The Data Analysis Course Module (DACM) Handbook* for ECO220Y1Y

*For easy navigation, **download the free pdf version**. Viewing in a browser is suboptimal.

Summer 2018: May - August 2018

Written by Jennifer Murdock[a,1]

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 full color figures in the popular press) you will *do* a lot of statistics and econometrics and deal with a lot of real data. DACM is an *immersive experience*: there is flood of data and research that create many opportunities for applying a host of ECO220Y course concepts. Next is a list of learning objectives.

1. Apply the methods you learn in ECO220Y "Quantitative Methods in Economics" to answer a variety of interesting research questions using real data
2. Build confidence in critically reading and assessing empirical research
3. Become data literate: understand how data are organized, documented, and readied for analysis and identify reputable sources
4. Become familiar with some major databases and journals that publish replication data
5. Link research questions with appropriate data and methods
6. Replicate tables and figures, presented in published research, using the original data
7. Understand why researchers sometimes complicate analyses (e.g. transforming or adjusting variables, conditioning on variables) by trying it multiple ways, including the simplest approach
8. Analyze subsets of data (e.g. developed and developing countries, males and females)
9. Fluently employ software (Excel) to summarize variables, describe relationships, and draw statistical inferences
10. Develop excellent habits with respect to precision, documentation, and error-checking
11. Use simulation methods to help with tricky conceptual questions
12. 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

[a]Associate Professor, Teaching Stream, Department of Economics, University of Toronto

[1]Thomas Russell, PhD Student, Department of Economics, University of Toronto, provided invaluable insights, suggestions, and feedback throughout the many stages of drafting this handbook and organizing the supporting data.

Contents

2 The Data Analysis Course Module (DACM) Syllabus

Course: All sections of ECO220Y1Y, Summer 2018, University of Toronto, Dept. of Economics
DACM course site: https://portal.utoronto.ca/
Piazza: https://piazza.com/utoronto.ca/summer2018/eco220y1y (sign-up link)
ECO220Y1Y Professors: Prof. Yu (Primary), Prof. Murdock (DACM only)
DACM Head TA: Thomas Russell (thomas.russell@mail.utoronto.ca)

The Data Analysis Course Module (DACM) is required for all students in ECO220Y1Y: all sections, all instructors. Overall, it counts for 15 percent of your ECO220Y1Y course grade and includes tutorials and online tests. DACM also provides crucial training in course concepts and exposure to case studies. Section 1 explains the goals and learning objectives for DACM. Over the Summer 2018 term, each student attends twelve interactive DACM tutorials and completes five online tests for DACM. These are grouped into five modules: Modules A through E. Each module contains either two or three submodules (e.g. A.1 and A.2) and practice test questions. Each module finishes with an online test. This handbook provides a complete plan for all modules.

2.1 DACM Tutorials: Prepare & Bring Laptops

You attend twelve interactive tutorials spaced from early May through mid August. Each tutorial is one hour long – some are back-to-back for a total of two hours – and is held in a large lecture hall. These are led by a TA who illustrates the steps in Excel on the data projector at the front of the room. You are meant to work along on your own laptop or to share a laptop with a classmate. Also, one or two helper TAs will circulate in the room to help you if you get stuck.

The content of tutorials is *not a surprise*: details for each submodule are listed under "Interactive tutorial materials." Each submodule also includes required background readings, which can involve studying figures and tables from published research. ***Complete the required readings and background BEFORE the tutorial.*** In fact, it is strongly recommended that you read through the entire submodule before the tutorial. Also, each submodule clearly lists the datasets you will need: ***download the data BEFORE the tutorial***. Our lecture halls do not typically have sufficient wireless capacity to accommodate a large group of people simultaneously downloading big data files.

Section 2.3 provides the complete schedule for all DACM tutorials and your specific tutorial assignment. You are generally expected to attend your assigned section. However, so long as we do not encounter space constraints, you *may* attend with another section.

After Module A, the tutorials are more challenging from a data analysis perspective. Module A provides the most detail and slowest pace as people get used to interactive tutorials. That said, you may find Module A intense if it is your first experience with the basics of using Excel. Also, Module A has important material and is neither short nor easy. You may find Module B the most challenging and the most work.

2.1.1 Pacing of Tutorials, Using Piazza & the DACM Head TA

Given the group setting of the tutorials combined with interactive materials, some will find the pace too fast and others will find it too slow.

What if I cannot keep up with the pace of tutorials? If you find the tutorials too fast paced, one strategy is to prepare more. In addition to the required background reading, spend time reading through the tutorial itself ahead of time. Open and browse the associated data files to gain some familiarity. If you are generally keeping up but get stuck in the middle of a tutorial, raise your hand to flag one of the helper TAs. Also, if lots of people are stuck where you are, the lead TA can backtrack and walk the class through a challenging part again. However, the tutorial cannot stop until everyone is caught up, which would impose a big cost on the group. If you leave a tutorial confused or lost, you still have options: attend a TA Aid Centre, review this detailed handbook, ask a classmate for help, post on Piazza. You could even attend the tutorial again with another section.

What if I find the pace of tutorials too slow? While we worked to include challenging and interesting material in every tutorial, some people may find the pace of tutorials generally too slow. If you are in that spot, you can complete the ***for homework*** portions during tutorial. You can also explore replicating parts of tables and or figures you are not asked to do in tutorials (you will likely encounter those on practice tests questions anyway). You could also peer mentor other people around you in tutorial. Teaching is the best way to really learn something and is likely to benefit you even more than the people you are directly helping.

Use Piazza. We use Piazza (https://piazza.com/utoronto.ca/summer2018/eco220y1y/home) to facilitate communication outside of tutorials. The TAs and professors periodically check Piazza to ensure proper usage, flag some postings, and possibly answer some questions. However, ***Piazza's emphasis is on student-to-student Q&A***. Piazza is a complement to face-to-face interactions. ***Piazza is a substitute for e-mail.***

For private/confidential matters, our Head TA's e-mail is thomas.russell@mail.utoronto.ca. For e-mails asking for a reply, if Thomas can answer briefly *without* explaining course content or revealing something of general interest, then he will reply within three business days. ***For any question that would interest other people (e.g. a question about an upcoming test, a clarification about tutorial materials, how to get unstuck on a hard practice test question, etc.), you must post on Piazza if you are hoping for an electronic reply.***

2.2 DACM Marking Scheme

Overall, DACM counts for ***15 percent of your ECO220Y1Y course grade***. The total points available over the five online DACM tests (Modules A, B, C, D, and E) is 100. At the end, your overall DACM mark is the sum of the points you earned over the five tests. Note that each online test can be worth anywhere from 15 to 25 points.

Section 2.3 provides the complete schedule for all five online tests. As you see, for each test you have a window of time to complete it. ***There are no make-up tests and you cannot submit a test***

late: failing to submit by the due date results in a mark of zero for that test. Do not leave the test until the last day, giving yourself no time to deal with health, personal, family or technological challenges. Once the deadline passes, the only option is to work extra hard on other graded course work to pull up your grade and compensate for the missed test.
For accessibility concerns visit http://www.studentlife.utoronto.ca/as. We *can* provide accommodations for students registered with Accessibility Services (e.g. 135 minutes for a 90 minute test).

2.2.1 DACM Online Tests

To prepare for the tests, make sure that you can fluently replicate all steps from the interactive tutorials, have completed all items flagged for homework, and have solved all the practice test questions. You can expect test questions to be variations on questions in tutorials and practice tests. Also, DACM modules are cumulative: retain your fluency with methods first introduced in earlier modules.

The five online tests are completed on portal. Questions are planned to be short-answer. A common question format requires you to type an exact numeric answer. Make sure to avoid typing errors, improper rounding, or failing to type your answer in the requested format. You can expect to do substantial data analysis using software (Excel) to answer.

Prior to starting each test, make sure you have ready access to the data files and the course aid sheets (http://homes.chass.utoronto.ca/~murdockj/eco220/AS220_Summer_2018.pdf). You should also have your notes, textbook, and this handbook handy before starting. If you are well-prepared, a typical test should take about half of the allotted time to complete.

Reasonable collaboration on online tests is allowed. However, your questions will vary from other people so only real collaboration (not copying) is helpful.

Section 2.3 explains when each test opens, the time limit, and when it is due. ***Once you begin, you have a maximum of 90 minutes to finish.*** Start each test well before the deadline and when you have the uninterrupted time needed to finish it.

After the due date and the marking is complete, you can see the questions and your answers by going to MyGrades, clicking on the title of the test (which will open the assessment details in a new window), and then clicking on your grade (under Calculated Grade). For concerns about the marking of your online test, our Head TA's e-mail is thomas.russell@mail.utoronto.ca.

2.2.2 Keeping it Real: Marking of DACM Online Tests

A reality of data analysis (and coding more generally) is that one seemingly tiny mistake can make your whole answer wrong, and not necessarily in a tiny way. Undergraduates typically have little experience with this harsh reality. Instructors often give partial credit for doing some things right. While we also take that approach with most graded work in ECO220Y, it does not make sense for DACM. In data analysis, there is no such thing as 80 percent correct: either all steps are correct and the analysis is correct or it is not. Of course, we want to support you in your learning. Some DACM online test questions are deliberately designed to involve very few steps. However, we also include

more realistic questions that require a good number of properly executed steps to answer correctly. ***In all cases, only completely correct answers earn marks: you cannot submit a regrade request based on losing all marks for your rounding errors (or your other errors, no matter how seemingly minor) on DACM online tests.*** Fortunately, there are many questions over the five DACM online tests: making a mistake in one question does little harm to your mark. However, if you often make mistakes, these lost points can add up. An important learning objective in DACM is being able to analyze data accurately.

2.2.3 Am I responsible for this material beyond the DACM tests?

Yes. DACM is an integral part of ECO220Y1Y, not an appendage. While it directly counts for 15 percent of your course grade, it indirectly counts for more. By working through these case studies, interactive tutorials, practice questions, and DACM tests, you can expect to deepen your understanding of the course material and improve your performance on regular term tests, assignments, and the common final examination. You may even see some of the same case studies again.

2.3 Summer 2018 Calendar of DACM Events

Summer 2018 Calendar of DACM Events		
	Day of the Week, Date, Time, and Location	
Event	**LEC0101**	**LEC0201**
Module A:		
Module A.1 Tutorial	Thurs, May 10, 11:10-noon, MP 103	Thurs, May 10, 2:10-3pm, MP 102
Module A Aid Centre	Thurs, May 10, 12:10-1pm, MP 103	Thurs, May 10, 3:10-4pm, MP 102
Module A.2 Tutorial	Thurs, May 17, 11:10-noon, MP 103	Thurs, May 17, 2:10-3pm, MP 102
Module A Aid Centre	Thurs, May 17, 12:10-1pm, MP 103	Thurs, May 17, 3:10-4pm, MP 102
Module A 90-Minute Online TEST opens at 4pm Thurs, May 17. **Due by Tues, May 22 at 6pm.**		
Module B:		
Module B.1 Tutorial	Thurs, May 31, 11:10-noon, MP 103	Thurs, May 31, 2:10-3pm, MP 102
Module B Aid Centre	Thurs, May 31, 12:10-1pm, MP 103	Thurs, May 31, 3:10-4pm, MP 102
Module B.2 Tutorial	Thurs, June 7, 11:10-noon, MP 103	Thurs, June 7, 2:10-3pm, MP 102
Module B.3 Tutorial	Thurs, June 7, 12:10-1pm, MP 103	Thurs, June 7, 3:10-4pm, MP 102
Module B Aid Centre	Fri, June 8, 12:10-1pm, MP 103	Fri, June 8, 1:10-2pm, MP 102
Module B 90-Minute Online TEST opens at 2pm Fri, June 8. **Due by Tues, June 12 at 6pm.**		
Module C:		
Module C.1 Tutorial	Thurs, July 5, 11:10-noon, MP 103	Thurs, July 5, 2:10-3pm, MP 102
Module C.2 Tutorial	Thurs, July 5, 12:10-1pm, MP 103	Thurs, July 5, 3:10-4pm, MP 102
Module C Aid Centre	Fri, July 6, 12:10-1pm, MP 103	Fri, July 6, 1:10-2pm, MP 102
Module C 90-Minute Online TEST opens at 2pm Fri, July 6. **Due by Tues, July 10 at 6pm.**		
Module D:		
Module D.1 Tutorial	Thurs, July 26, 11:10-noon, MP 103	Thurs, July 26, 2:10-3pm, MP 102
Module D.2 Tutorial	Thurs, July 26, 12:10-1pm, MP 103	Thurs, July 26, 3:10-4pm, MP 102
Module D Aid Centre	Fri, July 27, 12:10-1pm, MP 103	Fri, July 27, 1:10-2pm, MP 102
Module D 90-Minute Online TEST opens at 2pm Fri, July 27. **Due by Tues, July 31 at 6pm.**		
Module E:		
Module E.1 Tutorial	Thurs, Aug 9, 11:10-noon, MP 103	Thurs, Aug 9, 2:10-3pm, MP 102
Module E.2 Tutorial	Thurs, Aug 9, 12:10-1pm, MP 103	Thurs, Aug 9, 3:10-4pm, MP 102
Module E.3 Tutorial	Fri, Aug 10, 11:10-noon, MP 103	Fri, Aug 10, 1:10-2pm, MP 102
Module E Aid Centre	Fri, Aug 10, 12:10-1pm, MP 103	Fri, Aug 10, 2:10-3pm, MP 102
Module E 90-Minute Online TEST opens at 3pm Fri, Aug 10. **Due by Tues, Aug 14 at 6pm.**		

3 Getting Started with Excel, Finding the Data & Lynda.com

This handbook directly supports the use of **Microsoft Excel 2016** and the Analysis ToolPak add-in. Current U of T students can install Office 365, which includes Microsoft Excel 2016, for free. Carefully review the details at the page "Student Advantage and Office 365 ProPlus": http://help.ic.utoronto.ca/content/3/1965/en/student-advantage-and-office-365-proplus.html. Notice the separate links depending on your device (e.g. windows, mac) and that you must **first uninstall older versions of Office**. After successfully installing Microsoft Excel 2016, add the Analysis ToolPak add-in (which Microsoft does not automatically put in as many people use Excel for more basic functions). To start, open Excel and then follow these instructions: https://support.office.com/en-US/article/Load-the-Analysis-ToolPak-6a63e598-cd6d-42e3-9317-6b40ba1a66b4. The Analysis ToolPak add-in is required to use Excel for certain types of statistical analyses (e.g. multiple regression) that are important in DACM.

This handbook is accompanied by a large number of data files that you must download and analyze to complete interactive tutorials, practice tests, and online tests. The **data files are located on the DACM portal site** and are in Microsoft Excel 2016 format. (As a backup, the data files are also on a shared Google folder: `https://drive.google.com/open?id=0B3u6ZB1vQPGjODlYT1lQTnU3Tms`).

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

- **EXCEL TIPS** pepper this handbook.
- TAs will demonstrate how to use Excel in tutorials, as you follow along in real time.
- Our course textbook – "Business Statistics" by Sharpe et al. 2017, 3rd Canadian Edition – includes a "Technology Help" section at the end of each chapter. It has Excel support, with screen shots. It also supports some other software packages (Minitab, SPSS, and JMP).
- Also, some prominent universities maintain public help pages for analyzing data with Excel (and other software): for example, Princeton (`http://www.princeton.edu/~otorres/Excel/`).

Starting in April 2018, all U of T students have *free* access to **Lynda.com**, which offers high-quality instructional videos: http://main.its.utoronto.ca/news/free-access-to-lynda-com-online-courses/. I have browsed a variety of videos from basic uses of Excel to more advanced topics (multiple regression and simulation) and they are expertly done. We do not require that you watch these videos or use Lynda.com – everything you need to know is covered in this handbook and in the DACM tutorials – but Lynda.com is definitely a good supplemental source for you, especially if you like video instruction.

For those interested in more powerful software, the textbook offers some choices. Stata (not mentioned in our textbook) is my favorite. You will see Stata histograms throughout this handbook. (In fact, all analyses were done in Stata and then redone in Excel.) However, if you choose not to use Microsoft Excel 2016, you accept full responsibility for figuring out how to complete all analyses.

Regardless of what you use, ***you have significant responsibility for ensuring: (1) that you have access to an appropriately set up computer that allows you to analyze the data provided in DACM and (2) that you know how do properly complete the analyses with***

the software you are using. Figuring out how to get software to do what we want is an important real-life skill and one you are expected to develop via practice.

3.1 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). As you will see, even with just Excel we can replicate cutting-edge research and results. It is certainly good enough for ECO220Y1Y despite its limitations. However, there are a ***few key limitations of Excel*** worth highlighting:

- Excel is used interactively as opposed to writing code and running it. This can be very dangerous. With a set of code you can easily fix errors and re-run the whole thing. Also, with code you have documentation of exactly how you got from the raw data to the final result. When working interactively, you must be very careful and take extra steps to document your work.
- Excel is clunky to use for multiple regression (but it is doable) and it cannot do much 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 functions return an error when applied to a missing value. Other functions return a zero. Neither of these responses to a missing value 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. **You *really* have to be careful about handling missing values in Excel.** The cases when a zero is returned instead of a missing value are particularly dangerous: zero is a real number and hence subsequent functions will treat any zeros as zeros (not missing).

To illustrate, look at Figure 1, which shows data containing 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. When you create a new variable measuring salary in $1,000s of dollars (salary_1000), Excel records a zero for Xiadong and Marcus. If you computed the mean of Column D, it would be wildly incorrect, yet you would get no error message. This is the most dangerous situation illustrated. The natural log function returns an error, which is the least dangerous situation. Column F shows what happens if you try to compute the salary per hour: it shows how Excel can respond 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: Illustrative examples of issues with missing values. The left side shows what Excel functions can return instead of a missing value: 0, #NUM!, or #DIV/0!. The right side shows the functions behind each cell. (Note: To see the formulas, you can use the shortcut **Ctrl +** `` ` `` or the FORMULATEXT() function.)

There are workarounds but they are not fun. For functions that return some sort of error when encountering a missing value, a workaround is offered on page 23 in step 1j in Module A.2. For the more dangerous situation where the function returns a numeric value (zero), the workaround is even more crude. 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 and your subsequent analysis will be correct.

Remember to be hyper-careful in Excel when a variable contains missing values and you wish to apply any functions to it to create new variables. Unfortunately, this situation arises frequently in real life. Hence, you will need to use workarounds.

A Module A: Interactive Tutorial Materials & Test Prep

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

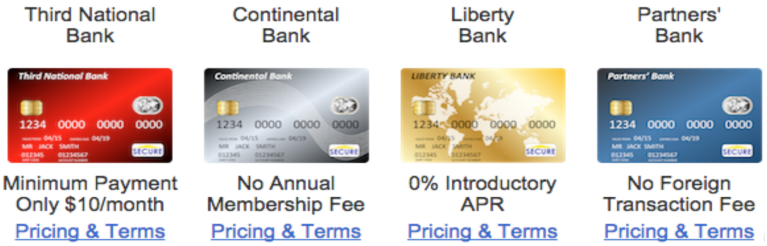
Main course concepts: Recognizing variable types: quantitative (interval) and categorical (nominal). Recognizing data structures: cross-sectional, times series, and panel. Using tabulations and cross tabulations to analyze categorical (nominal) data. Working with academic research papers.

Source materials (full citations in Section F): We replicate parts of the analysis from an academic journal article "Millennial-Style Learning: Search Intensity, Decision Making, and Information Sharing," abbreviated Carlin et al. (2017). We survey types of data from a variety of sources (City of Toronto, Google Finance, and the OECD).

Most relevant required readings: Chapters 2 and 4 including "Technology Help: Displaying Categorical Data on the Computer" on pp. 74-76. Background reading for Carlin et al. (2017):

- In an online experiment, 1,603 respondents from Amazon's Mechanical Turk (https://www.mturk.com/mturk/welcome) watched a short video and then chose among four credit cards offers:

> 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 is accompanied by 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. There are two video versions: "baseline" and "implemental." The **baseline video** is a humorous cartoon explaining what to watch out for when choosing a credit card. The **implemental video** adds a recap of the key takeaways to the baseline video.
  2. The graphic above illustrates **superfluous taglines**, e.g. "No Annual Membership Fee." Other respondents have the same choice but with **no taglines**. These taglines are superfluous because they do *not* help you choose among the four cards. Continuing the example, none of the four cards had an annual membership fee.

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

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

Textbook case studies (extra practice): "Loblaws" on pp. 73-74; "KEEN Footwear" on p. 74

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 tutorial materials:

1. Consider Carlin et al. (2017) and ***open*** the file cred_card.xlsx.

   (a) ***Browse*** the data in worksheet cred_card and the data description in worksheet readme.

   (b) Which kind of data are these? ***Verify*** that these are cross-sectional data with 1,603 observations and 43 variables. The unit of observation is a person (respondent).

   EXCEL TIPS: For large data sets, scrolling up-and-down and side-to-side is not efficient. Jump to the last cell in a column with a non-missing value with the shortcut **Ctrl** + ↓. (When counting observations, remember n is the row number of the last observation minus one (row one is the variable name).) 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.

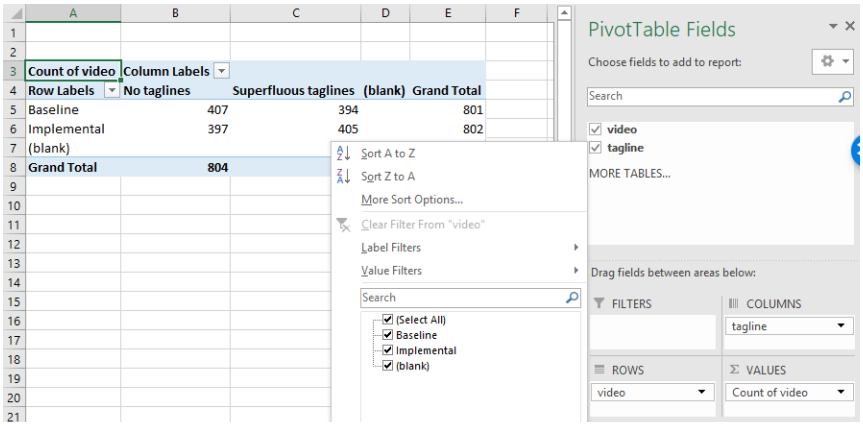
   (c) Which kind of variables are in these data? ***Verify*** there is one identifier variable (resp_id) and that most variables are nominal (categorical). ***Note*** that choicetime, choiceclicks, numcards, and age are clearly interval (quantitative) variables. Some are in a gray area: starttime, endtime, and the ten variables measuring opinions on a 1 to 7 Likert scale (e.g. confidence and lik_share_fam). However, we often treat Likert scale variables as interval. (For example, U of T summarizes course evaluations with the mean Likert scale response.)

