIFT. **Hint 4**: Remember that you can only obtain a non-missing value for the natural log of cumulative production if the cumulative production is greater than zero. This hint is relevant only for Shift 2

because of the abrupt scale up of production, which means that the first week of Shift 2 production (W41/Y1) has over 100 cars produced even though there was zero previous production experience in Shift 2. You cannot include the observation with zero previous Shift 2 production because the natural log of zero does not exist. (Remember Section 3.2 on page 4.)

- You will obtain (0.013) instead of (0.011).

- (c) Continuing with the previous part, run a multiple regression where you also include a time trend as an x variable. Report the coefficient on cumulative production and the coefficient on the time trend. Also, report the R-squared.

Q5. Recalling Currie and Schwandt (2016) and Figure 2, use [mort_in_figure_2.table_a2.xlsx](#).

- (a) Run the simple regression that corresponds to Panel B (females) in Figure 2 for *only* the year 1990. What is the value of the F test statistic?
- (b) Run the appropriate multiple regression to compare the relationship between life expectancy at birth and the poverty rank of county group in 1990 versus 2010 for females. What is the R-squared? What is the P-value for the test that the slope in 1990 for females equals the slope in 2010 for females? Also, what is the point estimate of the difference in the slopes and its s.e.?

Q6. Recall Currie and Schwandt (2016) and Figure 3. In Figure 3, age group 0-4 combines age group 0 with age group 1-4. Consider the question: Are there systematic differences across these two age groups (suggesting caution in combining them into one group)? To answer, use [mort_in_disaggregate_age_groups.xlsx](#). These data refer to data at the level of a year of age (e.g. 4 year olds) as opposed to more aggregate age groups (e.g. 0-4 year olds). To narrow the question, focus on females in 2010. Run an appropriate multiple regression to produce a graph like those shown in Figure 3 *except* that instead of comparing 1990 with 2010 for a particular sex and a particular age group, compare age group 0 with age group 1-4 for the same sex (females) in the same year (2010). Use the *unadjusted* variables.¹⁶ (The reason you do not need the adjusted values is that this question is asking you to compare two age groups in the *same year*.) What is the adjusted R-squared? What is the point estimate of the intercept and slope for age group 0? Age group 1-4? For the test of a difference in the slopes between the two age groups, what is the absolute value of the t test statistic and the P-value? For the test of a difference in the intercepts, what is the absolute value of the t test statistic and the P-value?

Q7. Use [skiing_2018.xlsx](#) for women's downhill skiing at the 2018 PyeongChang Winter Olympics.

- (a) For $\widehat{time_seconds} = b_0 + b_1 * start_order + b_2 * start_order_squared$, what are the OLS coefficients (i.e. b_0 , b_1 , and b_2)? What does the shape of the relationship between finish time and start order look like? Are there outliers?
- (b) Continuing with the previous part, what is the slope of the relationship between finish time and start order for a skier in position 2? How about a skier in position 20?

¹⁶You may be confused as to why [mort_in_disaggregate_age_groups.xlsx](#) even contains adjusted variables given that it seems not to be combining age groups: i.e. it is disaggregate. However, even to get to these data there has been aggregation of children of different ages (e.g. newborns with 10 month olds; 2 year olds with 4 year olds).

- (c) Run a multiple regression explaining finish time using start order, start order squared, and skill as measured by FIS points. What is the value of the s_e (include the units of measurement)? How about the R^2 ? The adjusted R^2 ?

Q8. Use [skiing_2010.xlsx](#). Run a multiple regression explaining finish time using start order, start order squared, skill as measured by FIS points, the skier's year of birth, and a dummy (indicator) variable for the skier KRIZOVA, Klara. If a skier starts in position two, has twelve FIS points, is born in 1983, and finishes in 108.07 seconds, according to the regression, this is _____ seconds slower than predicted.

Q9. Recall Levinson (2016) and use [calif_energy_regressions.xlsx](#).

- (a) For natural gas, what is the *exact* center point for the band for homes constructed from 2001-2004 in Figure 4 on page 151 for the “No controls” case?
- (b) Continuing, add controls for the natural log of the size of the house, the natural log of the number of residents, and central air conditioning. What is the center point for the band for homes constructed from 2001-2004 for this “Some controls” case?

For extra practice, additional questions, with an ^e superscript (e for extra), are next.

Q^e1. Recalling Currie and Schwandt (2016), Figure 3 and Table A3, use [mort_in_figure_3_table_a3.xlsx](#).

- (a) Consider males aged 0-4 years. Making the year 1990 the reference category (aka the omitted category), run the multiple regression that corresponds with the first row of results in Table A3. Verify that your results match those in the table.
- (b) Repeat the analysis in the previous part but make the year 2010 the reference category. Verify that these results also match those in the table.
- (c) Suppose that Table A3 compared 2000 with 2010, not 1990 with 2010. In other words, suppose Column (9) is for the year 2000, not 1990. Conduct an appropriate analysis to find the values for Columns (9), (10) and (11) for females aged 5-19.

Q^e2. Consider Currie and Schwandt (2016) and use [mort_in_figure_3_table_a3.xlsx](#).