   (d) 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 that are coded to take a value of one if yes and zero if no.

   EXCEL TIPS: 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). To ensure that you get all observations, select the entire column. Put this in a new worksheet that references the original data. For example: `=AVERAGE(cred_card!D:D)`. Note: When naming worksheets, avoid special characters (such as !, $, %, quotes, or spaces) and make sure the first and last characters are a letter (a - Z) and *not* a number (those can cause cryptic error messages later on).

(e) ***Replicate*** Table A.1 (summarizing the experimental design). This type of table is called 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 to create a pivot table in a new worksheet. In the PivotTable Fields area: check the boxes for both variables. Drag (if necessary) 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 (if necessary). (Dragging variables to an area automatically checks the boxes, so you could skip checking the boxes.) Clicking on a field under PivotTable Fields yields a drop-down menu where you can uncheck "(blank)" to clean up the presentation of the table when there are no blanks. (If you select only the rows with data, not the entire columns, you can avoid this step.)



(f) Figure 6 shows some key results. ***Replicate*** the values and ***report*** the exact height of each grey bar. (You should obtain 0.425061, 0.649874, 0.368020, and 0.511111, respectively.)

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

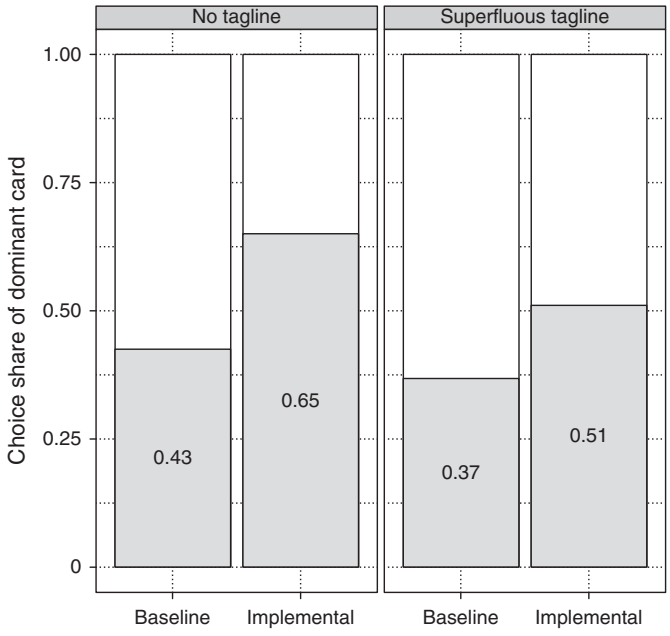
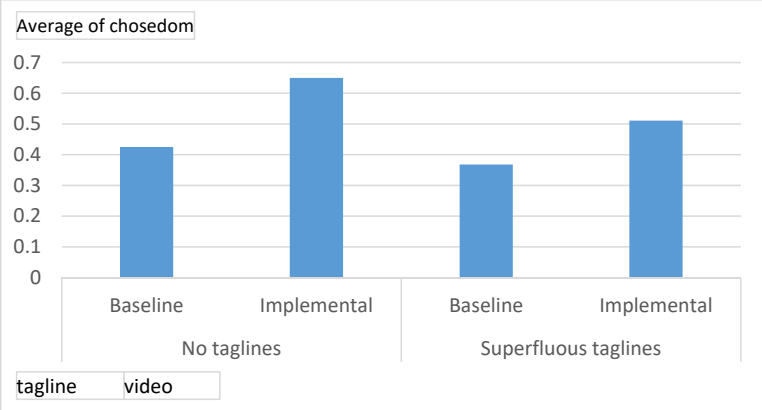


Figure 6: Carlin et al. (2017), p. 13.

EXCEL TIPS: Use a pivot table. Select three variables: video, tagline and chosedom. Insert a pivot table in a new worksheet. Check the boxes for all three variables. Drag (if necessary) tagline and video to the ROWS area, with tagline first; drag (if necessary) 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. (The mean of a 0/1 variable is the *share* of 1's.)

Row Labels	Average of chosedom
No taglines	**0.536069652**
Baseline	0.425061425
Implemental	0.649874055
Superfluous taglines	**0.440550688**
Baseline	0.368020305
Implemental	0.511111111
Grand Total	**0.488459139**

EXCEL TIPS (PC ONLY): Select three variables: video, tagline and chosedom. Click the PivotChart button (Insert tab with other charts) and select PivotChart & PivotTable in a new worksheet. Check the boxes for all three variables. Drag (if necessary) tagline and video to the AXIS (CATEGORIES) area. Drag chosedom to the Σ VALUES area and using the drop down menu, select Value Field Settings, and choose Average.



(g) ***Review*** the first column of results in Table 6 on page 15. ***Replicate*** the first four rows of the "Chosen (%)" column. Also, ***report*** the exact percentages.

EXCEL TIPS: Select the variable chosen_terms and insert a pivot table in a new worksheet. Drag chosen_terms to the ROWS area and drag another copy of chosen_terms to the Σ VALUES area. Select count. In the pivot chart itself, you can click on values of row labels and move them up or down to match the order in Table 6. One way to obtain percentages is to copy-and-paste-special (values only) 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 behind a cell. For example, cell E2 below contains `=FORMULATEXT(D2)`.

	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

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.

(h) ***Note*** that Columns A and B in the Excel output in part 1g 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.

(i) (OPTIONAL) Next, ***explore*** how Excel can help you do the replication in 1g with less work (but more skill).

EXCEL TIPS: Select the variable chosen_terms and insert a pivot table in a new worksheet. In the pivot table fields environment, drag chosen_terms to ROWS and drag a copy of chosen_terms to Σ VALUES. Select count. In the drop-down menu for chosen_terms in the sum values area, select Value Fields Settings, click the Show Values As tab, and select % of Column Total. It formats the proportions (values from 0 to 1) as percents (values from 0 to 100). BE CAREFUL if you plan to use these values in any subsequent calculations: this formatting hides the real number in the cell.

2. Tutorial time permitting (otherwise, for homework), ***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*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.

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

Main course concepts: Using histograms and descriptive statistics (e.g. mean, median, s.d.) to summarize quantitative (interval) variables. Summarizing subsets of data.

Source materials (full citations in Section F): 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. To enable 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).

Most relevant required readings: Chapter 5 including "Displaying and Summarizing Quantitative Variables" on pp. 127-8. Background reading on measuring air pollution:

- To measure air pollution, Zheng and Kahn (2017) use PM10, which is the concentration of particulate matter (i.e. small particles such as dust) with a diameter less than or equal to 10 micrometers, abbreviated as μm where 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 less than or equal to 2.5 micrometers. The units of measurement of PM10 or PM2.5 are micrograms per cubic meter air, which is abbreviated as $\mu g/m^3$.

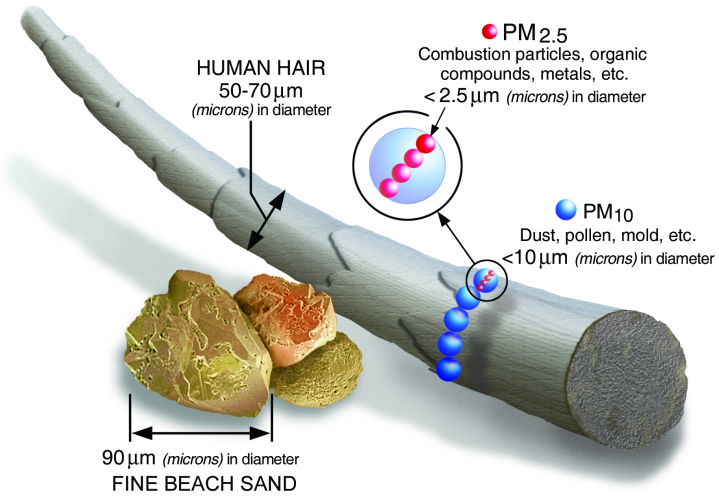


Figure 1: U.S. EPA (online), "Particulate Matter (PM) Basics", retrieved July 17, 2017.

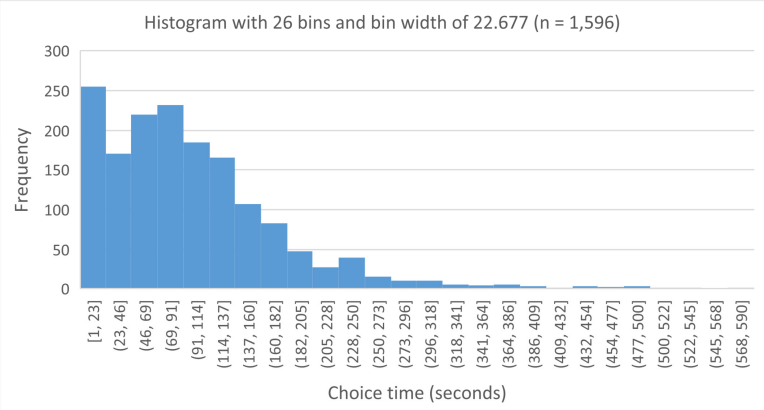
- To give some sense of these, according to WHO (2016), the annual mean of PM10 in 2012/13 in: Beijing, China is 108 $\mu g/m^3$; Toronto, Canada is 14 $\mu g/m^3$; Los Angeles, U.S. is 20 $\mu g/m^3$; Rome, Italy is 28 $\mu g/m^3$; Tokyo, Japan is 28 $\mu g/m^3$; Delhi, India is 229 $\mu g/m^3$.
- To reinforce Chapter 5's explanation of histograms – a tool to visually summarize the distribution of an interval variable – consider that 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 clearly specify any special treatment of outliers), you *do* have some artistic license.

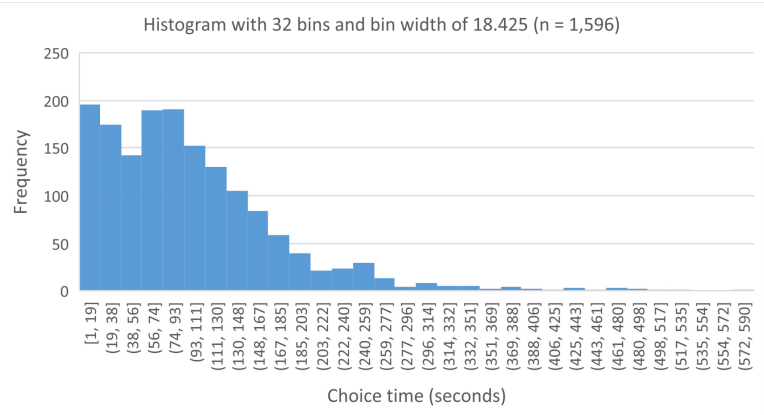
The goal is to *visually summarize* the distribution. Various software packages make different artistic decisions. Excel 2016 even makes different choices depending on which tool you use. This bullet list shows some popular formulas giving *suggestions* about the bins.

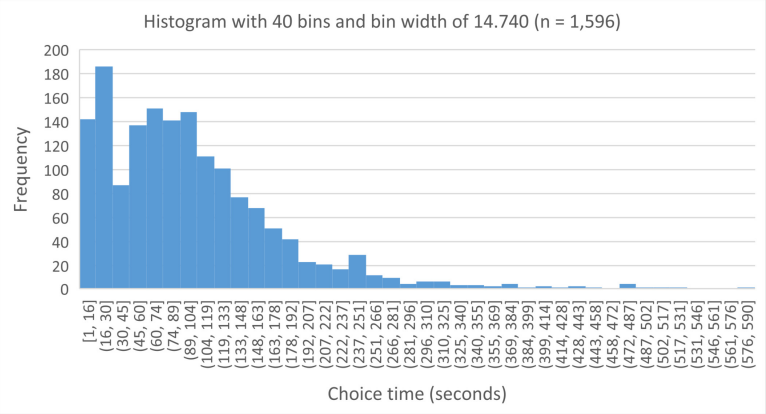
- Number of bins $= \sqrt{n}$. E.g., with 400 observations of a variable, choose ≈ 20 bins.
- Number of bins $= \frac{10ln(n)}{ln(10)}$. E.g., with 4,000 observations of a variable, choose ≈ 36 bins.
- Or, a combination. E.g., Stata uses $= MIN\left\{\sqrt{n}, \frac{10ln(n)}{ln(10)}\right\}$. (In practice, this means that you use $\sqrt{n}$ whenever the number of observations is less than 862.)
- Number of bins $= 1 + 3.3log_{10}(n)$. 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 standard deviation of 11.48, choose a bin width of ≈ 4.2.

Note: While all of these formulas suggest more bins (narrower bins) for bigger sample sizes, they make different suggestions. You may choose something a bit lower or higher, making a *subjective* judgment about how to most clearly summarize a particular distribution.

- To illustrate, recall cred_card.xlsx from Module A.1 and the variable choicetime, which records the time (in seconds) each participant spent choosing a card. 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 $\frac{3.5*77.4}{\sqrt[3]{1,596}} \approx 23$, $\frac{10ln(1,596)}{ln(10)} \approx 32$ bins, or $\sqrt{1,596} \approx 40$ bins. Remember, the number of bins must be an integer. Also, once you choose one of them (width or number), you have no choice about the other. In the case of the choicetime variable, all three choices give a similar looking histogram that shows a bi-modal and positively skewed distribution.







Additional readings (not required): Zheng and Kahn (2017) is published in the *Journal of Economic Perspectives*, which targets a general audience and is appropriate for undergraduates. While you do *not* need to read the article, if you are interested, you can understand much of it. (We cover multiple regression topics, needed to understand Table 1, at the end of our course.)

Textbook case studies (extra practice): "Solar Power in Ontario" on pp. 125-6; "Hotel Occupancy Rates" on p. 126; "Value and Growth Stock Returns" on p. 126

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).

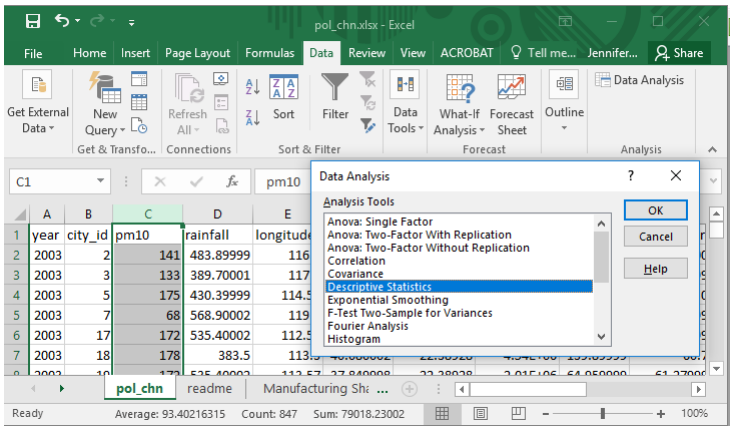
Interactive tutorial materials:

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

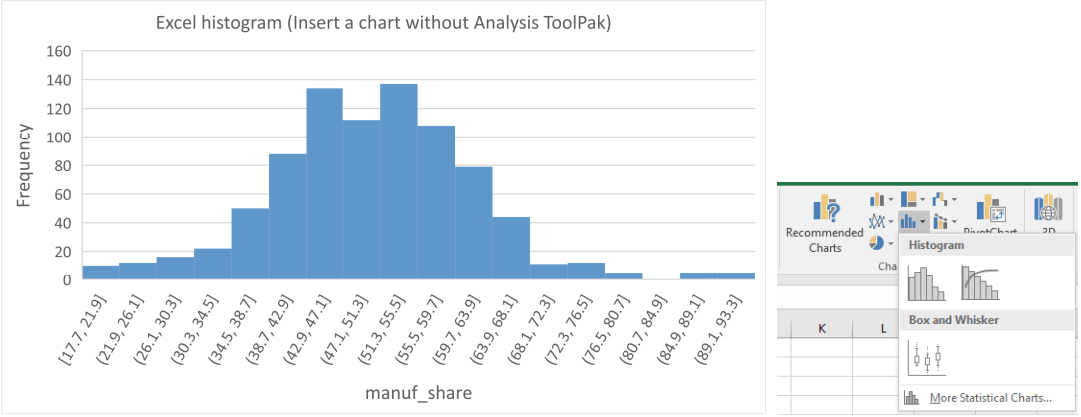
   (a) ***Browse*** the data in worksheet pol_chn and the data description in worksheet readme.

   (b) ***Compute*** the usual suite of descriptive statistics for the variable pm10. ***Verify*** that you obtain mean 93.4 $\mu g/m^3$, median 91.2 $\mu g/m^3$, s.d. 32.3 $\mu g/m^3$, and range 521 $\mu g/m^3$ for the 846 non-missing observations pm10.

   EXCEL TIPS: Find the Data Analysis button in the Data tab. (If it does not appear, you forgot to install the Data Analysis ToolPak add-in: see Section 3 on page 8.) Select Descriptive Statistics. Set the Input Range as the entire column ($C:$C) and check the boxes for Labels in First Row and for Summary statistics. Output to a new worksheet.



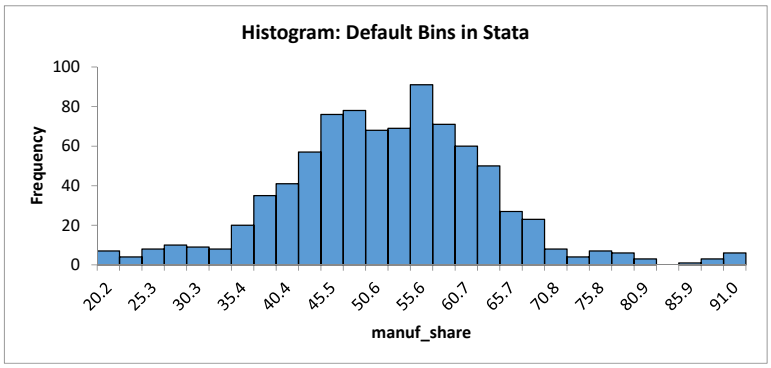
(c) ***Replicate*** the histogram below (not worrying about adding the titles).



EXCEL TIPS: In the 2016 version, Excel has recognized the utility of histograms and added them (and box plots) to the standard set of charts you can make without add-ins (like the Analysis ToolPak). Select the entire column for the manuf_share variable and, under the Insert tab, click the button with the miniature histogram (shown above). To be readable, format your bins on the x-axis (not too many decimal places). One way, which works for both pcs and macs, is to format the original manuf_share variable by selecting that entire column, Format Cells, select Number, and select 1 decimal place. You can do this after you have created the chart and the chart will automatically update.

EXCEL TIPS: If you draw a histogram using the standard chart tool in Excel 2016, by default, it chooses width of each bin $= \frac{3.5*sd}{\sqrt[3]{n}}$. Notice the width of the bins for the histogram in part 1c is 4.2: for example, $51.3 - 47.1 = 4.2$. The sample size is $n = 850$ and the standard deviation of manufacturing share is 11.48. Plugging in, we obtain $4.2 = \frac{3.5*11.48}{\sqrt[3]{850}}$.

(d) Now we explore how our choices of bins 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 to using the Data Analysis ToolPak to create histograms. If you use a pc (not a mac) you *can* use the histogram chart like in part 1c. However, the mac version does *not* allow you to change how the histogram is displayed (e.g. you cannot change the number or width of bins), which is important. The ToolPak version has flexibility for everyone. However, the histogram chart we learned about in part 1c *is* still useful for everyone for a quick and basic histogram.

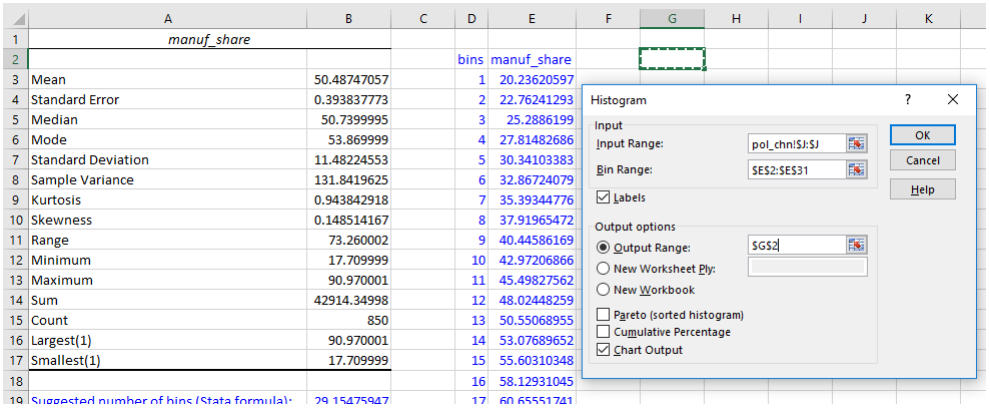
(e) Continuing, ***read*** the bullets below explaining the cells in blue:

i. 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: number of bins $= MIN\left\{\sqrt{n}, \frac{10ln(n)}{ln(10)}\right\}$.
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))`.

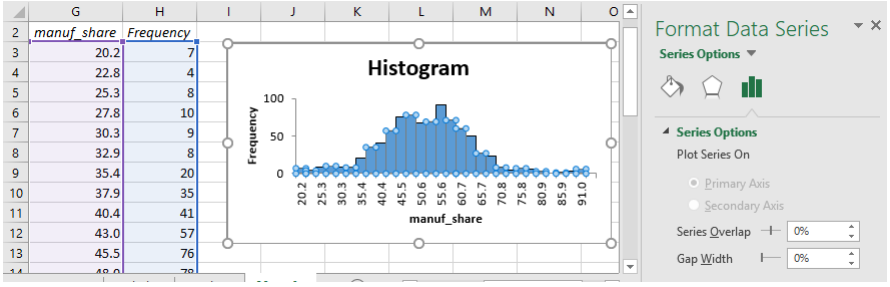
ii. Next, ***review*** the rounded suggested number of bins (must be an integer) and suggested width. ***Note*** that once you determine the number of bins, you can find the width of each bin by taking the range of your variable (which, recall, is the maximum value minus the minimum value) and dividing it by the number of bins.
EXCEL TIPS: Notice how the worksheet does this in cell `B21` with `=B11/B20`.

iii. 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.

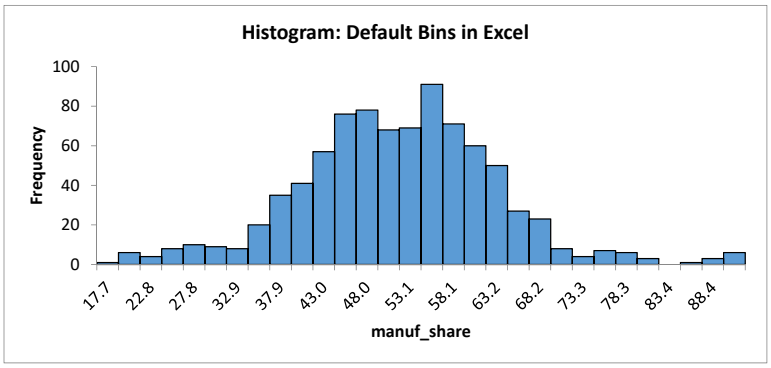
(f) ***Replicate*** the histogram in part 1d using the bins as defined in Stata.
EXCEL TIPS: Create a copy of the worksheet Manufacturing Share Histogram. Delete everything to the right of the two columns in blue that set up the right endpoints of each bin (i.e. *keep* the blue columns that define the bins). Click the Data Analysis button in the Data tab and select Histogram. Select the Input Range from the original data (`pol_chn!$J:$J`), select the Bin Range from your current worksheet (include the column label), check the Labels box, select `$G$2` as the Output Range (which makes the histogram appear in your current worksheet), and check the Chart Output box.



Next, fine-tune your histogram. Click the chart area to Format Data Series to set the Gap Width to zero. Select the numbers in the table produced with the histogram to round them. In this example, formatting to the first decimal place is a good choice.

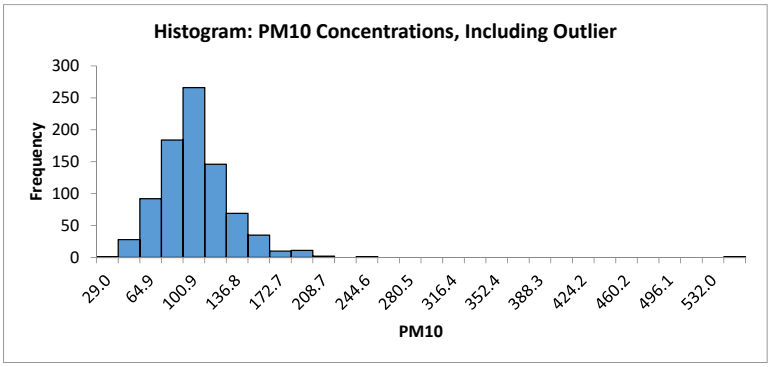


(g) ***Create*** a histogram of manuf_share using the default bins in the Data Analysis ToolPak for Excel: $\sqrt{n}+1$, which is one more than Stata in this example. ***Verify*** that your histogram, after fine-tuning, looks similar to the one below.



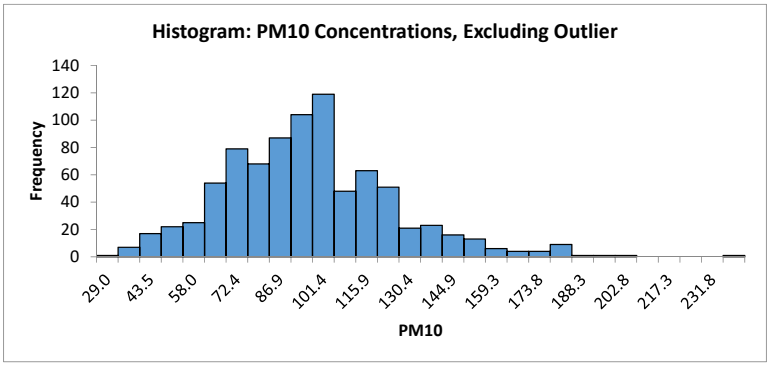
EXCEL TIPS: Use histogram under data analysis but leave the field for bin range blank. Also, Excel sets n to equal number of rows selected, ignoring missing values. Hence, ONLY select the range of values in the data (*not* `pol_chn!$J:$J`) or Excel will assume that your sample size is equal to the maximum number of rows permitted in worksheet. This would lead to a histogram with an unreasonably large number of very skinny bins.

(h) ***Create*** a histogram of PM10. ***Verify*** that your histogram looks similar to the one below.

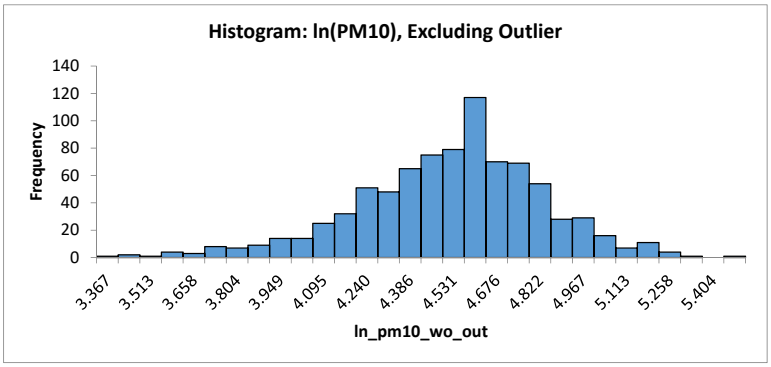


(i) ***Create*** a histogram of PM10 excluding the outlier with PM10 of 550 for the city with id 315 (Karamay) in 2003. ***Verify*** that your histogram looks similar to the one below.

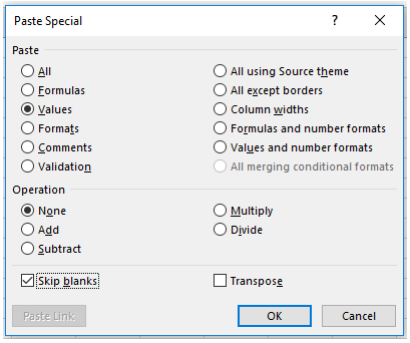
EXCEL TIPS: One way to remove the outlier is to copy the entire pm10 column and paste it right after the last variable in pol_chn. Rename it something like pm10_wo_out. Manually erase the outlier(s) in pm10_wo_out (i.e. clear that cell leaving a blank cell).



(j) ***Create*** a histogram of the natural log of PM10 (excluding the outlier). ***Watch out*** for the missing values and ***verify*** that your histogram looks similar to the one below.



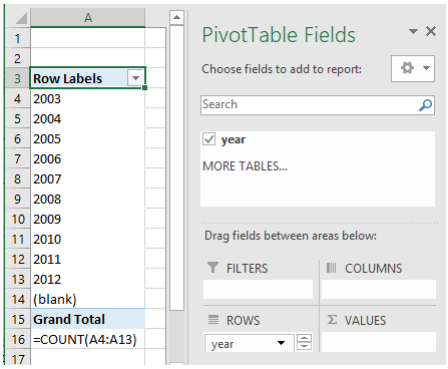
EXCEL TIPS (IMPORTANT!): Create a new column named ln_pm10_wo_out. Using `LN()` produces errors (`#NUM!`) for the five observations where PM10 is a missing value (the four originals and the cleared outlier). Unfortunately, Excel functions cannot return a truly blank cell. (In contrast, other software typically treats the natural log of a missing value as a missing value.) One option is to manually clear the cells containing `#NUM!`. Another option is to copy the entire ln_pm10_wo_out column to a new worksheet using Paste Special and choosing Paste Values. You can sort that new worksheet, which will put the offending cells at the end, and then select the rest of the values for the histogram.
EXCEL TIPS (PC ONLY): Another option is to use `=IF(ISERROR(LN(R2)),"",LN(R2))` where pm10_wo_out is in Column R. This tells Excel that if it encounters an error in computing the natural log it should record a blank in that cell. However, `""` is a character and you will get errors if you try to apply numeric functions (like descriptive statistics) to characters. Hence, you need to copy and Paste Special choosing Paste Values *and* checking the box the skip blanks to create another copy of that new column. This new column will *appear* to be identical, but, it is not because the copy contains no strings.



(k) Use pol_chn.xlsx to review the useful pivot tables from Module A.1.

i. How many different years and how many different cities are there in these data? ***Verify*** there are 10 unique years: 2003 through 2012. ***Verify*** there are 85 unique cities.

EXCEL TIPS: The most conceptually simple approach is to use two pivot tables. Insert a pivot table with one variable selected: year or city_id. By default it lists the unique values as row labels. Use the COUNT function, *not* including (blank), to count the row labels. It should return a value of 10.



ii. Tutorial time permitting (otherwise, for homework), across all 85 cities in China, what is the average pollution level and manufacturing share each year?
EXCEL TIPS: Use a pivot table. Insert a new worksheet. In the first three columns, copy-and-paste the variables year, pm10, and manuf_share. The shortcut **Ctrl + Click** lets you select more than one non-adjacent column for copying. (This gets the variables adjacent to each other without modifying the original data.) Insert a pivot table. 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 display the 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**

(l) ***For homework,*** consider the subtle differences across software packages. In this handbook you will see histograms produced by Stata (more powerful statistical software). Stata and Excel draw histograms differently. To illustrate, recall the Carlin et al. (2017) article and cred_card.xlsx from Module A.1. Consider 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).[2]

i. Review the three frequency histograms (next page), which use the exact same data. All three have a bin width of exactly 5.5. Replicate the first Excel histogram.

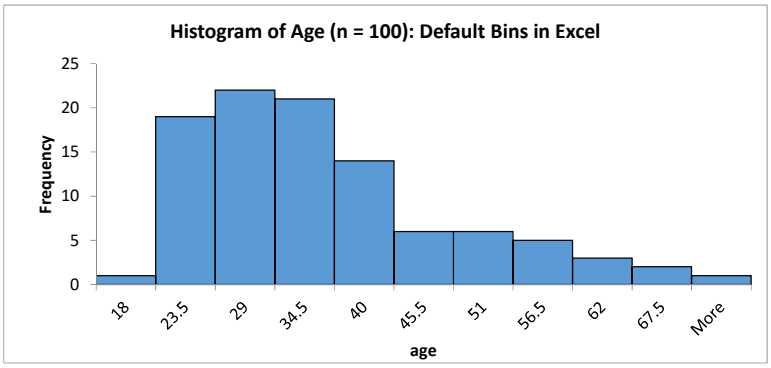
ii. Consider the question: if the data, bin widths, and type of histogram are exactly the same, why aren't these histograms identical?

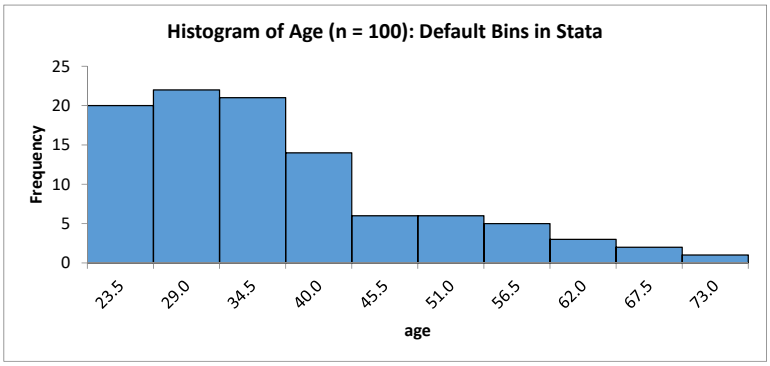
- When an observation is on the border between two bins, Excel and Stata make different choices, which explains the difference between the second and third histograms. Because age is an integer and the bin width is 5.5, this situation arises frequently. For example, in the first 100 observations, five people are 29 years old, which is on the boundary between two bins. Stata puts borderline cases in the bin to the right. Excel puts borderline cases in the bin to the left. In the bin from

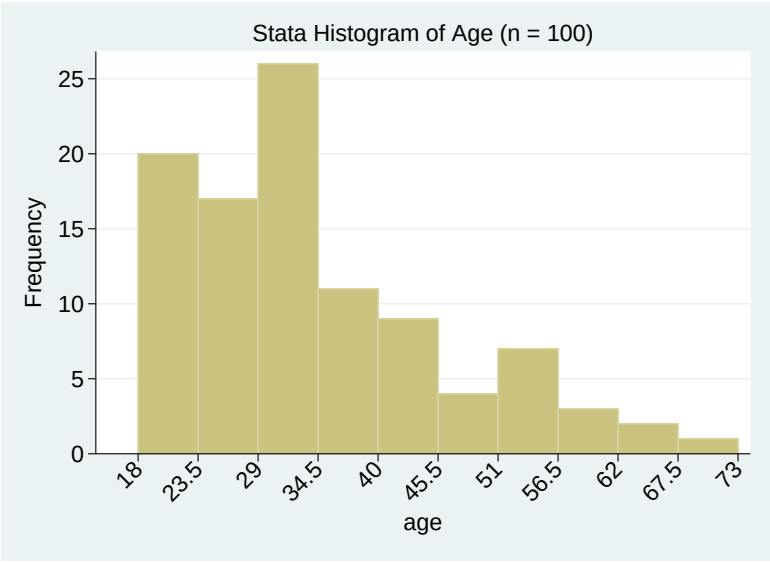
[2]Excel and Stata also differ in the rules for sorting. If you re-sort by resp_id in Excel, it will put the observations in a different order. For simple sorts, for example numeric data with no characters, Stata and Excel do the same thing.

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 includes an extra bin. Excel has the first bin end at 18. Because it puts borderline cases in the bin on the left, that bin contains the one person who is 18 years old in this sample (nobody is younger than 18). In contrast, the second Excel histogram uses the default bins in Stata and the first bin ends at 23.5 so it has the 20 people who are 18, 19, 20, 21, 22 or 23 years old.







iii. Consider the question: which histogram is correct?

- There is no such thing as *the* correct histogram (even though there are many possible wrong-headed ones). Histograms give an overall visual summary of the distribution of an interval variable. Remember to focus on the *overall* picture presented by a histogram. For example, it would be wrong to conclude that the

distribution is bi-modal given the Stata histogram: that is just a little blip. There is no systematic evidence supporting the inference that the age distribution is bi-modal. Slightly different, and still reasonable, choices when drawing the histogram makes that little blip disappear. In contrast, all three histograms clearly show that the age distribution is positively skewed and most respondents on Amazon's Mechanical Turk are quite young, which is not surprising as it is a fairly new online technology that many older people may not know about.

(m) ***For homework,*** 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.

A.3 Practice test questions for Module A

Q1. Recall the caution about missing values in Section 3.1 on page 9. In Module B, we will study a major database: Penn World Table (PWT), version 9.0. In preparing the data files to go with this handbook, original data from many sources were cleaned up to make them easier for you to understand and to work with. Some data required a lot of work. The PWT 9.0 data were excellently prepared and we really only needed to remove missing values and remove some variables we do not use. For this question use pwt90.xlsx, which are the data exactly as downloaded from https://www.rug.nl/ggdc/productivity/pwt/ (retrieved April 12, 2018).

(a) How many unique countries are in these data?

(b) How many unique years are in these data?

(c) Create a new variable measuring the natural logarithm of the variable rgdpe. What is the standard deviation of that new variable?

(d) 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.

(a) To asses how income varies among respondents, tabulate the variable hh_inc (i.e. create a frequency table). For each income category, include both a count of observations taking each unique value and the percent of observations taking each unique value.

(b) 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"?

Q3. Recalling Carlin et al. (2017), use cred_card.xlsx. Create a cross-tabulation (which is also called a contingency table) 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 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.

Q4. Recalling Carlin et al. (2017), use cred_card.xlsx. Do the results in Figure 6 vary by sex?

(a) One option would be to create Figure 6a using the subset of females and to create Figure 6b using the subset of males. You could compare these with each other. Or, you could make one new figure that also breaks things down by sex (Figure 6 already breaks things down by video and taglines). Create a **pivot table** like the pivot table supporting Figure 6 except for eight bars: the first four for females and the last four for males. Which best summarizes what your pivot table shows: "males have the opposite reaction to the implemental video compared to females," "females have the opposite reaction to the superfluous taglines compared to males," "males have the opposite reaction to both the implemental video and the superfluous taglines compared to females," or "males and females have similar reactions to the implemental video and superfluous taglines."

(b) Fill in the blanks with the appropriate 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.

Q5. Recalling Carlin et al. (2017), use cred_card.xlsx.

(a) Use Descriptive Statistics in the Data Analysis Toolpak to 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) Compute the percentiles to complete the table below. (You may use the original variable including the outlier as percentiles are not sensitive to outliers.)

Select Percentiles for Age

5th	
10th	
25th	
50th	
75th	
90th	
95th	

EXCEL TIPS: Use the function PERCENTILE.INC.

(d) 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.)

(e) Draw a histogram of age excluding the outlier. Describe the shape of the histogram.

(f) Draw a histogram of the natural log of age excluding the outlier. Describe the shape of the histogram.

(g) For age (with the outlier), copy the variable (age) to a fresh worksheet and SORT it in ascending order (which will allow you to avoid Excel's inability to reasonably handle missing values in data). Verify that you have 1,600 observations. Compute the sample mean using the formula $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$ using the SUM function in Excel. Compute the sample variance using the formula $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{n-1}$ by creating a column that is $(x_i - \bar{X})^2$ and using the SUM function. Before continuing, double-check that you get the exact same sample mean and sample s.d. as Excel's descriptive statistics. Using your work, fill in the blanks with the appropriate numbers: The outlier (the 200-year-old respondent) contributes _____ to the sum in the formula for $\bar{X}$ and _____ to the sum in the formula for s^2. A respondent who is 18 years old contributes _____ to the sum in the formula for $\bar{X}$ and _____ to the sum in the formula for s^2. A respondent who is 30 years old contributes _____ to the sum in the formula for $\bar{X}$ and _____ to the sum in the formula for s^2. Of these three, the respondent who is _____ years old [answer with 18, 30 or 200] is closest to the mean.

Q6. Open goog_s_p_tsx.xlsx on the S&P/TSX Composite Index for trading days from July 1, 2000 - June 30, 2017 by Google Finance (full citation in Section F). The index summarizes performance of a large number of Canadian stocks.

(a) Construct a time series plot of the S&P/TSX Composite Index for the period from April 1, 2003 through December 31, 2003. During this period, the index _____ [answer with: is trending up, is trending down, or shows no clear trend up or down].

EXCEL TIPS: Select the observations under the variables date and s_p_tsx_ind corresponding to the indicated time period. Insert a line chart.

(b) Construct a time series plot of the volume of shares traded for the period from April 1, 2003 through December 31, 2003. During this period, the volume of shares traded _____ [answer with: is trending up, is trending down, or shows no clear trend up or down].

(c) Create a new variable named lev_chg_ind: the trading-day-to-trading-day level change in the S&P/TSX Composite Index. (For example, if the index were 7,444 yesterday and is 7,472 today, the level change is positive 28.) During the period from April 1, 2003 through December 31, 2003, the worst trading day on the Canadian stock market (as measured by the level decline in the S&P/TSX Composite Index from the previous trading day) is: _____ (give month) _____ (give day), _____ (give year), when the index declined by _____ from the previous trading day.

(d) Create a new variable named pct_chg_ind: the trading-day-to-trading-day percent change in the S&P/TSX Composite Index. (For example, if the index were 7,444 yesterday and is 7,472 today, the percent change is positive 0.376 percent.) During the period from April 1, 2003 through December 31, 2003, the best trading day on the Canadian stock market (as measured by the percent increase in the S&P/TSX Composite Index from the previous trading day) is: _____ (give month) _____ (give day), _____ (give year), when the index increased by _____ percent from the previous trading day.

(e) Create a histogram of the variable lev_chg_ind for the period from April 1, 2003 through December 31, 2003. This histogram is best described as _____ [answer with: positively skewed, negatively skewed, bimodal, or Normal (bell shaped)].

(f) Create a new variable named up that is equal to 1 if the S&P/TSX Composite Index went up since the previous trading day and is equal to 0 otherwise. (This kind of variable is called a dummy variable, also known as an indicator variable.) Complete the table below.
EXCEL TIPS: Use the `IF` function to create the dummy variable. For example, `=IF(J2>0,1,0)` would return a value of 1 if the cell J2 is positive and a value of zero otherwise.

Year	Percent of Trading Days S&P/TSX Composite Index Goes Up
2001	
2002	
2003	
2004	
2005	

Q7. Recall Zheng and Kahn (2017) and the data pol_chn.xlsx.

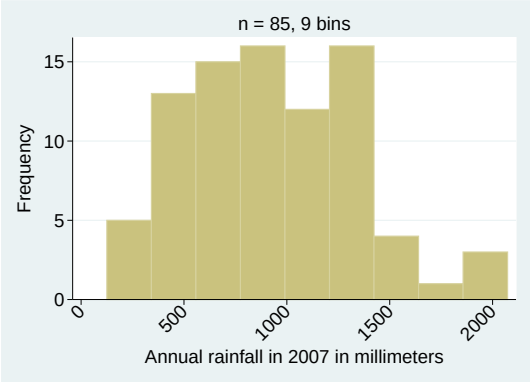
(a) Which kind of data are these? What is the unit of observation? How many observations are there? How many variables? How many identifier variables? How many nominal (categorical) variables? How many interval variables?

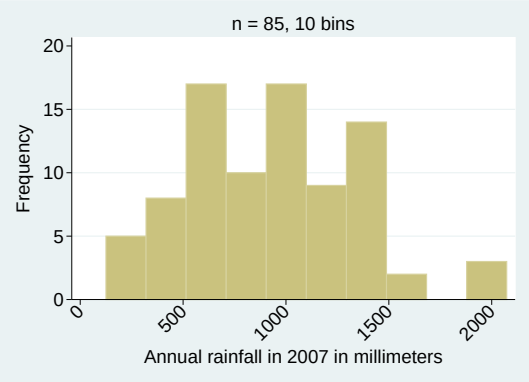
i. Suppose we dropped all observations except for the year 2011 and kept all variables. Which kind of data would these be? Unit of observation? Number of observations?

ii. 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.