- (a) Run a simple regression for females aged 0-4 in 1990. (i.e. replicating the blue line for females in Panel A of Figure 3 on page 108).
- (b) Assess whether we can conclude that the slope is less than 0.02.¹⁷ If so, at which significance levels? (Hint: This requires testing $H_0 : \beta = 0.02$ versus $H_1 : \beta < 0.02$.)
- (c) Find the 90% confidence interval estimate of the slope coefficient.

Q^e3. Recalling Johnson (1996), use [pct_body_fat.xlsx](#). In a multiple regression analysis with three explanatory (x) variables, what happens as you increase the sample size? Specifically, which of these should you expect to go up, go down, or remain unchanged: SST, SSR, SSE, s_y , s_e , R-squared, s_{b1} , s_{b2} , s_{b3} , P-value for the F test, P-value for the t tests for each coefficient?

¹⁷Why 0.02? It is a specific value to try: the classic test of statistical significance, where the null specifies a value of 0, is not the only test in the world. Also, 0.02 does have some meaning. Looking at the blue line for females in Panel A of Figure 3 on page 108, a slope of 0.02 would correspond to the female 0-4 mortality being about twice as high for the poorest county versus the richest county: $slope = \frac{\Delta y}{\Delta x} \approx \frac{4-2}{100-0} = 0.02$.

Scrutinize your intuition by running three multiple regressions where the y variable is percent body fat and the x variables are age, weight, and height and where the first regression uses observations 1 - 10 (small sample), the second uses observations 11 - 50 (medium sample), and the third uses observations 51 - 200 (larger sample).

- Q^e4.** Recall Levitt et al. (2013). Indicate which level of data – weekly or daily – yields a higher R-squared value in Table 1 (p. 142), Panel A, Column (1). Do the concepts discussed in Section 19.4 of the textbook “Working with Summary Values” apply in this context?
- Q^e5.** Recalling Currie and Schwandt (2016), use [mort_in_figure_3.table_a3.xlsx](#)
- Run the simple regression illustrated in Panel B of Figure 3 for males in 1990. What is the 95% confidence interval estimate of the slope?
 - Suppose that instead of being the three-year mortality rate per 1,000 population, the y-variable were the three-year mortality rate. In that case, what is the 95% confidence interval estimate of the slope?
 - In 2010, how is mortality related to poverty for males aged 20-49 versus males aged 50+? First, we comparing 2010 data with 2010 data: hence, there is NO need to use adjusted mortality, which is only necessary to compare different years because composition may have changed over time. However, there is a new challenge: the level of deaths is much higher in the older age group so the slopes are not directly comparable. To see that, suppose that deaths per 1,000 goes up by 1%: if the level of deaths were 5 then that would be an increase to 5.05 (a 0.05 change) but if the level of deaths were 60 that would be an increase to 60.6 (a 0.60 change). The 0.60 change appears much bigger than the 0.05 change, but if we recognize the much higher level of deaths in the older age group, these two changes are comparable in percentage terms. Hence, we can take the natural log of the y-variable to interpret $100 \times$ “the slope coefficient” as the percentage change in mortality given a one unit increase in the poverty ranking. But, the scatter plots are already straight! Hence, first verify that the two scatter plots (aged 20-49 and aged 50+) for males in 2010 are still straight after the natural log transformation of the y-variable. Run two simple regressions (one for each age group) after the natural log transformation of the y-variable. What is the point estimate of the slope coefficient for males aged 20-49? For males 50+?
- Q^e6.** Recall Currie and Schwandt (2016) and Figure 3. In Figure 3, the county groups are organized into quantiles based on their poverty rates. An alternative way to rank counties is by median income. Do the main results hold up using an alternative measure of richness/poorness? To answer, use [mort_in_median_income_quantile.xlsx](#). These data refer to counties being sorted into quantiles based on median income rather than poverty rates (as used for the main results). To narrow the question, focus on males aged 0-4. Run an appropriate multiple regression to produce a graph like that in Figure 3, Panel A (age group 0-4) for males *except* that you focus on comparing 1990 and 2010 only (leave out 2000) and you use quantiles based on median income instead of poverty rate. For 1990, what is the point estimate of the intercept and slope? For 2010? For the test of a difference in the slopes between the two years, what is the absolute value of the t test statistic and the P-value? Do these results contradict those in Figure 3, Panel A (age group 0-4) for males?

Answers for Module E practice questions:

- A1.** (a) Compute an F test statistic of 3.68070610 and a P-value of 0.06754125. This correlation is statistically significant at a 10% significance level, but we do not have sufficient proof to meet a 5% significance level. We have some evidence that the correlation is not zero, but it is hardly overwhelming evidence.
- (b) Compute an F test statistic of 8.80846052 and a P-value of 0.00306638. The slope coefficient is statistically significant at a 1% significance level.
- (c) i. The t test statistic is 7.75136451.
- ii. The t test statistic is 2.29316057. The P-value is 0.01212493. We can conclude at a 5% significance level that β_3 is larger than 1, but we haven't met a 1% burden of proof.
- iii. For a 99% confidence level, LCL is 0.93764127 and the UCL is 1.90262069.
- iv. The critical value is -2.36997678: must obtain a t test statistic below -2.36997678 to conclude that β_3 is less than 2 at a 1% significance level.
- (d) The critical value is 1.94212641: must obtain an F test statistic of at least 1.94212641 for the multiple regression to be statistically significant overall at a 0.1% significance level.
- A2.** (a) The coefficient on weight is negative. The P-value is 0.522.
- (b) The coefficient on weight is positive. The P-value is less than 0.0001.
- (c) The R-squared is the same in both. The s_e is higher in Regression #1. The SST is higher in Regression #1. The F statistic is the same in both.
- A3.** (a) Verify that you get the same coefficient, s.e., number of observations, and R-squared as reported in Table 1, Panel B, Column (1) (p. 142).
- (b) Verify that you get the same coefficients, s.e.'s, number of observations, and R-squared values as reported in Table 1, Panel B, Column (2) (p. 142) EXCEPT that you get 0.011 as the s.e. on the cumulative production coefficient (0.014 is the robust s.e.).
- A4.** (a) Check your "slope" coefficient, sample size and R-squared against Table 2 (p. 169).
- (b) Check your "slope" coefficient, sample size and R-squared against Table 2 (p. 169). If you are having trouble, make sure you remembered to only include weeks when at least 100 cars are produced *in the second shift*. Your sample size should be 39.
- (c) The coefficient on cumulative production is -0.1567641. The coefficient on the time trend is 0.0002902. The R-squared is 0.7821. (This explains why, after Table 1, the authors do not bother to also report the results with a time trend: it doesn't make much difference.)
- A5.** (a) $F = 205.94$ (with $k = 1$ and $n - k - 1 = 97$)
- (b) The R-squared is 0.7913. The P-value is 0.044. The point estimate of the difference in slopes is 0.0082542 and the s.e. is 0.0040636.
- A6.** Run a multiple regression with $n = 40$ and $k = 3$: the x variables are quantile, an age dummy (either 0 yrs or 1-4 yrs), and an interaction between the age dummy and quantile. (It does not matter which of the two age groups you make the omitted (reference) category.) The adjusted R^2 is 0.9878. For age group 0, the point estimate of the intercept is 4.520717 and the slope is

0.0317943. For age group 1-4, the point estimate of the intercept is 0.3862545 and the slope is 0.0029378. For testing for a difference in slopes between the two age groups, the absolute value of the t test statistic is 8.06 and the P-value is less than 0.0001. For testing for a difference in intercepts between the two age groups, the absolute value of the t test statistic is 19.29 and the P-value is less than 0.0001. Female infants are much more likely to die than 1-4 year olds. Further, there is far more inequality for these most vulnerable children: the death rates of females 0 years old are substantially higher in poorer counties compared to richer counties.

- A7.** (a) Insert a scatter chart – finish time (y-axis) and start order (x-axis) – and add a polynomial, order 2, trend line and display the equation: $\widehat{time_seconds} = 100.79 - 0.1197 * start_order + 0.0067 * start_order_squared$. It has a U-shape like the 2010 Vancouver Olympics. Note the negative coefficient $b_1 = -0.1197$. There are no outliers in these 2018 data.
- (b) The slopes are -0.0927 and 0.1497, respectively. (Use the Excel output, not the rounded values in the previous part, to get these accurate to four decimal places. See Part 3c on page 165 for more explanation.)
- (c) The value of the s_e is 0.81381 seconds. This measures the amount of “scatter” of finish time (the y-variable) left over after we control for start order, start order squared, and skill (FIS points). Given the s.d. of finish time is 2.135 seconds, these three variables explain a good chunk of the variation in finish time. This is consistent with the high R^2 of 0.8692 and Adjusted R^2 of 0.8547. About 87 percent of the variation in women’s downhill ski time at the 2018 Olympics is explained by variation in skiers’ skill (measured by the FIS score) and start order (included as a second-order polynomial).
- A8.** Use the coefficients from the multiple regression to obtain the predicted finish time: $107.12158 = 112.53115082 + 0.00073644*2 + 0.00154559*4 + 0.09671351*12 + -0.00331709*1983 + 19.99184608*0$. (The last term is zero because this skier is not Klara.) Hence, the residual is $0.9484 = 108.07 - 107.12158$. This means that this hypothetical skier finished nearly one second (0.9484 seconds) slower than predicted.
- A9.** (a) In Module E.2, we did this for electricity. Remember to take the natural log of the variable `gas_no.0`. The center point of that band is -0.120605543: see regression below.

Regression Statistics	
Multiple R	0.23517713
R Square	0.055308282
Adjusted R Square	0.054389992
Standard Error	0.505148952
Observations	12358