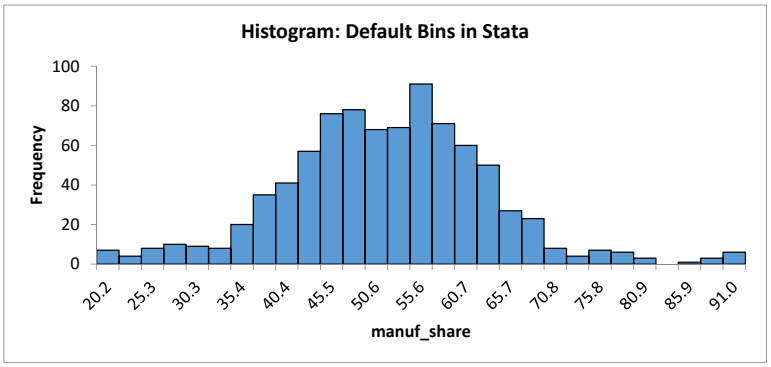
i. Why is the number of observations 85? In contrast, the number of observations is 850 for the variable manuf_share in these same data?

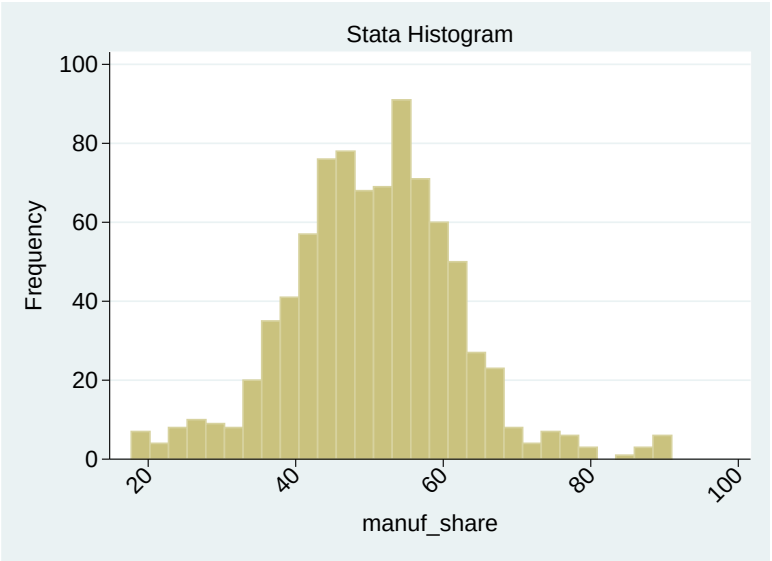
ii. What is the *suggested* number of bins according to $\left(MIN\left\{\sqrt{n}, \frac{10ln(n)}{ln(10)}\right\}\right)$, which happens to be the specific formula used by Stata?

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.

(d) Why do the Excel and Stata histograms (below) of manufacturing share in pol_chn.xlsx look the same but for age in cred_card.xlsx on page 25 looked different?





ANSWERS:

A1. (a) 182 unique countries

(b) 65 unique years

(c) 2.26078 (Note: The correct number of non-missing observations is 9,439.)

(d) 3.256812 (Note: The correct number of non-missing observations is 9,439.)

A2. (a) Create a pivot table using hh_inc.

Row Labels	Count of counter	Count of counter	Percent
Under $25,000	375	375	23.4
$25,000 - $49,999	501	501	31.3
$50,000 - $74,999	346	346	21.6
$75,000 - $99,999	197	197	12.3
$100,000 - $149,999	137	137	8.5
$150,000 or over	47	47	2.9
Grand Total	**1603**	1603	100.0

(b) Use a pivot table with hh_inc and chosedom, summarizing the average of chosedom. This tells the share of respondents in each income category that chose the dominant card. For example, 48.3% of those in the lowest income category chose 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**

A3. Use a pivot table to create the cross-tabulation (contingency table).

(a) Referring back to the worksheet readme in cred_card.xlsx, confidence measures how intensely the respondent agreed with the statement that "Choosing the best credit care 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) Of the 1,603 respondents, 512 respondents gave the same answer to the questions about about confidence and easiness. 249 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), 143 respondents found the choice at least fairly easy (5 or higher). Among those respondents indicating the choice was very easy (7), 86 respondents were at least fairly confident in their choice (5 or higher).

(c) 8.0 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, 30.5 percent quite strongly disagreed (2 or lower) with the confidence question. Among those respondents that quite strongly disagreed (2 or

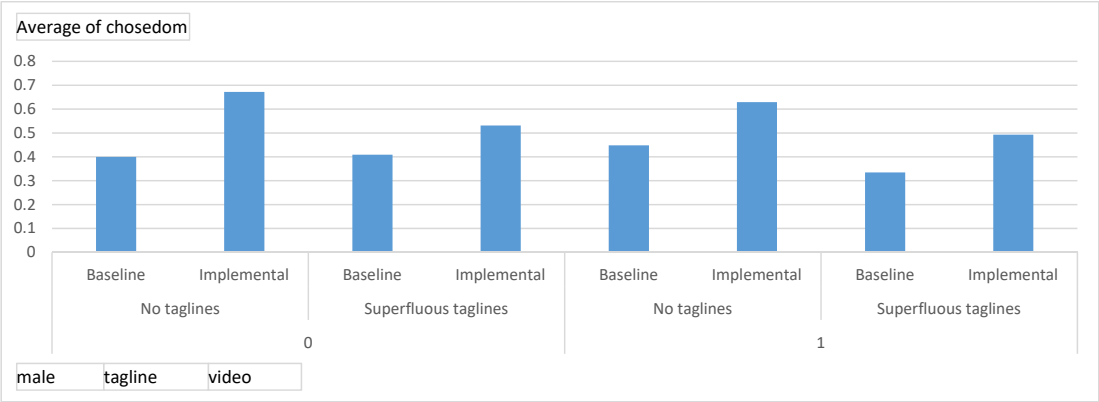
lower) with the confidence question, 91.4 percent quite strongly disagreed (2 or lower) with the easiness question. Among those respondents that were neutral (4) on the confidence question, 10.2 percent agreed (5 or higher) with the easiness question.

(d) The most common pair of answers to the easiness and confidence questions are 3 and 5, respectively. The second most common pair of answers to the easiness and confidence questions are 5 and 6, respectively.

(e) 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.

A4. (a) The figure below, created using a pivot chart, shows that males and females have similar reactions to the implemental video and superfluous taglines. In other words, breaking Figure 6 down further by sex does not ad much insight. Generally, researchers use the most simple figure that conveys the results in the data. If a simple figure is misleading (e.g. if there really were big differences by sex, combining the two sexes could be misleading), that suggests using a more complicated figure.

Average of chosedom	Column Labels		
Row Labels	0	1	Grand Total
Baseline	**0.404312668**	**0.390697674**	**0.397003745**
No taglines	0.4	0.448113208	0.425061425
Superfluous taglines	0.409090909	0.334862385	0.368020305
Implemental	**0.602094241**	**0.55952381**	**0.579800499**
No taglines	0.671875	0.629268293	0.649874055
Superfluous taglines	0.531578947	0.493023256	0.511111111
Grand Total	**0.504648074**	**0.474117647**	**0.488459139**



(b) For this part, answering requires looking at the pivot table created with the pivot chart above. 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.

A5. (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

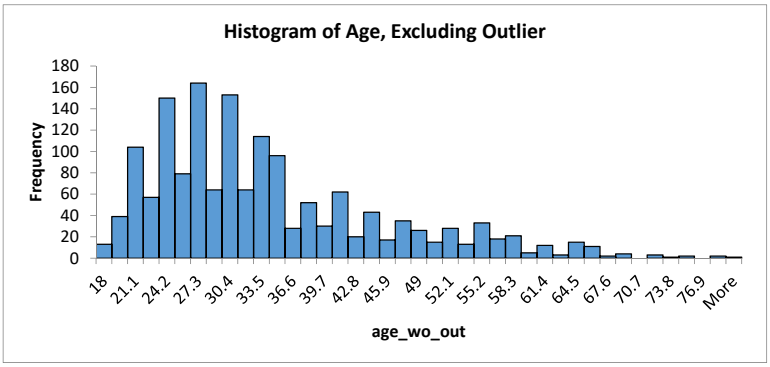
(c)

Select Percentiles for Age

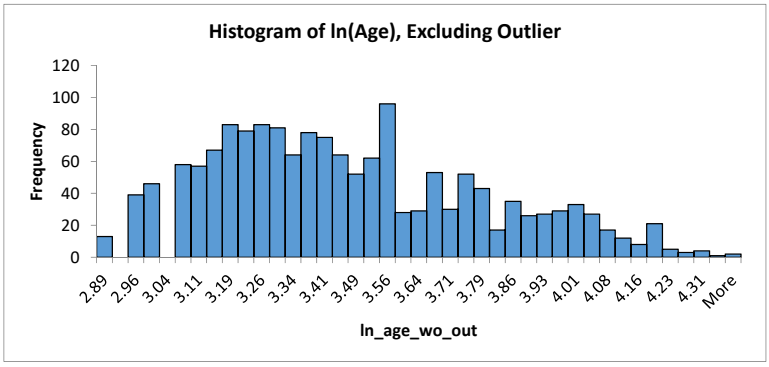
5th	20
10th	22
25th	25
50th	30
75th	39
90th	52
95th	57

(d) 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.

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



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



(g) The outlier (the 200-year-old respondent) contributes 200 to the sum in the formula for $\bar{X}$ and 27,728 to the sum in the formula for s^2. A respondent who is 18 years old contributes 18 to the sum in the formula for $\bar{X}$ and 240 to the sum in the formula for s^2. A respondent who is 30 years old contributes 30 to the sum in the formula for $\bar{X}$ and 12 to the sum in the formula for s^2. Of these three, the respondent who is 30 years old [answer with 18, 30 or 200] is closest to the mean.

A6. (a) During this period, the index is trending up.

(b) During this period, the volume of shares traded shows no clear trend up or down.

(c) During the period from April 1, 2003 through December 31, 2003, the worst trading day on the Canadian stock market (as measured by the level decline in the S&P/TSX Composite Index from the previous trading day) is: June 13, 2003, when the index declined by 96.3 from the previous trading day.

(d) During the period from April 1, 2003 through December 31, 2003, the best trading day on the Canadian stock market (as measured by the percent increase in the S&P/TSX Composite Index from the previous trading day) is: December 29, 2003, when the index increased by 1.52 percent from the previous trading day.

(e) This histogram is best described as Normal (bell shaped).

(f) After creating the requested dummy variable, use a pivot table to summarize the mean (which multiplied by 100 is the percent of up days) by year.

Year	Percent of Trading Days S&P/TSX Composite Index Goes Up
2001	48.8
2002	44.2
2003	60.3
2004	53.4
2005	59.0

A7. (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.

i. 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.

ii. 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) i. 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.

ii. 9 bins

iii. 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.

iv. The shape is fairly symmetric and nearly Normal (Bell). (This overview is unchanged whether we use 9 or 10 bins.)

(d) Because manufacturing share is measured very precisely (a percent with two decimal places), it turns out that none of the observations happen to lie on the boundary between two bins. Hence, the fact that Stata and Excel handle borderline cases differently is irrelevant. In contrast, for age, which is in an integer, borderline cases were fairly common and highlighted the different rules used by these two software packages.

B Module B: Interactive Tutorial Materials & Test Prep

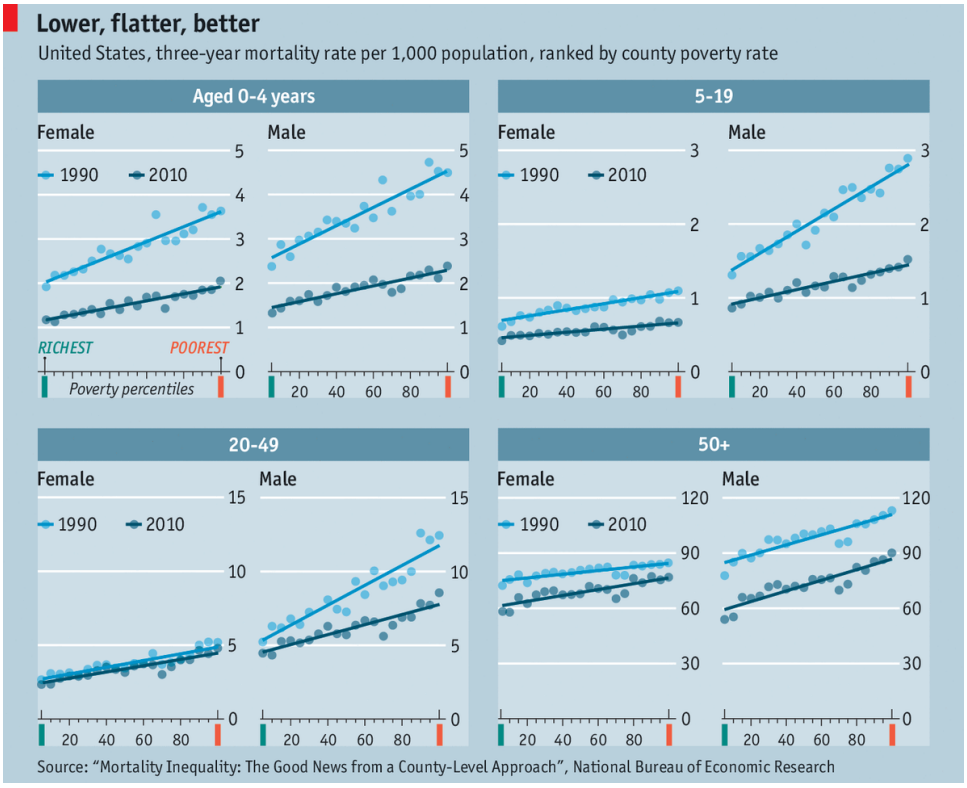
B.1 Module B.1: Association, Correlation, Regression & Composition Effects

Main course concepts: Describe relationships among quantitative variables using graphs, statistics, and regression. Confront "Simpson's paradox," which economists call "composition effects."

Source materials (full citations in Section F): 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*). Data from the U.S. Department of Energy www.fueleconomy.gov appears in practice questions.

Most relevant required readings: Sections 4.5 (Simpson's Paradox), 7.1-7.7. Background reading:

- *The Economist* reproduced a key figure from Currie and Schwandt (2016) adding formatting and captions. The y-axis is the three-year mortality rate per 1,000 population. The excerpts show how 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) [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 42 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

- To start, the numbers in boldface in Table B.1 are all you need to compute every other number. Deaths is the number of people in that county and in that age group who died in those years: for example, in the years 2000 to 2003, 6,300 people aged 45-59 died. Population is the total number of people in that age group living in the county in the start year: for example, in the year 2010, there are 52,000 people aged 40-44 in the county. From these two variables we can compute the deaths per 1,000 people: the three-year mortality rate per 1,000. For example, Table B.1 shows that from 1990 to 1993, for every 1,000 people aged 40-44, six people died.

- What if you wanted to combine these two 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 to compute 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 there is nothing mechanically wrong with this, there is a potentially serious problem if we wish to investigate how mortality is changing over time, which is what Currie and Schwandt (2016) wish to do.

- To see the problem, notice that BOTH age groups have declining mortality rates (i.e. things are improving between 1990 and 2010). However, a simple calculation (i.e. not adjusting) would paradoxically imply that when you combine the two groups, overall mortality rates are increasing (i.e. things are getting worse between 1990 and 2010). How can combining two positives lead to a negative? TWO things are changing: the mortality rates AND the

composition of the combined age groups. In the county in Table B.1, age group 45-59 is much larger and has a higher mortality rate generally than age group 40-44. Also, age group 45-59 is growing in terms of population size. Hence, when you combine these age groups, the combined group has an increasing fraction of the older age group (45-59).

- To deal with this, Currie and Schwandt (2016) construct *adjusted* mortality. Adjusted mortality controls for (i.e. holds fixed) the relative population sizes of groups 40-44 and 45-59 (i.e. holds the composition of the combined age group fixed) so that you can focus on the changes in mortality. The last column of Table B.1 gives the adjusted mortality for the 40-59 year olds and shows that when you combine two groups each with declining mortality rates, the combined group also has a declining mortality rate (i.e. this adjustment fixes the paradox).

- How to compute adjusted mortality Table B.1? First, for the original two age groups (40-44 and 45-59) construct adjusted population: hold it fixed at the 1990 level. Next, for the original two age groups, construct adjusted deaths: adjusted population times the mortality per 1,000 divided by 1,000, which is what the total deaths would have been without population growth (i.e. stayed at 1990 level). Next, sum the adjusted variables to obtain the total (adjusted) population and (adjusted) deaths for the combined age group (40-59). Finally, use the adjusted population and the adjusted deaths to compute the adjusted mortality per 1,000.

- 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.[3]

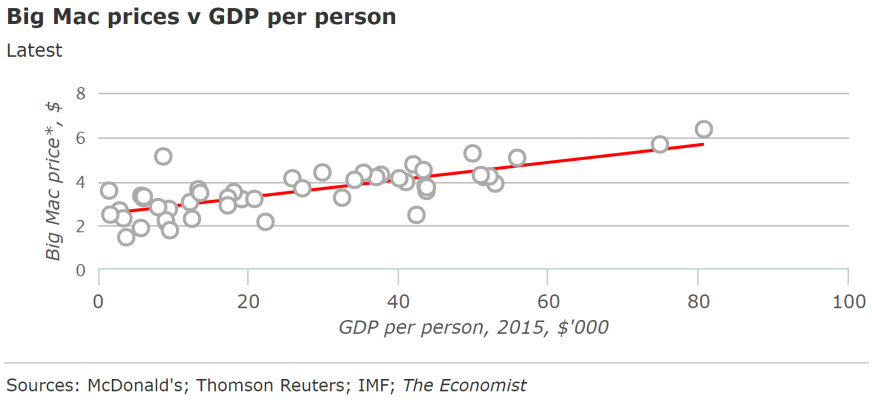


Figure 1: *The Economist* online showing the OLS results (red line) underlying their **January 2017** analysis, retrieved June 13, 2017. The regression line: y-hat = 2.487433 + 0.0394057x.

Additional readings (not required): The "The Big Mac Index" page maintained by *The Economist* has an interactive site including a scatter diagram. Wikipedia has a good entry for Simpson's paradox.

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.

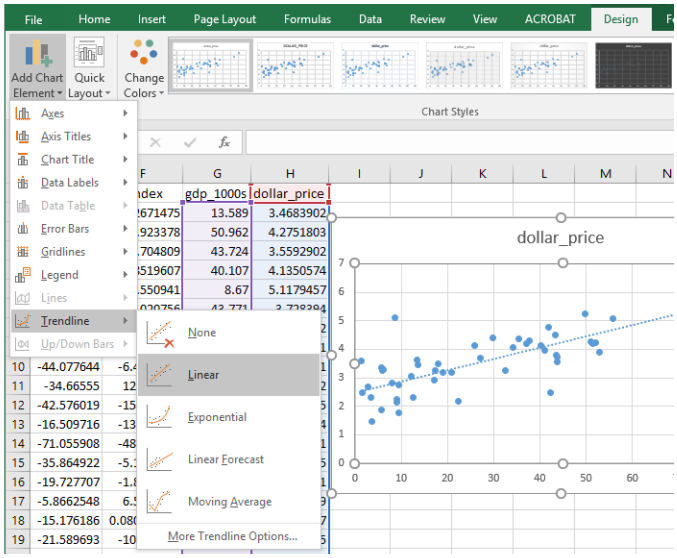
[3] *The Economist* uses the simple regression in Figure 1 to compute the adjusted Big Mac index. The basic idea is to control for differences in richness across countries. The details of that adjusted index are beyond the scope of our course: recent research exploring the cryptic remarks about how and why the adjusted index is constructed the way it is includes O'Brien and Ruiz de Vargas (2017) and Clements and Si (2017).

Interactive tutorial materials:

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

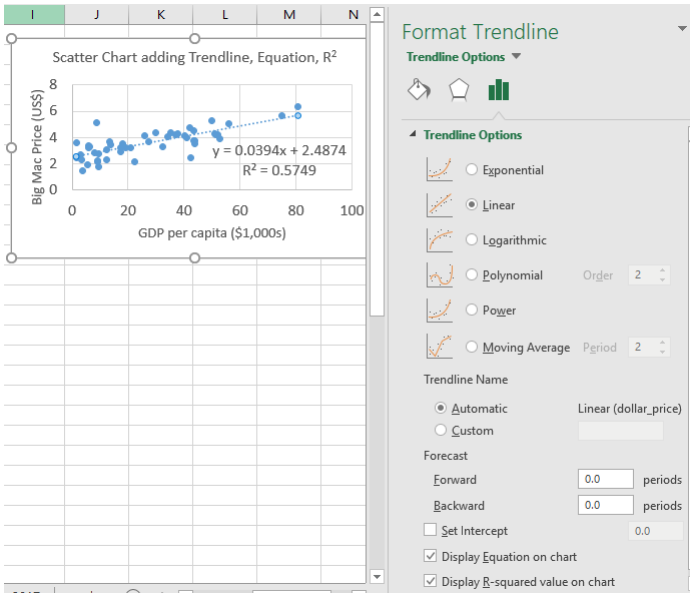
   (a) Using big_mac_jan_2017.xlsx, ***create*** a variable measuring GDP per capita in thousands of dollars (not dollars), giving it a meaningful name: gdp_1000s. ***Replicate*** the scatter plot, including the regression line in the graph, in Figure 1 on page 39.

   EXCEL TIPS: Copy all of the original data to a new worksheet. After creating the gdp_1000s variable, cut and insert it to be before the dollar_price variable: Excel scatter plots treat the first as the x variable and the second as the y variable. Select the entire two columns with gdp_1000s and dollar_price and, under the Insert tab, select Scatter from the chart choices. To include the regression line, add a Trendline, which is illustrated below.



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

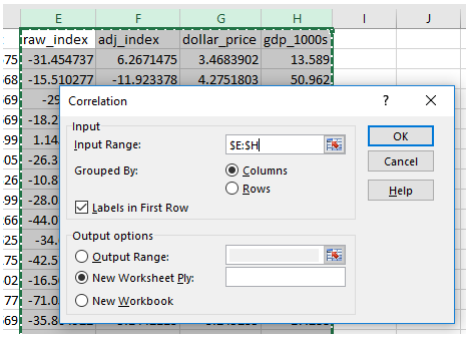
   EXCEL TIPS: To replicate the regression equation, click anywhere on the trendline in your scatter plot and select Format Trendline. Check the boxes for Display Equation on chart and Display R-squared value on chart.