ANOVA					
	df	SS	MS	F	Significance F
Regression	12	184.4293647	15.36911373	60.22959078	1.1813E-142
Residual	12345	3150.141094	0.255175463		
Total	12357	3334.570459			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	3.92204526	0.016386128	239.3515641	0	3.889925891	3.954164629
yr_2009	-0.226603168	0.009248196	-24.50241894	1.6782E-129	-0.244731077	-0.20847526
constr_40_49	-0.090898824	0.023742596	-3.828512493	0.000129557	-0.137438019	-0.044359629
constr_50_59	-0.054438341	0.019095999	-2.850772058	0.004368552	-0.091869481	-0.0170072
constr_60_69	-0.009677613	0.019380455	-0.499349091	0.617542376	-0.047666332	0.028311106
constr_70_74	-0.047759584	0.022457639	-2.126652043	0.033468783	-0.091780064	-0.003739104
constr_75_77	-0.03518764	0.025692201	-1.369584475	0.17084156	-0.085548366	0.015173087
constr_78_82	-0.088214107	0.02331719	-3.783222027	0.000155542	-0.133919441	-0.042508773
constr_83_92	-0.100505338	0.01995374	-5.036917367	4.79772E-07	-0.139617783	-0.061392892
constr_93_97	-0.109718755	0.025924868	-4.232181752	2.33108E-05	-0.160535546	-0.058901964
constr_98_00	-0.104126285	0.026454277	-3.93608505	8.3276E-05	-0.1559808	-0.05227177
constr_01_04	-0.120605543	0.028489603	-4.23331779	2.31935E-05	-0.176449615	-0.064761472
constr_05_08	-0.107782277	0.036970513	-2.915357877	0.003559197	-0.180250256	-0.035314297

(b) It is -0.367076457, which is taken from the regression below.

Regression Statistics	
Multiple R	0.41361024
R Square	0.17107343
Adjusted R Square	0.170065984
Standard Error	0.4732441
Observations	12358

ANOVA					
	df	SS	MS	F	Significance F
Regression	15	570.4564075	38.03042717	169.8090323	0
Residual	12342	2764.114051	0.223959978		
Total	12357	3334.570459			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0.725787057	0.085957457	8.443561334	3.42036E-17	0.557297014	0.8942771
ln_sq_feet	0.42507709	0.011740885	36.20485859	7.1328E-273	0.402063121	0.448091059
ln_num_res	0.136400795	0.008285721	16.46215171	3.00549E-60	0.120159487	0.152642102
central_ac	-0.004754465	0.009457434	-0.502722505	0.615168355	-0.023292514	0.013783584
yr_2009	-0.224053133	0.008685177	-25.79718572	5.6521E-143	-0.241077437	-0.207028828
constr_40_49	-0.069211278	0.022260341	-3.109174138	0.00188038	-0.112845024	-0.025577531
constr_50_59	-0.075325138	0.017935694	-4.199733601	2.69105E-05	-0.1104819	-0.040168375
constr_60_69	-0.092980088	0.018376749	-5.05965913	4.26046E-07	-0.129001388	-0.056958788
constr_70_74	-0.149928817	0.021339864	-7.025762521	2.24085E-12	-0.191758284	-0.10809935
constr_75_77	-0.164101925	0.024527287	-6.690586115	2.31879E-11	-0.212179239	-0.116024611
constr_78_82	-0.217196097	0.022401493	-9.695608281	3.77538E-22	-0.261106522	-0.173285672
constr_83_92	-0.254884849	0.019668045	-12.95933834	3.68942E-38	-0.29343729	-0.216332409
constr_93_97	-0.296568773	0.025264057	-11.73876281	1.18811E-31	-0.346090271	-0.247047275
constr_98_00	-0.319347642	0.025876376	-12.34128167	8.70442E-35	-0.370069381	-0.268625904
constr_01_04	-0.367076457	0.027863825	-13.1739437	2.28447E-39	-0.421693906	-0.312459008
constr_05_08	-0.387237201	0.035786034	-10.82090308	3.62868E-27	-0.457383417	-0.317090985

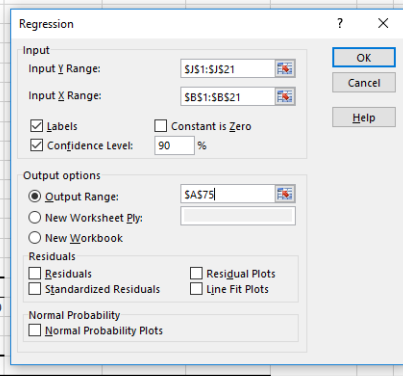
Answers to the additional questions for extra practice.

A^e1. (a) Verify that you obtain same 1990 slope, 2010 slope and P-value of the difference as reported in Table A3 for males aged 0-4 in Currie and Schwandt (2016). (The OLS point esti-

mates being: $\text{adjusted_mortality_hat} = 2.479075 + 0.0204704 \cdot \text{quantile} - 1.071161 \cdot \text{yr2010} - 0.011643 \cdot \text{quantile} \cdot \text{yr2010}$.)

- (b) Verify that you obtain same 1990 slope, 2010 slope and P-value of the difference as reported in Table A3 for males aged 0-4 in Currie and Schwandt (2016). (The OLS point estimates being: $\text{adjusted_mortality_hat} = 1.407914 + 0.0088274 \cdot \text{quantile} + 1.071161 \cdot \text{yr1990} + 0.011643 \cdot \text{quantile} \cdot \text{yr1990}$.)
- (c) Run a multiple regression with $n = 40$ and $k = 3$: the x variables are quantile, a year dummy, and an interaction between the year dummy and quantile. Comparing females aged 5-19 years between 2000 and 2010, Column (9) would be 0.0026731, Column (10) would be 0.0020822 and Column (11) would be 0.235.