(c) ***Construct a correlation matrix*** for the variables measuring: GDP per capita, dollar price of a Big Mac, raw index, and adjusted index. ***Check*** your work against the output below. (Note: It does not matter whether you measure GPD per capita in dollars or thousands of dollars because the coefficient of correlation is unit-free.)

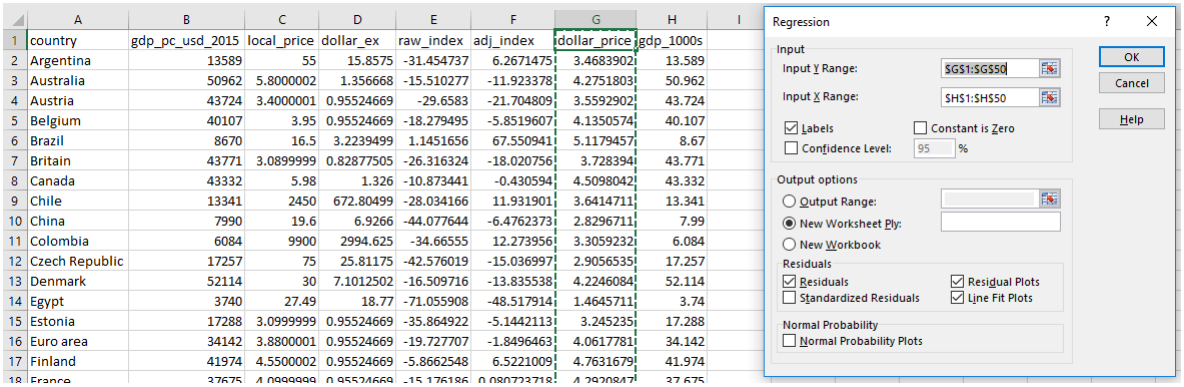
	raw_index	*adj_index*	*dollar_price*	*gdp_1000s*
raw_index	1.0000			
adj_index	0.6244	1.0000		
dollar_price	1.0000	0.6244	1.0000	
gdp_1000s	0.7582	-0.0211	0.7582	1.0000

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



(d) 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.

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.



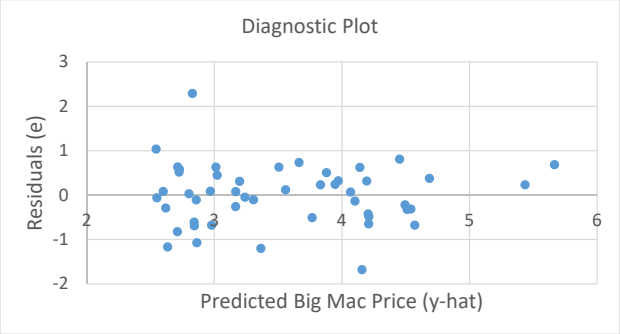
	country	gdp_pc_usd_2015	local_price	dollar_ex	raw_index	adj_index	dollar_price	gdp_1000s
2	Argentina	13589	55	15.8575	-31.454737	6.2671475	3.4683902	13.589
3	Australia	50962	5.8000002	1.356668	-15.510277	-11.923378	4.2751803	50.962
4	Austria	43724	3.4000001	0.95524669	-29.6583	-21.704809	3.5592902	43.724
5	Belgium	40107	3.95	0.95524669	-18.279495	-5.8519607	4.1350574	40.107
6	Brazil	8670	16.5	3.2239499	1.1451656	67.550941	5.1179457	8.67
7	Britain	43771	3.0899999	0.82877505	-26.316324	-18.020756	3.728394	43.771
8	Canada	43332	5.98	1.326	-10.873441	-0.430594	4.5098042	43.332
9	Chile	13341	2450	672.80499	-28.034166	11.931901	3.6414711	13.341
10	China	7990	19.6	6.9266	-44.077644	-6.4762373	2.8296711	7.99
11	Colombia	6084	9900	2994.625	-34.66555	12.273956	3.3059232	6.084
12	Czech Republic	17257	75	25.81175	-42.576019	-15.036997	2.9056535	17.257
13	Denmark	52114	30	7.1012502	-16.509716	-13.835538	4.2246084	52.114
14	Egypt	3740	27.49	18.77	-71.055908	-48.517914	1.4645711	3.74
15	Estonia	17288	3.0999999	0.95524669	-35.864922	-5.1442113	3.245235	17.288
16	Euro area	34142	3.8800001	0.95524669	-19.727707	-1.8496463	4.0617781	34.142
17	Finland	41974	4.5500002	0.95524669	-5.8662548	6.5221009	4.7631679	41.974
18	France	37675	4.0999999	0.95524669	-15.176186	0.0807237	4.2920847	37.675

(e) ***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 $s_e = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$ (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 1c 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.)

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

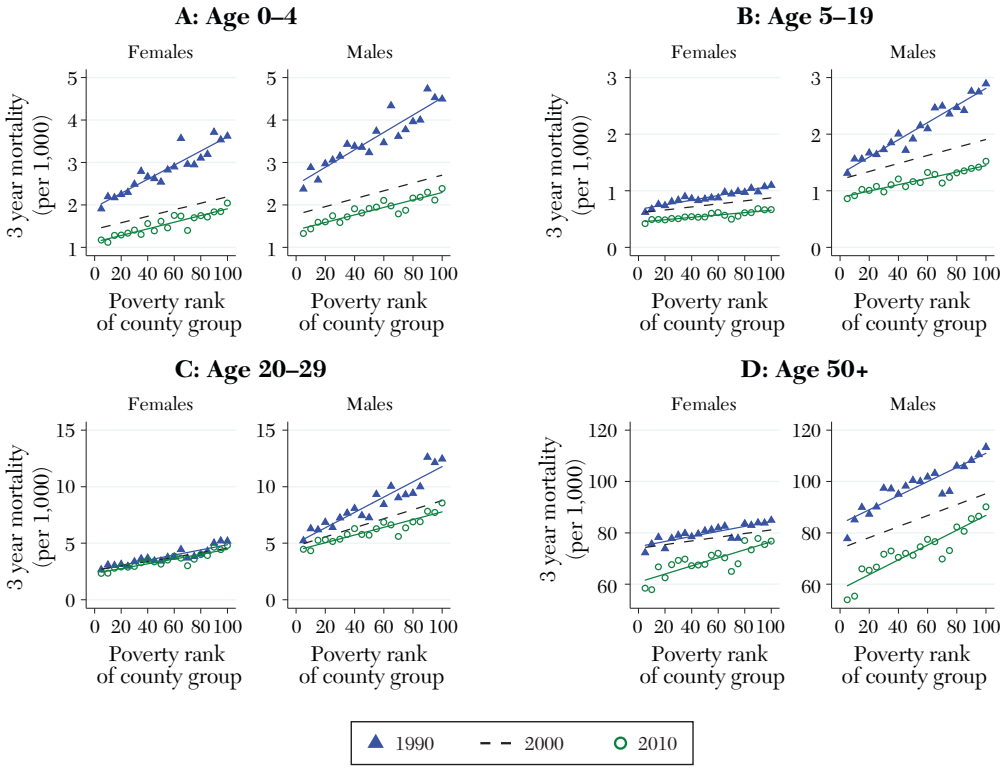
EXCEL TIPS: Go to the new worksheet with the output created in part 1d. Under RESIDUAL OUTPUT, select the Predicted dollar_price and Residuals columns and Insert a Scatter plot. (If your worksheet is missing the RESIDUAL OUTPUT, repeat the previous part, this time remembering to check the boxes under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



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

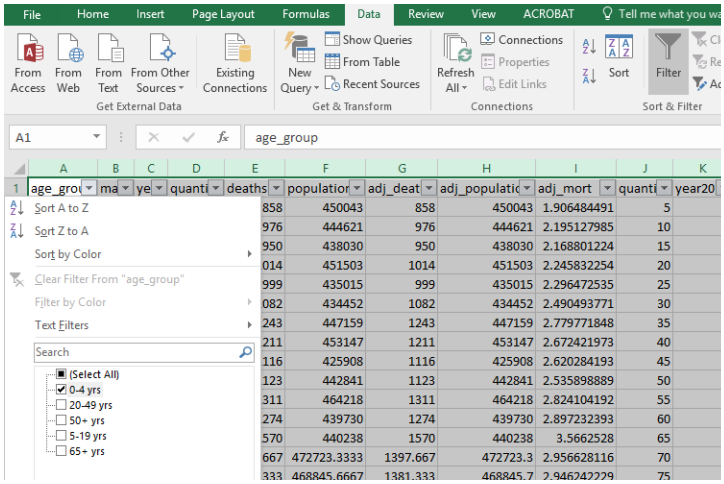


Source: Authors using data from the Vital Statistics, the US Census, and the American Community Survey.
Note: Three-year mortality rates for four different age groups are plotted across county groups ranked by their poverty rate. Mortality rates in 2000 and 2010 are age-adjusted using the 1990 population, 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. 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 (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.) ***Verify*** that you obtain: y-hat = 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.



(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 boldface, 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

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

Main course concepts: Natural logarithms in regression. Working with real data (PWT) to get economic results: estimating GDP per capita growth rates using population and GDP levels.

Source materials (full citations in Section F): We use a major database – Penn World Tables (PWT) – to replicate parts of the analysis from 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 as will we in tutorials. You will also encounter newer versions.

Most relevant required readings: Section 7.8. "Logarithms in Regression Analysis with Asiaphoria", which also assigns *parts* of Pritchett and Summers (2014) and an NBER digest of it. After reading those, review Table 1, an except describing it, and the plan for Modules B.2 and B.3:

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-squared of the regression (which is of course the square of the correlation coefficient). We use the PWT8.0 (Feenstra, Inklaar, and Timmer 2013) data on local currency real GDP from national accounts (since we are not yet comparing levels) 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)

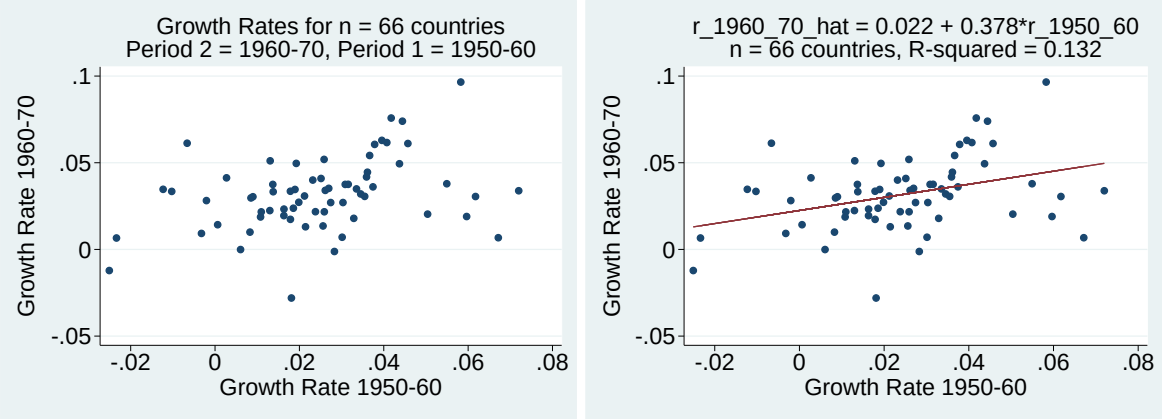


Figure 2: These scatter plots and OLS results visually illustrate the first row of results in Table 1. In these scatter plots and regression, the unit of observation is a country: each dot is a country. Module B.2 shows how to find each dot in the scatter plots above ("first-stage" regressions). Module B.3 shows how to run regressions like in Table 1, illustrated by the red line above ("second-stage" regressions).

- In Module B.2 we take the *first* step on the journey to replicate Table 1. We use regression to estimate GDP per capita growth rates for different decades and countries. In Module B.3 you will complete the replication using data like that we create together today but with all available countries. Today we complete the first step for just Canada and China. Today (and for homework) you will complete the data shown in Table B.2. The r_ abbreviates rate. For example, r_1970_80 records the growth rate of GDP per capita in the 1970s.

Table B.2: Part of a data that you construct in Module B.2 (blanks to be filled in during tutorial)

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

- As you will see today, the growth rates that will populate Table B.2 are obtained by running a simple regression where the y-variable is the natural log of GDP per capita. *Why* did Summers and Pritchett (2014) use the natural log of GDP per capita instead of just GDP per capita (without the natural log transformation)? There are TWO distinct reasons why: (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. Today, we illustrate these reasons interactively.

Datasets: For Pritchett and Summers (2014): asiap_pwt_80_one_decade.xlsx, where "asiap" abbreviates "Asiaphoria" from the title and "pwt_80" abbreviates Penn World Tables, version 8.0. It includes only the PWT data that Pritchett and Summers (2014) used in their analysis[4] leading to

[4]It excludes observations with missing data, labeled as outliers, or in decades with insufficient observations.

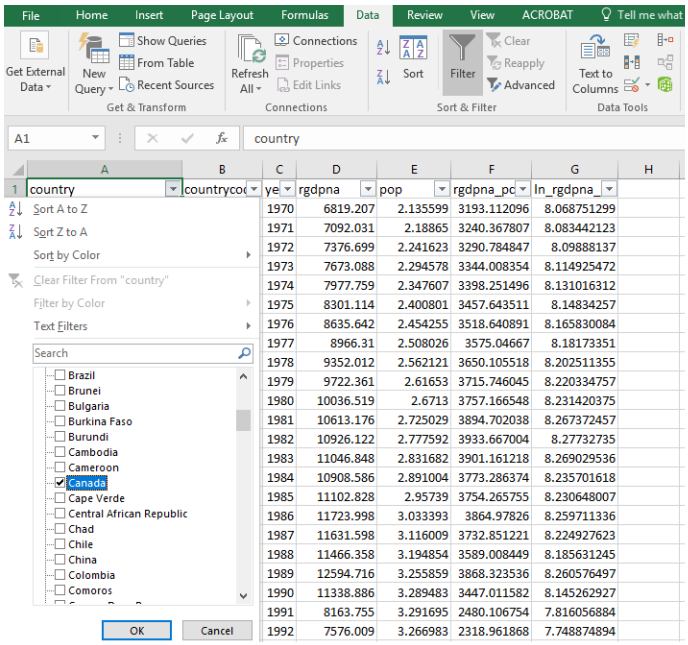
Table 1. Also, "one_decade" means that these are the data Pritchett and Summers (2014) used to estimate growth rates at the level of one decade (e.g. growth rates in the 1990s: 1990-2000).

Interactive tutorial materials:

1. To start work on filling in Table B.2, open asiap_pwt_80_one_decade.xlsx.

   (a) ***Create*** a new variable measuring real GDP per capita. Next, ***create*** another new variable that is the natural log of real GDP per capita.

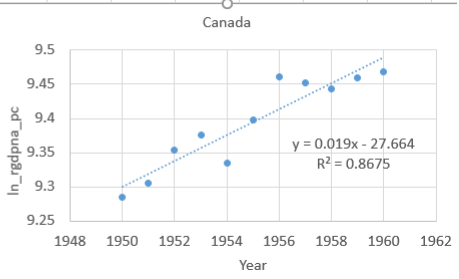
   (b) Starting with Canada, ***copy and paste*** all observations for Canada into a new worksheet.
   EXCEL TIPS: In the Data tab, click the Filter button. Uncheck the box for "(Select All)" and then check Canada. Next, select the entire worksheet and copy to a new worksheet. Conveniently, it will only copy the filtered rows (i.e. just Canada).



country	countrycod	ye	rgdpna	pop	rgdpna_pc	ln_rgdpna_
		1970	6819.207	2.135599	3193.112096	8.068751299
		1971	7092.031	2.18865	3240.367807	8.083442123
		1972	7376.699	2.241623	3290.784847	8.09888137
		1973	7673.088	2.294578	3344.008354	8.114925472
		1974	7977.759	2.347607	3398.251496	8.131016312
		1975	8301.114	2.400801	3457.643511	8.14834257
		1976	8635.642	2.454255	3518.640891	8.165830084
		1977	8966.31	2.508026	3575.04667	8.18173351
		1978	9352.012	2.562121	3650.105518	8.202511355
		1979	9722.361	2.61653	3715.746045	8.220334757
		1980	10036.519	2.6713	3757.166548	8.231420375
		1981	10613.176	2.725029	3894.702038	8.267372457
		1982	10926.122	2.777592	3933.667004	8.27732735
		1983	11046.848	2.831682	3901.161218	8.269029536
		1984	10908.586	2.891004	3773.286374	8.235701618
		1985	11102.828	2.95739	3754.265755	8.230648007
		1986	11723.998	3.033393	3864.97826	8.259711336
		1987	11631.598	3.116009	3732.851221	8.224927623
		1988	11466.358	3.194854	3589.008449	8.185631245
		1989	12594.716	3.255859	3868.323536	8.260576497
		1990	11338.886	3.289483	3447.011582	8.145262927
		1991	8163.755	3.291695	2480.106754	7.816056884
		1992	7576.009	3.266983	2318.961868	7.748874894

   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 (see below), insert a Scatter Chart. Recall from parts 1a and 1b of Module B.1 on page 40 that you can add a trendline and the OLS equation and R^2.

	A	B	C	D	E	F	G	H
1	country	countrycode	year	rgdpna	pop	rgdpna_pc	year	ln_rgdpna_pc
2	Canada	CAN	1950	150703.707	13.990079	10772.18413	1950	9.284722547
3	Canada	CAN	1951	157290.185	14.308846	10992.51365	1951	9.304969743
4	Canada	CAN	1952	170464.937	14.763062	11546.71958	1952	9.354156656
5	Canada	CAN	1953	178861.719	15.160247	11798.0742	1953	9.375691594
6	Canada	CAN	1954	176854.656	15.612427	11327.81316	1954	9.335016322
7	Canada	CAN	1955	193193.848	16.025907	12055.09604	1955	9.397242757
8	Canada	CAN	1956	211082.171	16.420037	12855.15806	1956	9.461500415
9	Canada	CAN	1957	216448.669	16.984243	12744.08692	1957	9.452822672
10	Canada	CAN	1958	220026.34	17.435405	12619.51414	1958	9.442999636
11	Canada	CAN	1959	228970.494	17.84481	12831.20941	1959	9.459635717
12	Canada	CAN	1960	236125.82	18.23894	12946.24688	1960	9.468561209
13	Canada	CAN	1961	242808.626	18.606591	13049.60301	1961	9.476512992
14	Canada	CAN	1962	259515.633	18.956929	13689.75075	1962	9.524402711



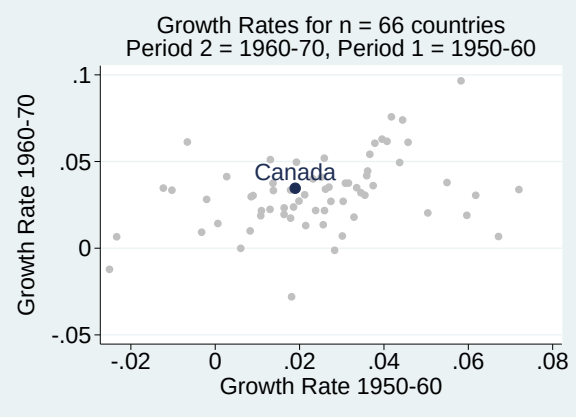
   ii. For the same years (1950-1960), ***run a simple regression*** of the natural log of real GDP per capita (y variable) on year (x variable). Using the worksheet "Your results

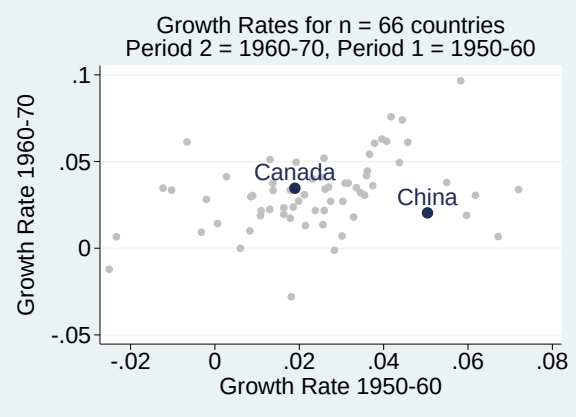
Canada and China", ***copy*** the "slope" coefficient into the cell for Canada for 1950-1960: r_1950_60, which is the growth rate. ***Verify*** that you obtain 0.0189558: an average GDP per capita growth of 1.9% annually in the 1950's in Canada.

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 0.0345877: an average GDP per capita growth of 3.5% annually the 1960's in Canada.

iv. ***Note*** that 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 45, 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 0.0503967: an average GDP per capita growth of 5.0% annually in the 1950's in China.

ii. ***Repeat*** part 1(c)i but for 1960-1970. ***Verify*** that you obtain 0.020365.

iii. ***Note*** that 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 45, which is illustrated above (right).

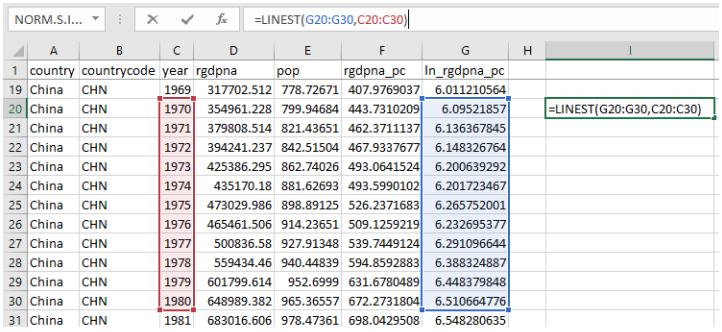
(d) We have not yet considered the other 64 countries (the light grey dots above) and we are not even done with Canada and China. ***Recall*** that Table 1 on 44 also considers growth rates in other decades (and two-decade periods). ***Note*** the blank cells still to be filled.

What we have filled in so far in Table B.2 on page 45

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 0.0387015, 0.0825142, 0.0909205, and 0.0973505, respectively.