- A^e2.** (a) Verify that you get $\hat{y} = 1.9404 + 0.0166 \cdot x$, $n = 20$. If your results do not match, make sure you are running your regression on the correct subset of the data: females aged 0-4 in 1990.
- (b) Verify that you obtain a t test statistic of -2.256239 and a P-value of 0.018366, which means that we can conclude at the 5% significance level that the slope is not as steep as 0.02 (quite a bit of inequality), but we do not have sufficient evidence to conclude that it is less steep than 0.02 at a 1% significance level.
- (c) Verify that you obtain $[0.0140277, 0.0192185]$. Note that there are two ways to compute these results in Excel. First, you can use your regression output from part 2a and the T.INV function to plug into $b \pm t_{\alpha/2} s_b$ with $\nu = n - k - 1$. Make sure your T.INV function returns a value of 1.734064, which is the correct value of $t_{\alpha/2}$ for a 90% CI with 18 degrees of freedom. Second, you can request a 90% confidence level when running the regression, which is illustrated below.



	df	SS	MS	F	Significance F
Regression	1	4.593946825	4.593946825	123.3557617	1.73526E-09
Residual	18	0.670346011	0.037241445		
Total	19	5.264292837			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 90.0%	Upper 90.0%
Intercept	1.940436074	0.089645479	21.64566584	2.44939E-14	1.752097912	2.128774235	1.784985112	2.095887036
quantile	0.016623105	0.001496692	11.1065639	1.73526E-09	0.013478672	0.019767538	0.014027746	0.019218464

- A^e3.** You should expect that the SST, SSR, and SSE all go up with an increase in the sample size: these are sums of squares (always positive) so more observations means bigger sums. You should expect no change in the s_y , s_e and R-squared with an increase in the sample size. These three relate to the underlying variability of percent body fat across males (s_y) and the amount of scatter about the regression line (s_e and R-squared): there is no reason to expect these to change in a systematic way as we increase the sample size. You should expect the values of s_{b1} , s_{b2} , s_{b3} , the P-value for the F test, and the P-value for the t tests to all decrease with

an increase in the sample size. These are all measures of sampling error and sampling error decreases as the sample size increases. Make sure to convince yourself of all of this by doing even more samples if necessary: there is nothing magic about the three regressions you were asked to run to represent three different sample sizes. Also, any given sample could produce statistics that deviate from expectation: look at more to convince yourself.

- A^e4.** The weekly data has a higher R-squared value (0.961) compared to the daily data (0.931). This is what we would expect as the concepts in Section 19.4 of the textbook “Working with Summary Values” do apply. If we aggregated up further to monthly data, we’d expect an even higher R-squared.
- A^e5.** (a) The 95% confidence interval estimate of the slope is (0.0132867, 0.0170393): in other words, the point estimate of the slope is 0.015163 and the margin of error is 0.0018763.
- (b) The 95% confidence interval estimate of the slope is (0.0000133, 0.000017): in other words, the point estimate of the slope is 0.0000152 and the margin of error is 0.0000019. Note that it is NOT necessary to run a second regression to obtain these results: they are a simple rescale of the original results recognizing the change in the units of the y variable.
- (c) Both scatter plots are still straight. The coefficient for 2010 for males aged 20-49 years is 0.0052268. The coefficient for 2010 for males aged 50+ years is 0.0030183. Hence, adjusting for the higher level of deaths, mortality inequality is actually less for the older males (remember, flatter is better). If we just looked at the slopes without the natural log transformation of the y variable, we would have gotten 0.0343925 for males aged 20-49 and 0.2127863 for males aged 50+, which would give the misleading impression that mortality inequality is greater for older males (remember, steeper is worse).
- A^e6.** Run a multiple regression with $n = 40$ and $k = 3$: the x variables are quantile, a year dummy (either 1990 or 2010), and an interaction between the year dummy and quantile. (It does not matter which of the two years you make the omitted (reference) category.) For 1990, the point estimate of the intercept is 4.266096 and the slope is -0.013231. For 2010, the point estimate of the intercept is 2.448319 and the slope is -0.0108919. For testing for a difference in slopes between the two years, the absolute value of the t test statistic is 0.90 and the P-value is 0.375. The results are different from Figure 3 in that there is no significant difference in the slopes for males aged 0-4 years between 1990 and 2010: there is a significant difference if we do the analysis by poverty rates rather than median income. However, you should NOT say that the results contract each other because the slopes are positive in the original analysis and negative in this new analysis. The switch in signs is not a contradiction and is to be expected given that poverty rates and median incomes are negatively correlated.