EXCEL TIPS: The function LINEST is a short-cut. 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 0.038701455. (Similarly, the function RSQ returns the R^2.)



	A	B	C	D	E	F	G	H	I
1	country	countrycode	year	rgdpna	pop	rgdpna_pc	ln_rgdpna_pc		
19	China	CHN	1969	317702.512	778.72671	407.9769037	6.011210564		
20	China	CHN	1970	354961.228	799.94684	443.7310209	6.09521857		=LINEST(G20:G30,C20:C30)
21	China	CHN	1971	379808.514	821.43651	462.3711137	6.136367845		
22	China	CHN	1972	394241.237	842.51504	467.9337677	6.148326764		
23	China	CHN	1973	425386.295	862.74026	493.0641524	6.200639292		
24	China	CHN	1974	435170.18	881.62693	493.5990102	6.201723467		
25	China	CHN	1975	473029.986	898.89125	526.2371683	6.265752001		
26	China	CHN	1976	465461.506	914.23651	509.1259219	6.232695377		
27	China	CHN	1977	500836.58	927.91348	539.7449124	6.291096644		
28	China	CHN	1978	559434.46	940.44839	594.8592883	6.388324887		
29	China	CHN	1979	601799.614	952.6999	631.6780489	6.448379848		
30	China	CHN	1980	648989.382	965.36557	672.2731804	6.510664776		
31	China	CHN	1981	683016.606	978.47361	698.0429508	6.548280635		

ii. Tutorial time permitting (otherwise, for homework), ***obtain*** the values of r_1970_80, r_1980_90, r_1990_00, and r_2000_10 for Canada. ***Verify*** that you obtain 0.0281345, 0.0194571, 0.0205556, 0.0091647, respectively.

(e) ***For homework,*** consider repeating all these steps for each country (not just Canada and China)! In Module B.3 you will be given the results for the remaining 140 countries.

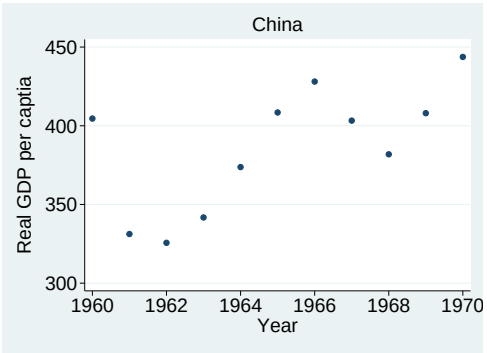
2. "*Why* did Summers and Pritchett (2014) use the natural log of GDP per capita instead of just GDP per capita (without the log transformation)?" There are TWO distinct reasons: (1) growth *rates* can be directly compared across countries (richer versus poorer) and across time (when a country was poorer 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 annually in the 1960's 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).

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 7.71667, which corresponds to an average GDP per capita growth of $8 USD annually in the 1960's in China. This is an estimate of the *level* of growth each year.

v. ***Note*** that compared to $527, $8 looks tiny, but, ***consider*** that China was much poorer per capita than Canada in the 1960s. ***For homework,*** study this further explanation. 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 China versus $15,451 USD in Canada. Recall your results from earlier in today's tutorial where you found that, in the 1960s, GDP per capita grew at an average annual rate of 2.0% in China and 3.5% in Canada. These are estimates of the *rate* of growth each year. 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*527/15,451=3.4% (which is very close to 3.5%) for Canada and by looking at 100*8/408=2.0% (which, rounded, is the same as 2.0%) for China.

(b) Tutorial time permitting (otherwise, for homework), 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 2000's in China.

- ***For homework,*** note how $542 overstates growth in the early 2000's and understates growth in the late 2000's: 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 today's tutorial where you found that GDP per capita grew at an average annual rate of 9.7% in China in the 2000s.

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

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

Source materials (full citations in Section F): Same as Module B.2.

Most relevant required readings: Same as Module B.2. Remember to review Table 1 on page 44 and the associated excerpt. Recall the twelve simple regressions (where the y variable was the natural log of GDP per capita and the x variable was year) you ran in Module B.2 to obtain an estimate of the growth rate (the "slope" coefficient) in each of six decades for two countries: Canada and China. The complete results from Module B.2 are summarized below in Table B.3.

Table B.3: Decade growth rate estimates for Canada and China obtained in Module B.2

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	0.0281345	0.0194571	0.0205556	0.0091647
China	CHN	0.0503967	0.020365	0.0387015	0.0825142	0.0909205	0.0973505

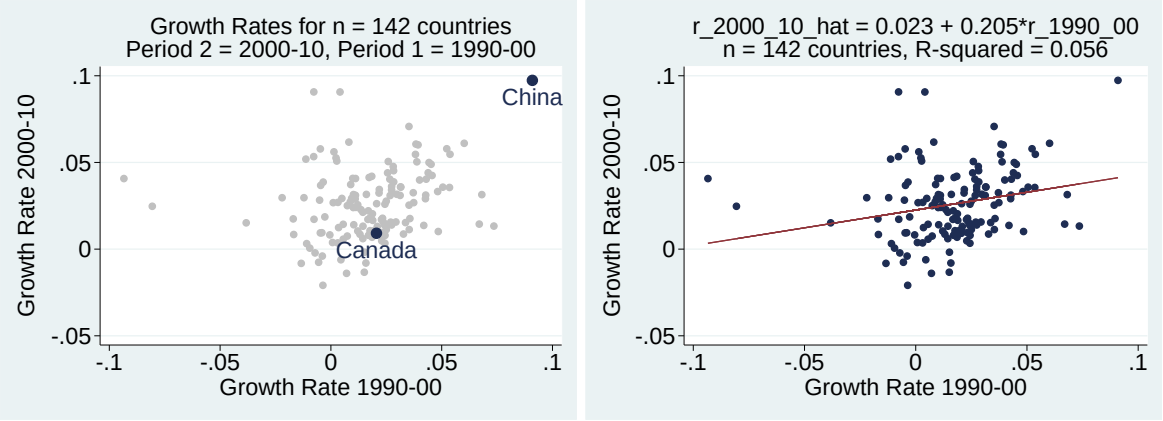


Figure 2: These scatter plots and OLS results visually summarize Modules B.2 and B.3. Recall that Module B.2 shows how to find each dot in the scatter plots ("first-stage" regressions): interactively we did this for Canada and China, which are highlighted in the left scatter diagram. Module B.3 shows how to run regressions like in Table 1, which is illustrated by the red line in the right scatter diagram ("second-stage" regressions).

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 (i.e. these are like Table B.3 except for all countries, not just Canada and China).

Interactive tutorial materials:

1. ***Browse*** asiap_rates_pwt_80.xlsx. It includes all available countries, not just Canada and China.
    - ***For homework,*** verify that 142 simple regressions were 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.2) were run to obtain all the data shown in asiap_rates_pwt_80.xlsx.

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 1990's and countries' growth rates in the 2000's. ***Verify*** that the *unit of observation* in your graph is a country (i.e. each dot corresponds to a different country).

3. ***Review*** Table 1 on page 44. 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).

   - ***For homework,*** reflect on the interpretations offered next. The value 0.205 means that, on average, countries with an additional one percentage point of growth in the 1990's had only an additional 0.205 percentage points of growth in the 2000's. In other words, the slope is positive but fairly flat: if we look across a wide range of countries, fast growth in the 1990's is *not* a guarantee of fast growth in the 2000's. Similarly, looking at the cross-section of 142 countries, slow growth in the 1990's is *not* a guarantee of slow growth in the 2000's. The very low value of the R-squared means that only 5.6 percent of the variation across countries in growth rates in the 2000's is explained by variation across countries in growth rates in the 1990's.

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

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

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

   EXCEL TIPS: Use the `RANK` function. Specifically, 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. In Data Analysis, use Correlation on your two new variables (rank_1990_00 and rank_2000_10).

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 (just a copy of it).

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

(c) ***For homework,*** replicate all of the "Correlation" and "Rank Correlation" results. (Note: Because it just so happens that Israel and the United Kingdom literally 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 equally at 12 it will come out to be 0.218.)

EXCEL TIPS: Select each subset using the filter tool and copy that subset to a new worksheet (that you name something like "OECD Countries") and put the outputs of your analyses in the associated worksheet.

B.4 Practice test questions for Module B

QUESTIONS:

Q1. 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?

(d) Compute the correlation between the raw index and the adjusted index excluding Switzerland. Compare and contrast the correlation with and without Switzerland.

Q2. Consider the 2015 Big Mac index and Questions (11) - (20) on the October 2015 term test.

(a) Using big_mac_jul_2015.xlsx, replicate the output given in the preamble to those questions.

(b) Solve multiple choice questions (11) - (20).

Q3. 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).

Q4. 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.

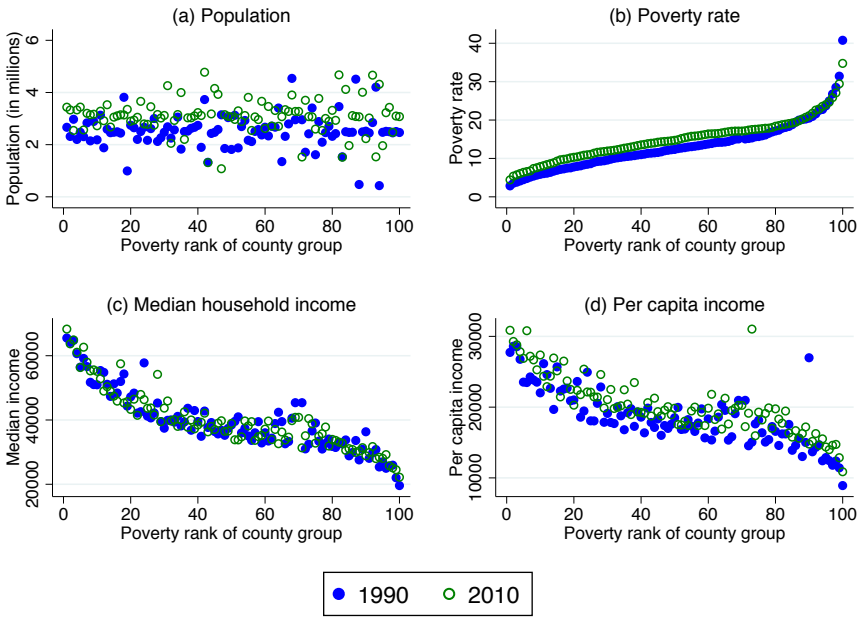
(b) Compute the value of the R-squared.

(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.

Q5. 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.

(a) Replicate the scatter diagram in Panel (c) in Figure A2 for the year 1990 only.

(b) 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*.

(c) Create a scatter plot of median income versus per capita income using only the year 2010 data. How is that relationship best described?

(d) 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.)

Q6. Following Currie and Schwandt (2016), fill in all of the missing values in the table below.

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						

Q7. 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) Repeat part (a) for Argentina.

(c) 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 and/or Argentina.

Q8. Recall Pritchett and Summers (2014) and use asiap_pwt_80_one_decade.xlsx for all subparts. These questions focus on Bangladesh between 1960-2010.

(a) For Bangladesh between 1960-2010, plot population and real GDP on the same graph.
EXCEL TIPS: Select year, population, and real GDP and insert a line chart. To make the graph readable, add a secondary axis (double click on one of the series and click secondary axis under series options). You should also add axis titles (click the plus sign to add chart elements). Finally, it is a good idea to use different series markers for people looking at a black-and-white version (under marker when formating data series).

(b) Visually inspect your graph. Identify which variable is growing at a faster rate. Given this, is real GDP *per capita* increasing or decreasing for Bangladesh between 1960-2010?

(c) Compute the coefficient of correlation between population and real GDP for Bangladesh between 1960-2010.

(d) 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.

Q9. 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.

(a) 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"?

(b) In that same row, what does 142 mean?

(c) For that same row, what is the value of the x variable for Canada? What is the value of the y variable for Canada?

(d) For the first row under the panel heading "Gap of two decades," what is the value of the y variable for Canada?

Q10. Recall Pritchett and Summers (2014) and use asiap_rates_pwt_80.xlsx for all subparts, which focus on replicating the results in Table 1 on page 44.

(a) In Table 1 there are three panels: "Adjacent decades," "One decade apart," and "Two decades apart." 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.

(b) In Table 1 under the panel "Adjacent decades," why is the sample size only 66?

(c) In Table 1 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?

Q11. 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 44.

(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?

Q12. 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 44. 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.

Q13. Recall Pritchett and Summers (2014) and use asiap_pwt_90_all.xlsx for all subparts, which contains the most recent PWT data (version 9.0).

(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 44 in Pritchett and Summers (2014)?

ANSWERS:

A1. (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.)

(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.

A2. (a) Check that your Excel output matches the sample size, intercept, slope and R-squared values given in the preamble.

(b) Check your answers to the multiple-choice questions against the answer key.

A3. (a) There is an extremely strong, negative, and non-linear association between CO2 emissions and fuel efficiency.

(b) Yes, the diagnostic plot shows a very clear U-shaped pattern: it is catching the non-linearity 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).

A4. (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 that 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.

A5. (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 non-linear. 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 non-linear. 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 non-linear. 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 non-linear.

(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.

A6. 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

A7. (a) The results for Brazil:

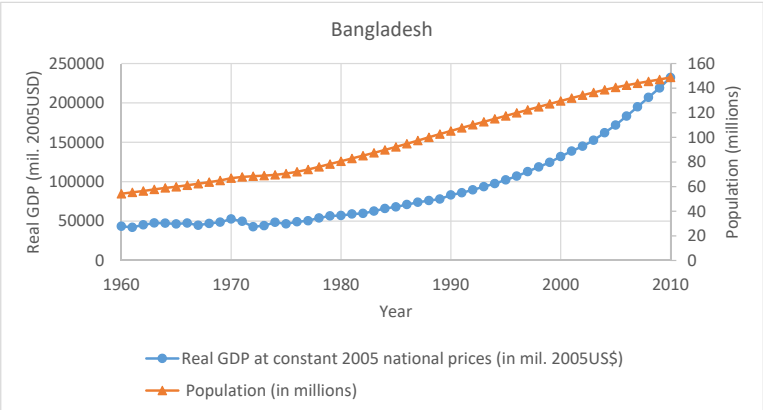
Half decade	Coefficient	n
1985 - 1990	0.0001	6
1990 - 1995	0.0159	6
1995 - 2000	0.0023	6
2000 - 2005	0.0151	6
2005 - 2010	0.0323	6

(b) 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

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

A8. (a) Check that your figure looks similar to the one below.



(b) The graph shows real GDP is growing faster than population, which implies that real GDP per capita is generally increasing.

(c) The coefficient of correlation is 0.9327.

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

A9. (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.

A10. (a) For the correlation, rank correlation, regression coefficient, and R-squared see Table 1 on page 44 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.

(b) 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.

(c) 66

A11. (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.

A12. (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.

A13. (a) Following the methods in Module B.2:

country	countrycode	r_1994_04	r_2004_14
Canada	CAN	0.0244016	0.0054735
Mexico	MEX	0.0166058	0.0072423
United States	USA	0.0224004	0.0029438

(b) Following the methods in Module B.3:

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	-0.3419709	0.4107	3

(c) Yes, these results look unusual compared to Table 1 on page 44. 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.

C Module C: Interactive Tutorial Materials & Test Prep

C.1 Module C.1: Sampling Distributions

Main course concepts: Using simulation to obtain a sampling distribution (the distribution of a *statistic* reflecting variation caused by sampling error). The Central Limit Theorem, an important theoretical result for the sampling distribution of the sample mean $\bar{X}$.

Source materials (full citations in Section F): We will work with the *population* of *all* Ontario public sector employees making $100K+ in either the "Universities" or "Colleges" sectors, using the public disclosures of 2016 salaries. This is abbreviated ON Univ. & Col. (2016).

Most relevant required readings: Chapter 10 and below, which includes a select review of Chapter 10 and background reading on Ontario's public disclosure of salaries.

- Section 10.3 of the textbook does a simulation of rolling dice 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 the simulation results. For example, for $n = 3$, the simulation involves throwing three dice 10,000 times and recording 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. The other two figures reproduced below show $n = 1$ and $n = 20$. It is important to keep straight the *sample size* (e.g. $n = 3$) versus the *number of simulation draws*, which is 10,000 in all of these figures. We denote the sample size as the usual n and the number of simulation draws as m. Today, you will simulate sampling distributions for $n = 10$, 25, and 100 and $m = 500$ and 10,000.

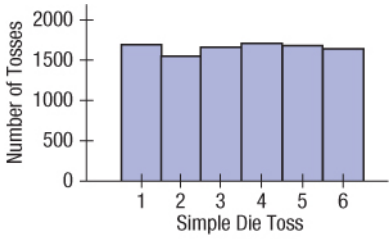


Figure 10.3 Simple die toss.

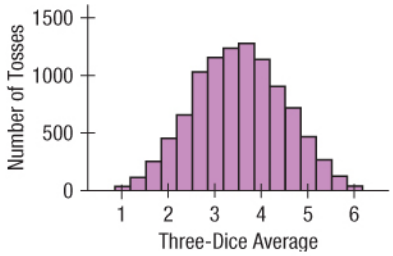


Figure 10.5 Three-dice average.

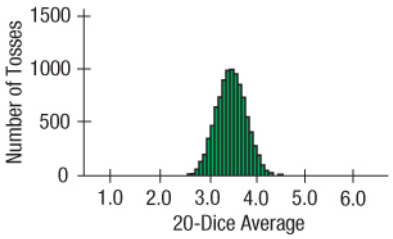


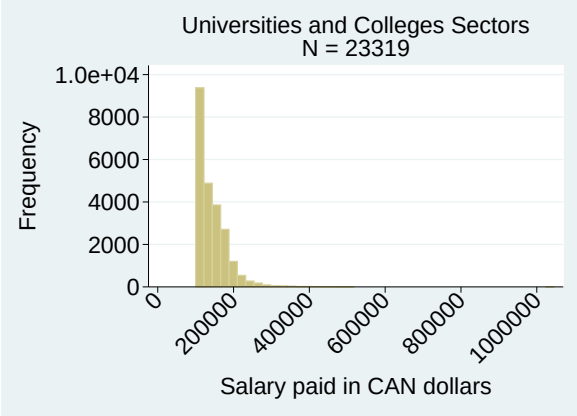
Figure 10.7 Twenty-dice average.

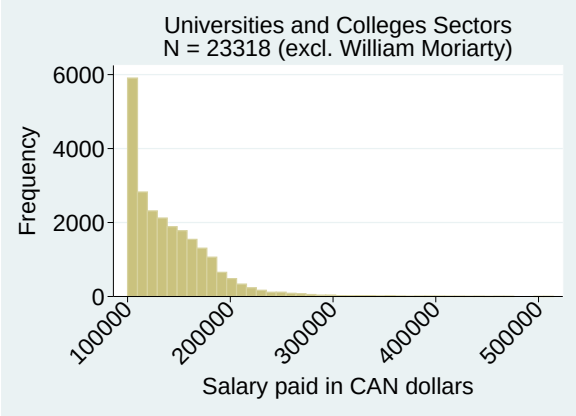
- The Public Sector Salary Disclosure Act of 1996 mandates that *all* organizations receiving funding from the Province of Ontario publicly disclose the name, job title, salary and taxable benefits of *all* employees paid $100,000 (CAN) or more in the previous calendar year.

- This includes a broad range of employees. Some examples: 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+).





Textbook case studies (extra practice): "Real Estate Simulation" on pp. 325-26

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 tutorial materials:

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 the 2016 calendar year. ***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. ***Consider*** that 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 tutorial time, 497 random samples are already drawn from the population. ***Follow*** these steps to draw the remaining three random samples (for a total of $3 + 497 = 500$ samples):

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

   (a) ***Generate*** one variable containing random numbers for each sample you wish to draw.
   EXCEL TIPS: Column A in the worksheet Drawing random samples contains all 23,319 salaries copied from the population. Column B contains random numbers from a Uniform[0,1] distribution. Column C contains 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 + ↓** discussed on page 12 (part 1b 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, which updates with every edit making it impossible to sort or to retrace your steps. Random #1, #2, and #3 should each contain *different* random numbers.

(b) ***Sort*** the population by the variable of random numbers.

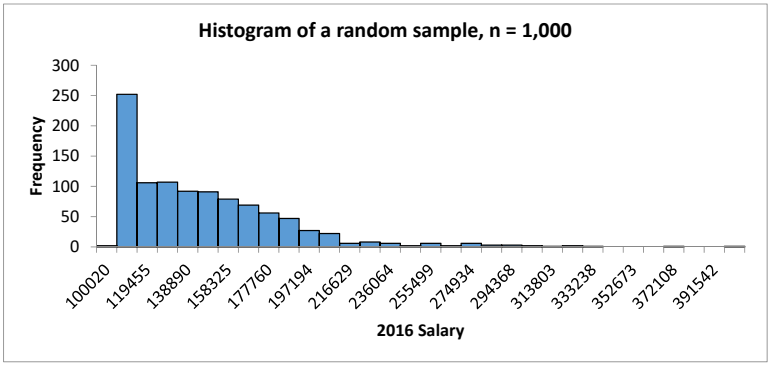
EXCEL TIPS: Select the entire worksheet Drawing random samples and sort by Random #1.

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

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 the distribution 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 below. (If your sample happens to include William Moriarty it will have a much longer right tail.)



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

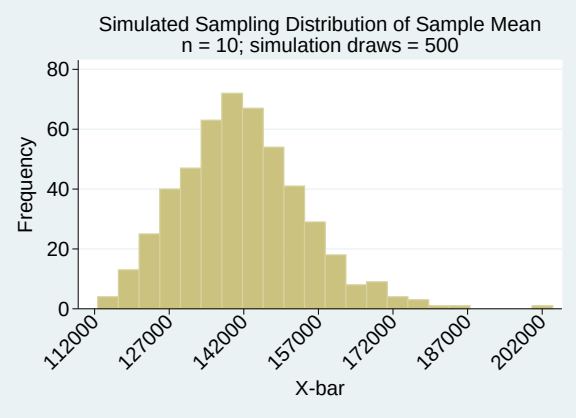
4. Now continuing with our simulation, we have already drawn many random samples. Let's start by ***considering*** $n = 10$. In other words, consider the first 10 observations from each of the 500 random samples. Recall that n denotes the sample size and m denotes the number of simulation draws.