F References

- Andreoni, James, and Lise Vesterlund. 2001. "Which is the Fair Sex? Gender Differences in Altruism." *The Quarterly Journal of Economics*, 116(1): 293-312. <https://doi.org/10.1162/003355301556419>
- Carlin, Bruce I., Li Jiang, Stephen A. Spiller. 2017. "Millennial-Style Learning: Search Intensity, Decision Making, and Information Sharing." *Management Science*, Online. DOI: 10.1287/mnsc.2016.2689. <https://doi.org/10.1287/mnsc.2016.2689>
- Carlin, Bruce I., Li Jiang, Stephen A. Spiller. 2014. "Learning Millennial-Style." *NBER Working Paper*, No. 20268. <http://www.nber.org/papers/w20268.pdf> (NOTE: Authors redid the experiment and collected fresh data for 2017 publication.)
- City of Toronto (online). Open Data. "Wellbeing Toronto." <http://www.toronto.ca/wellbeing>.
- "Wellbeing Toronto - Housing." Retrieved June 6, 2017 from <https://www1.toronto.ca/wps/portal/contentonly?vgnextoid=f5c12c077444d410VgnVCM10000071d60f89RCRD>
 - "Wellbeing Toronto - Demographics." Retrieved June 6, 2017 from <https://www1.toronto.ca/wps/portal/contentonly?vgnextoid=4482904ade9ea410VgnVCM10000071d60f89RCRD>
- Currie, Janet, and Hannes Schwandt. 2016. "Mortality Inequality: The Good News from a County-Level Approach." *Journal of Economic Perspectives*, 30(2): 29-52. DOI: 10.1257/jep.30.2.29. <https://www.aeaweb.org/articles?id=10.1257/jep.30.2.29>
- Dubner, Stephen J. 2015. "How Efficient Is Energy Efficiency? A New Freakonomics Radio Podcast." Appeared on the *Freakonomics* website on February 5, 2015. <http://freakonomics.com/podcast/how-efficient-is-energy-efficiency-a-new-freakonomics-radio-podcast/>.
- The Economist. 2016. "Death and money, Looking up: The link between income and mortality rates is weakening." Appeared in U.S. print edition on May 14, 2016. <https://www.economist.com/news/united-states/21698702-link-between-income-and-mortality-rates-weakening-looking-up>.
- The Economist (online). 2017. "Interactive currency-comparison tool: The Big Mac index." <http://www.economist.com/content/big-mac-index>; On Jun. 13, 2017, downloaded data posted on Jan. 12, 2017 from <http://infographics.economist.com/2017/databank/BMFile2000toJan2017.xls> and figure from <http://www.economist.com/content/big-mac-index>.
- Feenstra, Robert C., Robert Inklaar, and Marcel P. Timmer. 2015. "The Next Generation of the Penn World Table." *American Economic Review*, 105(10): 3150-3182. DOI: 10.1257/aer.20130954. <https://www.aeaweb.org/articles?id=10.1257/aer.20130954>
- Penn World Table 8.0, Released Jul. 2, 2013. DOI: 10.15141/S5159X. <https://www.rug.nl/ggdc/productivity/pwt/pwt-releases/pwt8.0>. Retrieved June 9, 2015.
 - Penn World Table 9.0, Released Jun. 9, 2016. DOI: 10.15141/S5J01T. <https://www.rug.nl/ggdc/productivity/pwt/pwt-releases/pwt9.0>. Retrieved July 31, 2019.

- Penn World Table 9.1, Released Apr. 30, 2019. DOI: 10.15141/S50T0R. <https://www.rug.nl/ggdc/productivity/pwt/pwt-releases/pwt9.1>. Retrieved July 31, 2019.
- Penn World Table 10.0, Released Feb. 18, 2021, Updated June 18, 2021. DOI: 10.15141/S5Q94M. <https://www.rug.nl/ggdc/productivity/pwt/pwt-releases/pwt100>. Retrieved July 6, 2021.
- Penn World Table 10.01, Released Jan. 23, 2023. DOI: 10.34894/QT5BCC. <https://www.rug.nl/ggdc/productivity/pwt/>. Retrieved May 1, 2023.

Fitzgerald, Jay. 2018. "Postdoctoral Fellowships and Career Choice in Science." *The NBER Digest*, July 2018: 6. <http://www.nber.org/digest/jul18/jul18.pdf>

Google (online). "Google Finance: Stock market quotes, news, currency conversions & more." <https://www.google.ca/finance>

- "S&P/TSX Composite index (INEXTSI:OSPTX)." Retrieved July 21, 2017, from <https://www.google.ca/finance/historical?cid=9291235>
- "Prices for Apple Inc. (NASDAQ: AAPL)." Retrieved June 3, 2017, from <https://www.google.ca/finance/historical?cid=22144>
- "Prices for Amazon.com, Inc. (NASDAQ:AMZN)." Retrieved June 3, 2017, from <https://www.google.ca/finance/historical?cid=660463>

Government of Ontario (online). "Public sector salary disclosure." <https://www.ontario.ca/page/public-sector-salary-disclosure>

- Government of Ontario. 2024. "Public sector salary disclosure 2024: all sectors and seconded employees." Published: Mar. 28, 2025. <https://www.ontario.ca/public-sector-salary-disclosure/2024/all-sectors-and-seconded-employees/>. Data downloaded on Jul. 30, 2025.
- Government of Ontario. 2023. "Public sector salary disclosure 2023: all sectors and seconded employees." Published: Mar. 28, 2024. <https://www.ontario.ca/public-sector-salary-disclosure/2023/all-sectors-and-seconded-employees/>. Data downloaded on Apr. 3, 2024.
- Government of Ontario. 2022. "Public sector salary disclosure 2022: all sectors and seconded employees." Published: Mar. 24, 2023. <https://www.ontario.ca/public-sector-salary-disclosure/2022/all-sectors-and-seconded-employees/>. Data downloaded on Apr. 29, 2023.
- Government of Ontario. 2021. "Public sector salary disclosure 2021: all sectors and seconded employees." Published: Mar. 25, 2022. <https://www.ontario.ca/public-sector-salary-disclosure/2021/all-sectors-and-seconded-employees/>. Data downloaded on Jul. 9, 2022.
- Government of Ontario. 2020. "Public sector salary disclosure 2020: all sectors and seconded employees." Published: Mar. 19, 2021. <https://www.ontario.ca/page/public-sector-salary-disclosure-2020>. Data downloaded on Jun. 18, 2021.
- Government of Ontario. 2019. "Public sector salary disclosure 2019: all sectors and seconded employees." Published: Mar. 20, 2020. Updated: Dec. 22, 2020. <https://www.ontario.ca/>

[page/public-sector-salary-disclosure-2019-all-sectors-and-seconded-employees.](#)

Data downloaded on Jun. 18, 2021.

- Government of Ontario. 2017. “Public sector salary disclosure 2016: all sectors and seconded employees.” Published: Mar. 31, 2017. Updated: Jun. 7, 2017. <https://www.ontario.ca/page/public-sector-salary-disclosure-2016-all-sectors-and-seconded-employees>.

Data downloaded on Jun. 30, 2017.

- Government of Ontario. 2016. “Public sector salary disclosure 2015: all sectors and seconded employees.” Published: Mar. 24, 2016. Updated: Dec. 20, 2016. <https://www.ontario.ca/page/public-sector-salary-disclosure-2015-all-sectors-and-seconded-employees>.

Data downloaded on May 25, 2017.

- Government of Ontario. 2015. “Public Sector Salary Disclosure Act: Disclosures for 2014.” Published: Mar. 24, 2016. Updated: Aug. 23, 2016. <https://www.ontario.ca/page/public-sector-salary-dis>

Data downloaded on May 25, 2017.

Heggeness, Misty, Donna Ginther, Maria Larenas, and Frances Carter-Johnson. 2018. “The Impact of Postdoctoral Fellowships on a Future Independent Career in Federally Funded Biomedical Research.” *NBER Working Paper*, April 2018: 1-44. <http://www.nber.org/papers/w24508>

Helliwell, John F., Richard Layard, Jeffrey D. Sachs, Jan-Emmanuel De Neve, Lara B. Aknin, and Shun Wang. 2021. *World Happiness Report 2021*. New York: Sustainable Development Solutions Network. <https://worldhappiness.report/ed/2021/>

Helliwell, John F., Richard Layard, Jeffrey D. Sachs, Jan-Emmanuel De Neve, Lara B. Aknin, and Shun Wang. 2022. *World Happiness Report 2022*. New York: Sustainable Development Solutions Network. <https://worldhappiness.report/ed/2022/>

Helliwell, John F., Richard Layard, Jeffrey D. Sachs, Lara B. Aknin, Jan-Emmanuel De Neve, and Shun Wang. 2023. *World Happiness Report 2023*. New York: Sustainable Development Solutions Network. <https://worldhappiness.report/ed/2023/>

Helliwell, John F., Richard Layard, Jeffrey D. Sachs, Jan-Emmanuel De Neve, Lara B. Aknin, and Shun Wang. 2024. *World Happiness Report 2024*. University of Oxford: Wellbeing Research Centre. <https://worldhappiness.report/ed/2024/>

International Ski Federation (FIS) (online). “Alpine Skiing: Search Results” page <https://data.fis-ski.com/alpine-skiing/results.html> specifying Sector as “Alpine Skiing,” Gender as “L,” Category as “Olympic Winter Games,” and Discipline as “DH.”

- Beijing, China, 2022 Olympic Winter Games. “Results, Analysis, Standings.” <https://medias3.fis-ski.com/pdf/2022/AL/5204/2022AL5204.pdf>. Contains reports titled “Entry List by Event,” “Draw List,” “Start List,” “Results,” “Penalty Calculation,” “Overall FIS WCSL List,” “Performance Analysis by Rank,” “Performance Analysis by Bib,” “FIS WCSL List,” “Medallists,” and “Medallists by Event.” Retrieved February 22, 2022.

- PyeongChang, South Korea, 2018 Olympic Winter Games. “Results, Analysis, Standings.” <http://medias1.fis-ski.com/pdf/2018/AL/5216/2018AL5216.pdf>. Contains reports titled “Entry List by Event,” “Draw List,” “Start List,” “Official Results,” “Penalty Calculation,” “FIS WCSL List,” “Overall FIS WCSL List,” “Performance Analysis by Rank,” “Performance Analysis by Bib,” “Medallists,” and “Medallists by Event.” Retrieved May 28, 2018.
- Sochi, Russia, 2014 Olympic Winter Games. “Results, Analysis, Standings.” <http://medias4.fis-ski.com/pdf/2014/AL/5326/2014AL5326.pdf>. Contains reports titled “Entry List by Event,” “Draw List,” “Start List,” “Official Results,” “Penalty Calculation,” “Performance Analysis by Rank,” “Performance Analysis by Bib,” “FIS WCSL List,” “Overall FIS WCSL List,” “Medallists,” “Medal Standings,” and “Medallists by Event.” Retrieved May 28, 2018.
- Vancouver, Canada, 2010 Olympic Winter Games. “Results, Analysis, Standings.” <http://medias3.fis-ski.com/pdf/2010/AL/5409/2010AL5409.pdf>. Contains reports titled “Draw List,” “Start List,” “Official Results,” “Penalty Calculation,” “Performance Analysis by Rank,” “Performance Analysis by Bib,” “FIS WCSL List,” “Overall FIS WCSL List,” “Medallists,” “Medal Standings,” and “Medallists by Event.” Retrieved May 28, 2018.
- Salt Lake City, United States, 2002 Olympic Winter Games. “Results, Analysis, Standings.” <http://medias4.fis-ski.com/pdf/2002/AL/6413/6413.pdf>. Contains reports titled “List of Competitors by FIS Points,” “Entry List by NOC,” “Start List,” “Official Results,” “Penalty Calculation,” “Performance Analysis by Rank,” “Performance Analysis by Bib,” “Downhill FIS WCSL List,” “Overall FIS WCSL List.” Retrieved May 28, 2018.