EXCEL TIPS: Browse the worksheet $n = 10, m = 500$. It automatically pulls the first 10 observations from each of the 500 random samples in worksheet 500 random samples, n=100.

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

EXCEL TIPS: Use the AVERAGE function and copy and paste.

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



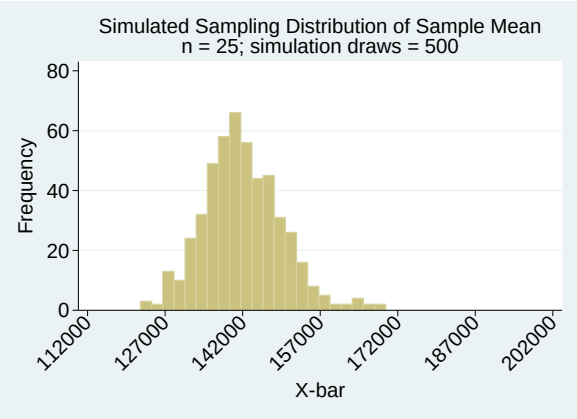
EXCEL TIPS: To save tutorial time, the worksheet $n = 10, m = 500$ is set up to automatically draw a histogram once the previous steps are complete. Note that your histogram will not be identical because your simulation has three random samples generated interactively and because Stata and Excel draw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

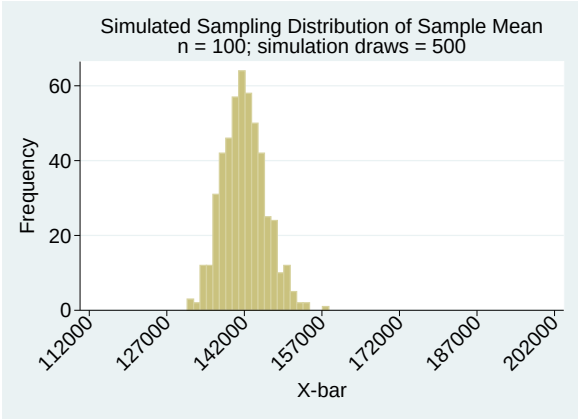
(c) ***Compute*** the mean of the 500 sample means. ***Compute*** the standard deviation of the 500 sample means. ***Verify*** that your simulation results are similar to what theory predicts: $\mu_{\bar{X}} = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

5. Using the worksheet $n = 25, m = 500$, ***repeat*** the steps you did for $n = 10$ but now with $n = 25$. Use the 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. This is the same idea as what your textbook does in Figures 10.3 to 10.7 reviewed at the start of this module. Keeping the scale of the x-axis constant, even as the sampling distribution itself is less spread out, helps people visually notice the important concept that as the sample size goes up, sampling error goes down.

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

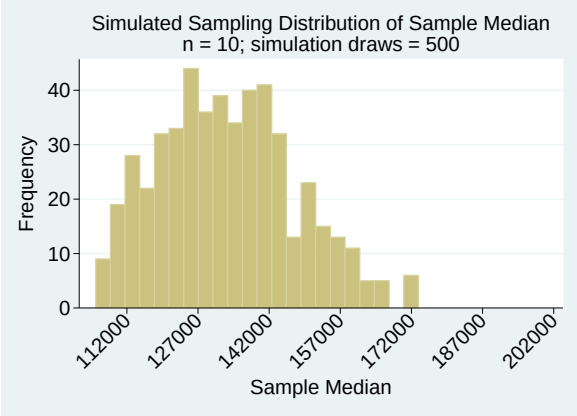




7. What about the sampling distribution of the *sample median*? In ECO220Y, we do not cover theoretical results for the *median*. Hence, simulation is particularly useful. ***Repeat*** the simulation you did for the sample mean for $n = 10$ with $m = 500$, but now for the sample median.

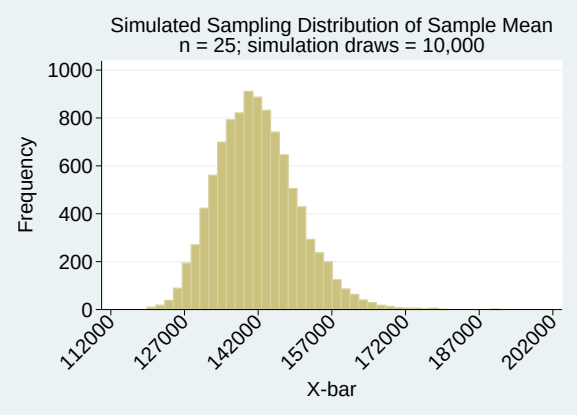
Verify that in this case, sampling error is larger for the sample median than $\bar{X}$ (i.e. that the simulated SD of the sample median is slightly larger than the simulated SD of the sample mean).

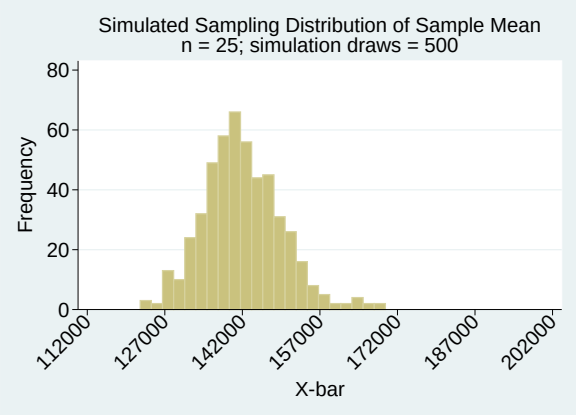
EXCEL TIPS: Go to the worksheet Median $n = 10, m = 500$. Using the MEDIAN function, compute the sample median of each of the 500 samples (with $n = 10$). Next, compute the mean and s.d. of the 500 sample medians.



8. Tutorial time permitting (otherwise, for homework), ***consider*** that so far you have varied the sample size from $n = 10$ to $n = 25$ to $n = 100$ for $\bar{X}$ and you compared $\bar{X}$ with the sample median for $n = 10$. 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 the sample mean for $n = 25$ and $m = 10,000$. ***Verify*** that it looks similar to the histogram below. ***Note*** that this more powerful simulation ($m = 10,000$) *is* telling the same story about the sampling distribution of the sample mean when $n = 25$ as the smaller ($m = 500$) simulation we did originally, which is shown again below for easy comparison.

EXCEL TIPS: In the column labeled X-bar, compute the sample mean for each row using the AVERAGE function. (Make sure not to include Column A, which is not a salary.) Worksheet $n = 25, m = 10000$ has been set-up to create a histogram using the same bins as earlier today to enable direct comparison. (Note: A larger number of simulation draws would ordinarily imply more bins, like shown in the Stata histogram.)

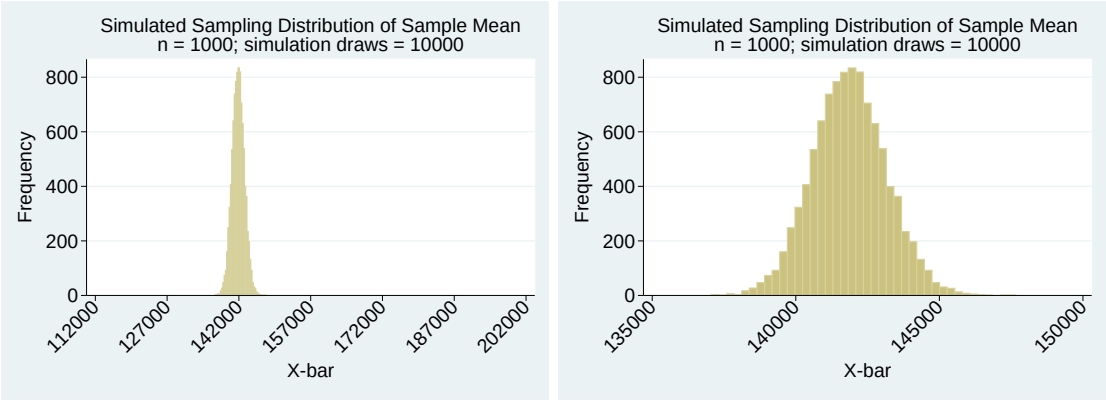




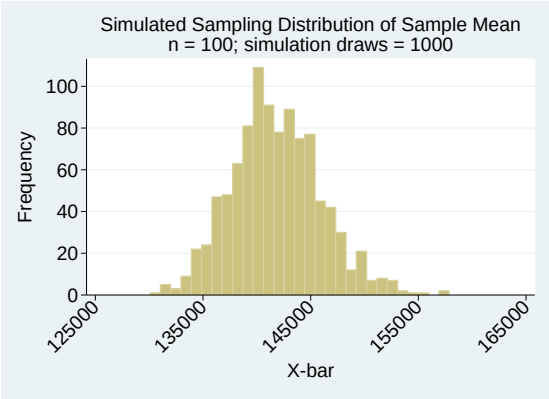
9. ***For homework,*** study the following points, which require reading, thinking, and synthesizing what you have learned in this Excel tutorial, but no further actions in Excel.

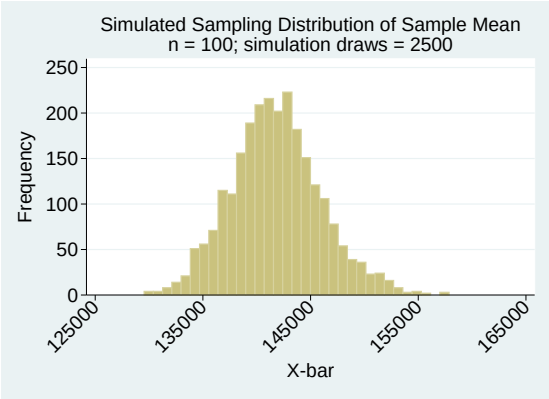
   (a) Recall that a simulation helps learn the *shape* of the sampling distribution of $\bar{X}$ when we cannot use the Central Limit Theorem (CLT). The CLT, a useful theoretical result about the shape of the sampling distribution of $\bar{X}$, requires a *sufficiently large* sample size. However, because the salary population is extremely skewed, today's tutorial 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 $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

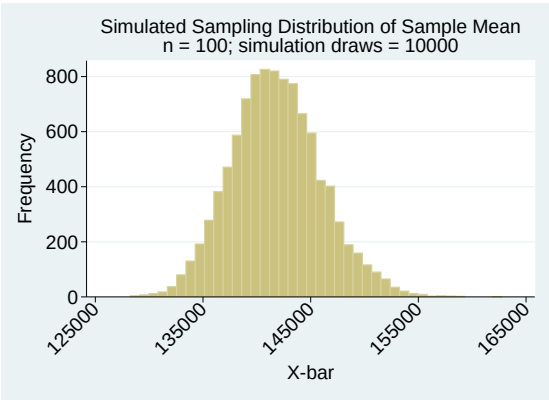
   (b) Consider $n = 1,000$. This simulation would be cumbersome in Excel. However, it helps complete the story of how the *shape* of the sampling distribution of $\bar{X}$ approaches Normal as the sample size increases. Even if the population is very skewed, a sample size of $n = 1,000$ is surely sufficiently 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 the positive skew in the population.

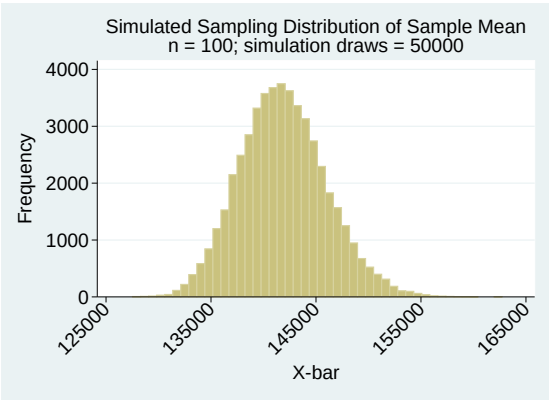


   (c) Consider further the important distinction between n (the sample size) and m (the number of samples drawn for the simulation). As seen in tutorial, changing n fundamentally changes the sampling distribution. In contrast, changing m simply affects how clearly we can see the sampling distribution. 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).

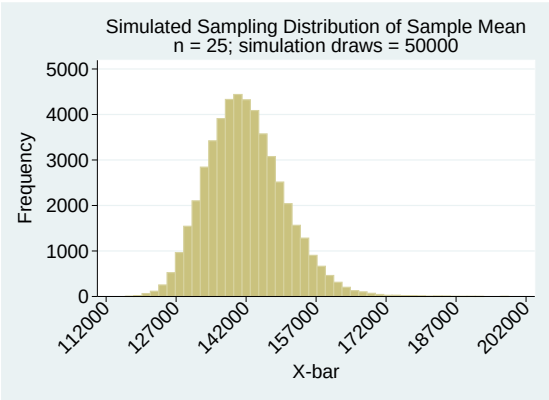


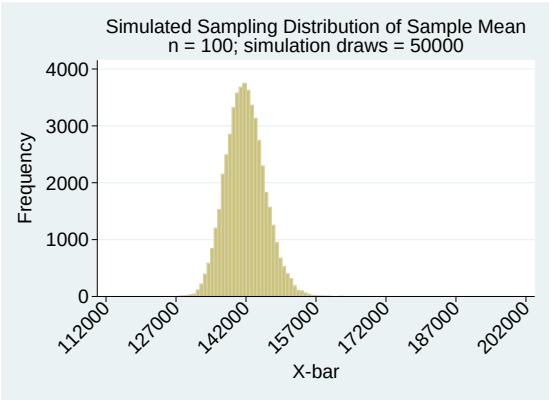






In contrast to the four graphs above, which only vary m and hence are very similar (i.e. different levels of focus but all the same giant panda), the graphs below show the dramatic difference in the sampling distribution when we change the sample size n (i.e. point the camera at a penguin instead of a giant panda).





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

Main course concepts: Analyzing categorical data with proportions and statistical inference.

Source materials (full citations in Section F): We will replicate parts of the analysis from an academic journal article "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment," which is abbreviated Karlan and List (2007).

Most relevant required readings: Chapter 11 including "Technology Help: Confidence Intervals for Proportions" on p. 355. 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 fund raising 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 fund raising.
- "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)*

			Match ratio		
	Control	Treatment	1:1	2:1	3:1
Implied price of $1 of public good:	1.00	0.36	0.50	0.33	0.25
Panel A	(1)	(2)	(3)	(4)	(5)
Response rate	0.018	0.022	0.021	0.023	0.023
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Dollars given, unconditional	0.813	0.967	0.937	1.026	0.938
	(0.063)	(0.049)	(0.089)	(0.089)	(0.077)
Dollars given, conditional on giving	45.540	43.872	45.143	45.337	41.252
	(2.397)	(1.549)	(3.099)	(2.725)	(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.

Additional readings (not required): Karlan and List (2007) is mostly written at an approachable level. In particular, see pp. 1774-1782, which includes the abstract, introduction, and sections entitled "I. Experimental Design," "A. Price Ratio," "B. Maximum Size of the Matching Grant," "C. Ask Amount," "D. Heterogeneous Treatment Effects," and the first portion of "II. Experimental Results."

Textbook case studies (extra practice): "Alberta Oil Sands" on p. 354

Datasets: For Karlan and List (2007): char_give.xlsx, where "char_give" abbreviates "Charitable Giving" from the title.

Interactive tutorial materials:

1. Open char_give.xlsx and ***browse*** the first two worksheets char_give and readme with the data and data description. Looking at the worksheet char_give, ***verify*** that the unit of observation of these cross-sectional data is a person (a prior donor) and that the sample size is 50,083.

2. Review the first results reported in Table 2A, Panel A, Columns (1) and (2).

   (a) ***Replicate*** the response rates of 0.018 and 0.022. ***Verify*** your precise answers are: $\frac{298}{16{,}687} = 0.017858213$ and $\frac{736}{33{,}396} = 0.02203857$.

   EXCEL TIPS: Without overwriting Columns A or B, copy the variables treatment and gave to worksheet CI Est. of Diff. in Resp. Rates. To make the results more readable, make two new variables to measure treatment and gave using word categories instead of 0's and 1's. For example, if Column D has the variable treatment, you can make a new variable named group with `=IF(D2=0,"Control","Treatment")`. Similarly, if Column E has the variable gave, you can make a new variable named response with `=IF(E2=0,"Did not give","Gave")`. Recall the shortcuts **Ctrl** + ↓ (+ **Shift**) discussed on page 12 (part 1b of Module A.1). Next, select the variables group and response and insert a pivot table in this worksheet. Drag group to COLUMNS and drag response to ROWS. From the pivot table fields environment, drag a second copy of response to Σ VALUES.

Count of response	Column Labels		
Row Labels	**Control**	**Treatment**	**Grand Total**
Did not give	16389	32660	49049
Gave	298	736	1034
Grand Total	**16687**	**33396**	**50083**

   Type the counts in the appropriate cells in CI Est. of Diff. in Resp. Rates. (You can make group 1 the control group and group 2 the treatment group.) Compute $\hat{P}$.

   (b) Using this formula from our aid sheets $\sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$, ***replicate*** the standard errors of 0.001 below the response rates for the control and treatment groups, which are displayed in the first results reported in Table 2A, Panel A, Columns (1) and (2). ***Verify*** your precise answers for each standard error (SE) are 0.00102522 and 0.00080335, respectively.

   EXCEL TIPS: Add formula in the appropriate cells in CI Est. of Diff. in Resp. Rates.

3. However, how much the match *changes* the response rate is the interesting question. This requires making an inference about the *difference* in proportions, not each proportion by itself (as in part 2). One method of statistical inference is confidence interval estimation (the other method is hypothesis testing). To make an inference about how much the match improves the response rate, ***recall*** this formula from our aid sheets, $(\hat{P}_2 - \hat{P}_1) \pm z_{\alpha/2}\sqrt{\frac{\hat{P}_2(1-\hat{P}_2)}{n_2} + \frac{\hat{P}_1(1-\hat{P}_1)}{n_1}}$.

   (a) ***Compute*** the standard error (SE) of the difference in response rates between the treatment group and control group. (Recall that the SE is the square root term in the formula.) ***Verify*** that you obtain 0.00130248.

EXCEL TIPS: Continue using the worksheet CI Est. of Diff. in Resp. Rates as a template, adding formulas in the appropriate cells.

(b) ***Compute*** the 95% confidence interval estimate of the difference in the response rate between the control group and treatment group. ***Verify*** that you obtain 0.00418035 as the point estimate of the difference and 0.00255281 as the margin of error (ME). Similarly, ***verify*** that you obtain $[0.00162755, 0.00673316]$ as the lower confidence limit (LCL) and upper confidence limit (UCL). (Be careful with computing the term $z_{\alpha/2}$: you must remember that confidence intervals are *two tailed*, which is why α is divided by two.)

EXCEL TIPS: Continue using the worksheet CI Est. of Diff. in Resp. Rates. Use the NORM.S.INV function to obtain $z_{\alpha/2}$. In the cell for alpha, type 0.05 to specify $\alpha = 0.05$. You may either use NORM.S.INV(1 - 0.05/2) or ABS(NORM.S.INV(0.05/2)).

4. Next, consider Panels B and C in Table 2A, which check for "heterogeneous treatment effects."

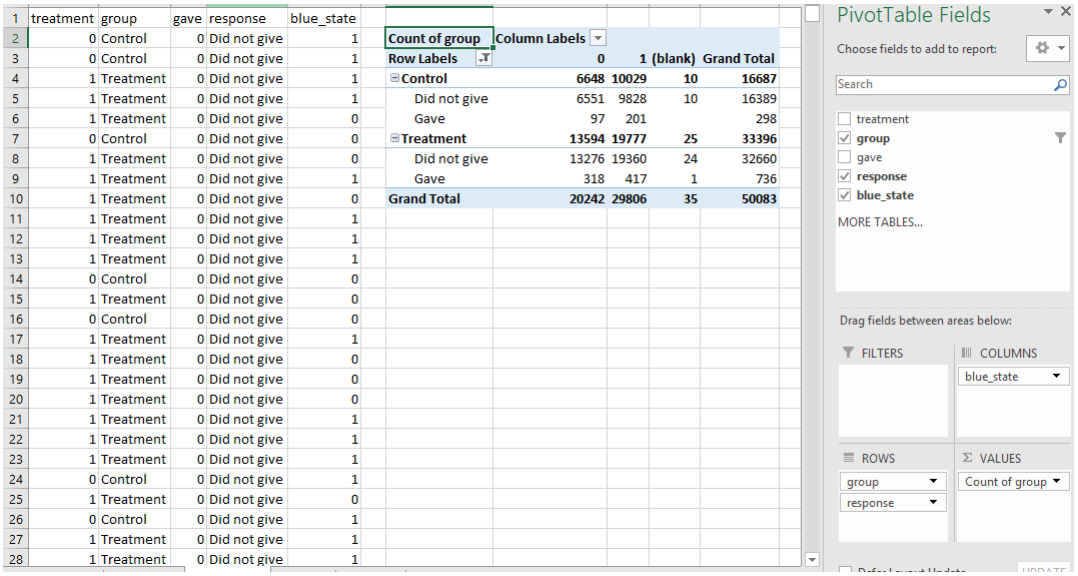
TABLE 2A—MEAN RESPONSES
(Mean and standard errors)

			Match ratio		
	Control	Treatment	1:1	2:1	3:1
Implied price of $1 of public good:	1.00	0.36	0.50	0.33	0.25
Panel A	(1)	(2)	(3)	(4)	(5)
Response rate	0.018	0.022	0.021	0.023	0.023
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Dollars given, unconditional	0.813	0.967	0.937	1.026	0.938
	(0.063)	(0.049)	(0.089)	(0.089)	(0.077)
Dollars given, conditional on giving	45.540	43.872	45.143	45.337	41.252
	(2.397)	(1.549)	(3.099)	(2.725)	(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
Panel B: Blue states					
Response rate	0.020	0.021	0.021	0.022	0.021
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)
Dollars given, unconditional	0.897	0.895	0.885	0.974	0.826
	(0.086)	(0.059)	(0.102)	(0.110)	(0.091)
Dollars given, conditional on giving	44.781	42.444	42.847	44.748	39.635
	(2.914)	(1.866)	(3.356)	(3.456)	(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
Panel C: Red states					
Response rate	0.015	0.023	0.021	0.024	0.026
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)
Dollars given, unconditional	0.687	1.064	0.987	1.103	1.101
	(0.093)	(0.085)	(0.157)	(0.148)	(0.135)
Dollars given, conditional on giving	47.113	45.490	47.667	46.110	43.161
	(4.232)	(2.607)	(5.848)	(4.392)	(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.