Johnson, Roger W. 1996. “Fitting Percentage of Body Fat to Simple Body Measurements.” *Journal of Statistics Education*, 4(1). <https://ww2.amstat.org/publications/jse/v4n1/datasets.johnson.html>

Karlan, Dean, and John A. List. 2007. “Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment.” *American Economic Review*, 97(5): 1774-1793. DOI: 10.1257/aer.97.5.1774. <https://www.aeaweb.org/articles?id=10.1257/aer.97.5.1774>

Levinson, Arik. 2016. “How Much Energy Do Building Energy Codes Save? Evidence from California Houses.” *American Economic Review*, 106(10): 2867-2894. DOI: 10.1257/aer.20150102. <https://www.aeaweb.org/articles?id=10.1257/aer.20150102>

- “Online Appendix” pages 1-15 (same link as above)
- Levinson, Arik. 2014. “How Much Energy Do Building Energy Codes Really Save? Evidence from California” *NBER Working Paper*, December 2014: 1-40. <http://www.nber.org/papers/w20797.pdf> (NOTE: These preliminary results are not identical but are substantively the same.)

Levitt, Steven D., John A. List, and Chad Syverson. 2013. “Toward an Understanding of Learning by Doing: Evidence from an Automobile Assembly Plant.” *Journal of Political Economy*, 121(4): 643-681. DOI: 10.1086/671137. <http://www.journals.uchicago.edu/doi/abs/10.1086/671137>

Organisation for Economic Co-operation and Development. “OECD Data.” <https://data.oecd.org/>

- “Air and GHG emissions.” DOI: 10.1787/93d10cf7-en. Retrieved from <https://data.oecd.org/air/air-and-ghg-emissions.htm> on June 3, 2017.
- “Crude oil import prices.” DOI: 10.1787/9ee0e3ab-en. Retrieved from <https://data.oecd.org/energy/crude-oil-import-prices.htm#indicator-chart> on June 3, 2017.
- “Gross domestic product (GDP).” DOI: 10.1787/dc2f7aec-en. Retrieved from <https://data.oecd.org/gdp/gross-domestic-product-gdp.htm> on June 3, 2017.
- “Renewable energy.” DOI: 10.1787/aac7c3f1-en. Retrieved from <https://data.oecd.org/energy/renewable-energy.htm> on June 3, 2017.

Organisation for Economic Co-operation and Development (online). “Water: Freshwater abstractions.” *OECD Environment Statistics*. Retrieved from <http://dx.doi.org/10.1787/data-00602-en> on May 31, 2018.

Picker, Les. 2015. “Digest: Asiaphoria Meets Regression to the Mean.” *The NBER Digest*, March 2015: 1-2. <http://www.nber.org/digest/mar15/mar15.pdf>

Pritchett, Lant, and Lawrence H. Summers. 2014. “Asiaphoria Meets Regression to the Mean.” *NBER Working Paper*, October 2014: 1-61. <http://www.nber.org/papers/w20573>

Statistics Canada (online). 2010. “Human Activity and the Environment: Freshwater supply and demand in Canada.” “Section 3: The demand for water in Canada.” Retrieved from <https://www.statcan.gc.ca/pub/16-201-x/2010000/part-partie3-eng.htm> on May 31, 2018.

U.S. Board of Governors of the Federal Reserve System. “China / U.S. Foreign Exchange Rate [AEXCHUS].” Retrieved on July 17, 2017 from FRED, Federal Reserve Bank of St. Louis. <https://fred.stlouisfed.org/series/AEXCHUS>.

U.S. Department of Energy. 2017. “Fuel Economy Guide: 2017 Datafile.” Retrieved from <https://www.fueleconomy.gov/feg/download.shtml> on Jun. 9, 2017.

World Health Organization (WHO) (online). “WHO Global Urban Ambient Air Pollution Database (update 2016).” Retrieved from http://www.who.int/phe/health_topics/outdoorair/databases/cities/en/ on July 17, 2017.

Zheng, Siqi, and Matthew E. Kahn. 2017. “A New Era of Pollution Progress in Urban China?” *Journal of Economic Perspectives*, 31(1): 71-92. DOI: 10.1257/jep.31.1.71. <https://www.aeaweb.org/articles?id=10.1257/jep.31.1.71>