(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 the variable blue_state into the worksheet CI Est. of Diff. in Resp. Rates. Create a pivot table with the variables: group, response, and blue_state. One way to do it: put the variables group and response under ROWS, blue_state under COLUMNS, and drag another copy of group to Σ VALUES.



	treatment	group	gave	response	blue_state
2	0	Control	0	Did not give	1
3	0	Control	0	Did not give	1
4	1	Treatment	0	Did not give	1
5	1	Treatment	0	Did not give	1
6	1	Treatment	0	Did not give	0
7	0	Control	0	Did not give	0
8	1	Treatment	0	Did not give	0
9	1	Treatment	0	Did not give	1
10	1	Treatment	0	Did not give	0
11	1	Treatment	0	Did not give	1
12	1	Treatment	0	Did not give	1
13	1	Treatment	0	Did not give	1
14	0	Control	0	Did not give	0
15	1	Treatment	0	Did not give	0
16	0	Control	0	Did not give	0
17	1	Treatment	0	Did not give	1
18	1	Treatment	0	Did not give	0
19	1	Treatment	0	Did not give	0
20	1	Treatment	0	Did not give	0
21	1	Treatment	0	Did not give	1
22	1	Treatment	0	Did not give	1
23	1	Treatment	0	Did not give	1
24	0	Control	0	Did not give	1
25	1	Treatment	0	Did not give	0
26	0	Control	0	Did not give	1
27	1	Treatment	0	Did not give	1
28	1	Treatment	0	Did not give	1

Count of group	Column Labels			
Row Labels	0	1	(blank)	Grand Total
Control	6648	10029	10	16687
Did not give	6551	9828	10	16389
Gave	97	201		298
Treatment	13594	19777	25	33396
Did not give	13276	19360	24	32660
Gave	318	417	1	736
Grand Total	20242	29806	35	50083

(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 0.00099266 with SE 0.00173192 and ME 0.00446114, which yields the 99% CI estimate $[-0.00341974, 0.00550618]$.

EXCEL TIPS: Using the pivot table results from part 4a, input values into Column B in the worksheet CI Est. of Diff. in Resp. Rates, which makes this step just a matter of changing the input values (X_1, n_1, X_2, n_2, and α).

(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 $[0.00557719, 0.01202645]$.

EXCEL TIPS: Using the pivot table results from part 4a (recalling that blue_state equal to zero is the same as red_state equal to one), input values into Column B in the worksheet CI Est. of Diff. in Resp. Rates, which makes this step just a matter of changing the input values (X_1, n_1, X_2, n_2, and α).

(d) ***For homework,*** consider what Karlan and List (2007) mean by "heterogeneous treatment effects." While this term may sound intimidating, it just means that the effectiveness of these schemes may vary across types of people. "Hetero" means different. You just checked if people in Red states responded differently to the treatments (the letters offering a match) compared to people in Blue states. You found some evidence that they did. This is an important issue in marketing: the effectiveness of marketing strategies often varies across groups (e.g. younger people versus older people, females versus males, etc.).

C.3 Practice test questions for Module C

QUESTIONS:

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. Sample statistics are computed as usual (for example, $\bar{X} = \frac{\sum_{i=1}^{30} x_i}{30}$). As another example, the last row of data in Simulation 2 shows the ten-thousandth random sample of 30 observations drawn from Population 2.

(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?

(e) Using worksheet Simulation 2, make an inference about the shape of Population 2.

(f) 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 $\bar{X}$?

(g) Repeat the previous part, but with a sample size of 30 ($n = 30$). How does this change your answers to the previous part?

(h) 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?

Q2. Recall ON Univ. & Col. (2016). For all subparts, use the worksheets 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 people do not visually detect skew. Consider the simulated sampling distribution of $\bar{X}$ for $n = 16$ from the previous part. The following 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?

(c) 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?

(d) According to the simulation results, which of the two sample statistics is less affected by sampling error as judged by the standard deviation? As judged by the range?

Q3. Recall the dice example (Section 10.3, which Module C.1 reviewed). Use mod_c_dice_roll.xlsx.

(a) 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?

Q4. 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?

Q5. 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)
Member activity			
Number of months since last donation	13.007	13.012	12.998
	(12.081)	(12.086)	(12.074)
Highest previous contribution	59.385	59.597	58.960
	(71.177)	(73.052)	(67.269)
Number of prior donations	8.039	8.035	8.047
	(11.394)	(11.390)	(11.404)
Number of years since initial donation	6.098	6.078	6.136
	(5.503)	(5.442)	(5.625)
Percent already donated in 2005	0.523	0.523	0.524
	(0.499)	(0.499)	(0.499)
Female	0.278	0.275	0.283
	(0.448)	(0.447)	(0.450)
Couple	0.092	0.091	0.093
	(0.289)	(0.288)	(0.290)
... [Part of the table has been excluded] ...			
State and county			
Red state–proportion that live in red state	0.404	0.407	0.399
	(0.491)	(0.491)	(0.490)
Red county–proportion that live in red county	0.510	0.512	0.507
	(0.500)	(0.500)	(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 (_____).

Q6. 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.

Q7. 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
Panel A	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Response rate	0.018	0.022	0.022	0.022	0.022	0.021	0.022	0.023
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)
Dollars given, unconditional	0.813	1.060	0.889	0.903	1.015	0.914	1.004	0.983
	(0.063)	(0.109)	(0.091)	(0.084)	(0.106)	(0.080)	(0.091)	(0.084)
Dollars given, conditional on giving	45.540	49.172	39.674	41.000	45.815	43.107	45.239	43.251
	(2.397)	(3.522)	(2.900)	(2.336)	(3.475)	(2.557)	(2.932)	(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
Panel B: Blue states								
Response rate	0.020	0.020	0.022	0.022	0.020	0.019	0.022	0.022
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Dollars given, unconditional	0.897	0.884	0.912	0.900	0.884	0.796	0.950	0.939
	(0.086)	(0.115)	(0.127)	(0.110)	(0.116)	(0.094)	(0.108)	(0.102)
Dollars given, conditional on giving	44.781	43.204	41.091	41.236	44.469	41.516	43.194	42.503
	(2.914)	(3.716)	(4.227)	(3.093)	(3.806)	(3.283)	(3.364)	(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
Panel C: Red states								
Response rate	0.015	0.023	0.023	0.022	0.025	0.024	0.022	0.024
	(0.001)	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)
Dollars given, unconditional	0.687	1.330	0.856	0.874	1.206	1.086	1.082	1.023
	(0.093)	(0.212)	(0.127)	(0.124)	(0.199)	(0.141)	(0.158)	(0.141)
Dollars given, conditional on giving	47.113	57.156	37.649	39.584	47.330	44.929	48.097	43.519
	(4.232)	(6.485)	(3.643)	(3.462)	(6.039)	(4.005)	(5.234)	(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

(a) 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.

(b) 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.

(c) 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

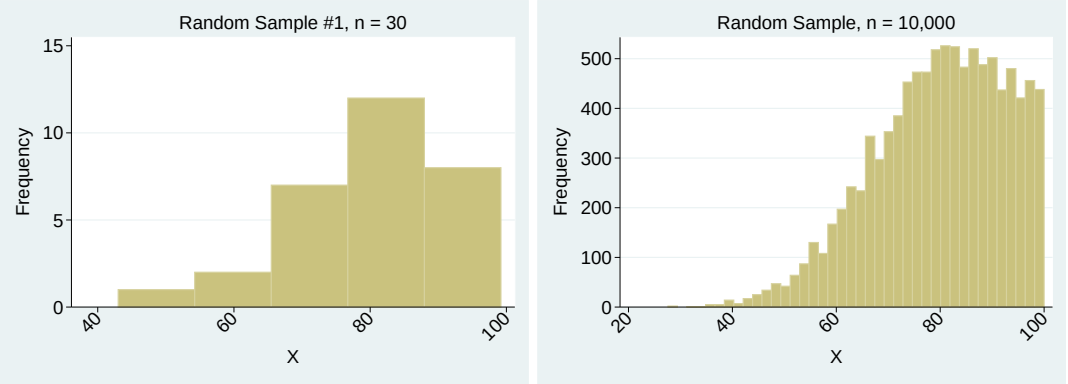
donation). Report the point estimate of the difference, the standard error of the difference, the margin of error of the difference, and the LCL and UCL.

Q8. Recall ON Univ. & Col. (2016). Use the worksheet Drawing random samples in on_univ_col_16.xlsx. A method of selecting random samples discussed in tutorials 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.

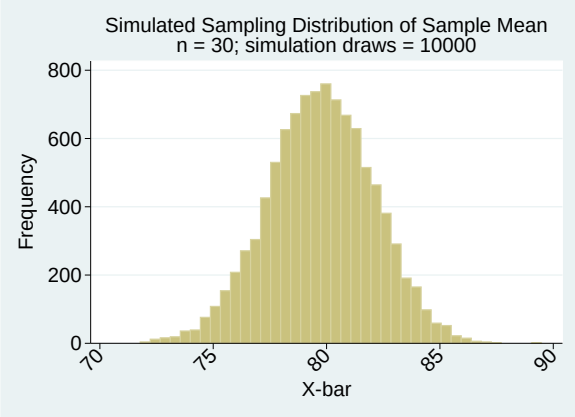
Draw #	Sample Mean
1	
2	
3	
...	...
9,999	
10,000	

ANSWERS:

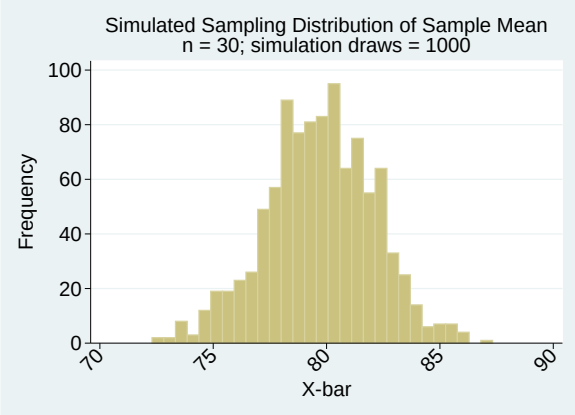
A1. (a) You could make a rough inference using Sample #1. However, in worksheet Simulation 1 we have 300,000 (=30*10,000) observations randomly drawn from Population 1 so it is easy to make an inference about that population using this very large sample. In fact, we do not need to use all 300,000 observations. For ease, we can use the first 10,000 draws from the population (i.e. $x1$). We can confidently infer with this large sample and the clear pattern in the histogram that the population is negatively skewed.



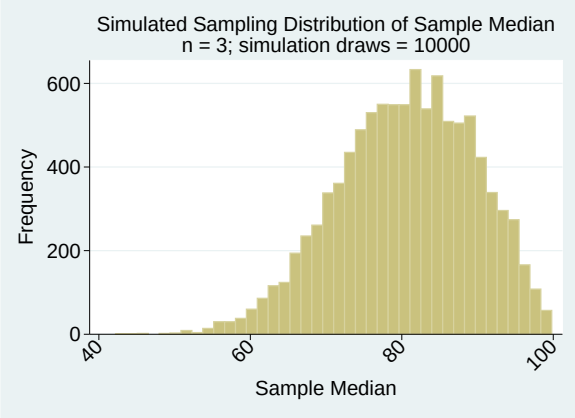
(b) The mean of the 10,000 sample means is 79.7475 and the s.d. of the 10,000 sample means is 2.327985. The shape looks Normal (see histogram below).



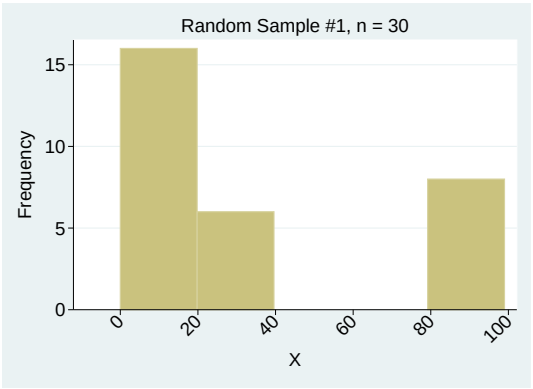
(c) It does not change our answers in any meaningful way. The shape looks the same and the mean and s.d. of the sample mean are comparable. The mean of the 1,000 sample means is 79.71958 and the s.d. of the 1,000 sample means is 2.383696.

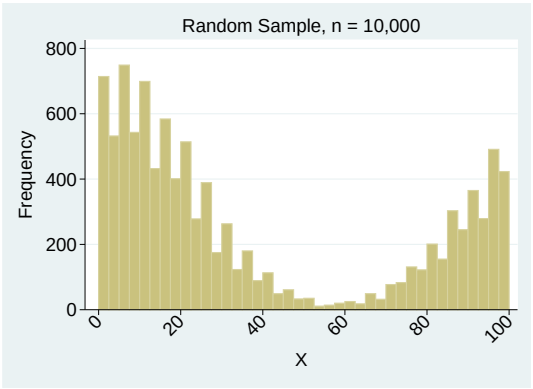


(d) The mean of the 10,000 sample medians is 80.54002 and the s.d. of the 10,000 sample medians is 9.037415. The shape is somewhat negatively skewed (see histogram below).

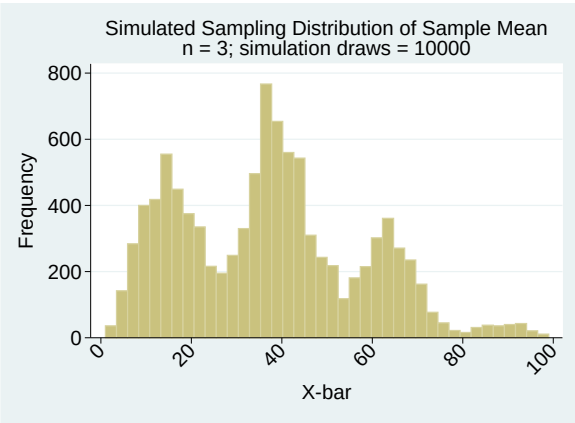


(e) You could make a rough inference using Sample #1. However, in worksheet Simulation 2 we have 300,000 (=30*10,000) observations randomly drawn from Population 2 so it is easy to make an inference about that population using this very large sample. In fact, we do not need to use all 300,000 observations. For ease, we can use the first 10,000 draws from the population (i.e. $x1$). We can confidently infer with this large sample and the clear pattern in the histogram that the population is bimodal.



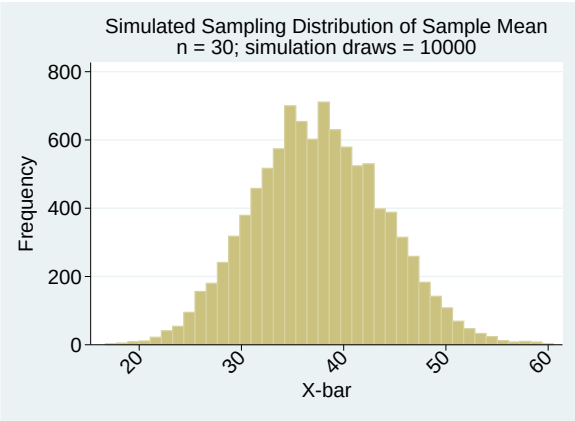


(f) The mean of the 10,000 sample means is 37.22233 and the s.d. of the 10,000 sample means is 19.99509. The shape is unusual: four modes!

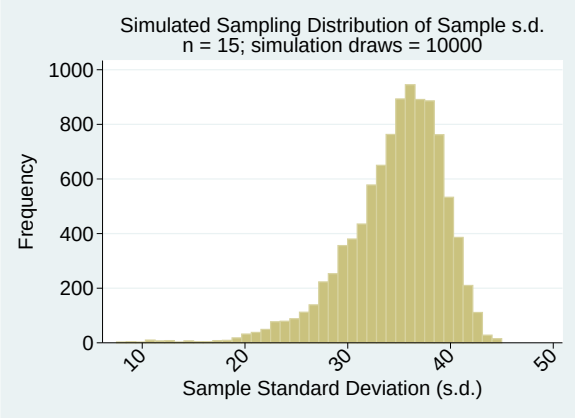


(g) The mean of the 10,000 sample means is 37.56397 and the s.d. of the 10,000 sample means is 6.394322. The shape looks almost Normal (see histogram below): there are still some

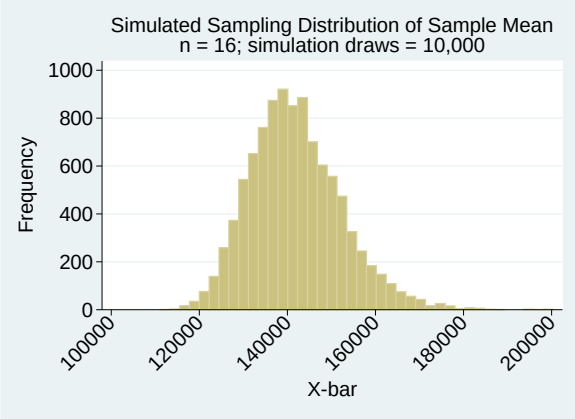
hints of modality but we've definitely gotten a lot closer to Normal with this lager sample size.



(h) 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).



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. =NORM.S.DIST(1,TRUE)-NORM.S.DIST(0,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

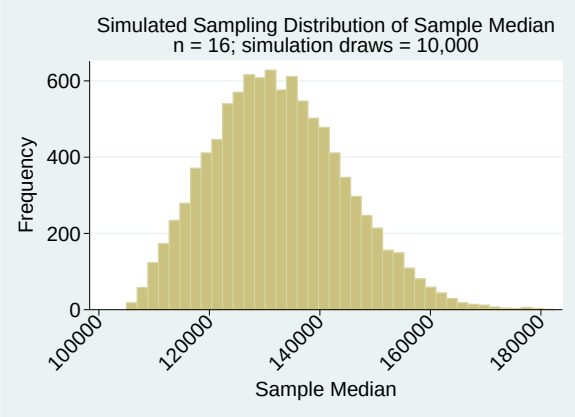
iii. =NORM.S.DIST(2,TRUE)-NORM.S.DIST(1,TRUE) = 0.135905122 (Given the symmetry of the Normal distribution: $P(-2 < Z < -1) = P(1 < Z < 2)$.)

iv. 0.1434 between -2 and -1 and 0.1157 between 1 and 2

v. =NORM.S.DIST(3,TRUE)-NORM.S.DIST(2,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

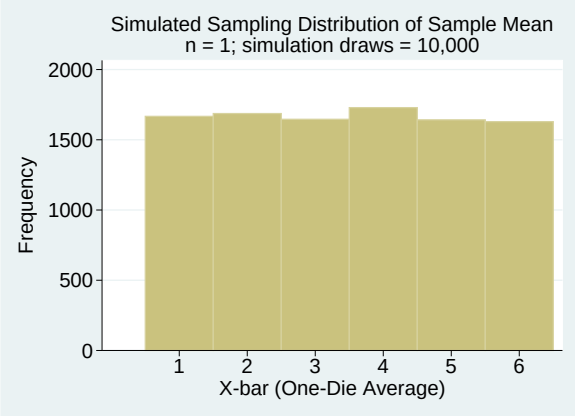
(c) The mean of the 10,000 sample medians is 132538.9 and the s.d. of the 10,000 sample medians is 12068.4. The shape is somewhat positively skewed.



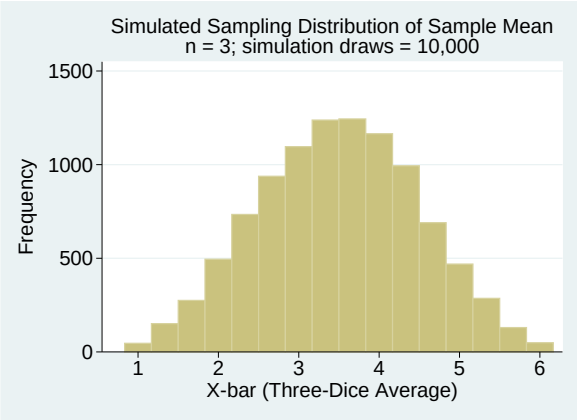
(d) The sample mean has a smaller s.d. than the sample median: by this metric the sample mean is affected less by sampling error. However, the sample mean has a bigger range given the possibility of having a very highly paid person in the sample (which drives up the mean but does not impact the median).

A3. (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).

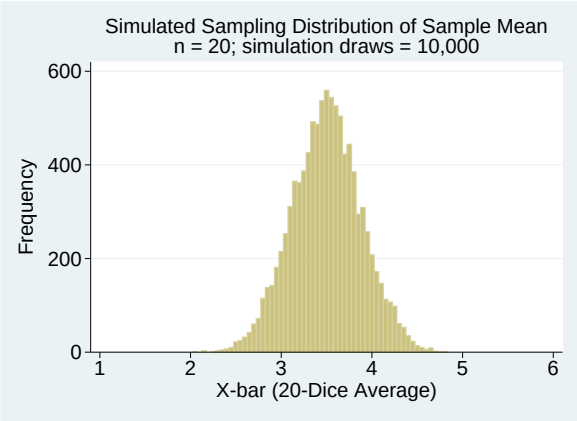
(b) Six unique values of $\bar{X}$ occur in the simulation. A value between 3 and 4 (including those endpoints) occurs 3,375 times.



(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.



A4. Table C.1 is easily replicated with a pivot table 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**

A5. (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.)

A6. 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.)

A7. (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.

A8. You should get the exact same result as the first row below. (Your other two rows cannot be compared against a fixed solution.)

Draw #	Sample Mean
1	$141,233.3

D Module D: Interactive Tutorial Materials & Test Prep

D.1 Module D.1: Hypothesis Testing (Part 1 of 2)

Main course concepts: Building skills in hypothesis testing. Hypothesis testing for comparing two proportions. Hypothesis testing for inference about a single mean.

Source materials (full citations in Section F): We will replicate parts of the analysis from an academic journal article "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment," abbreviated Karlan and List (2007). We will also work with the "Sparton Resources of Toronto" case study.

Most relevant required readings: Chapters 12 and 13. Also, recall Karlan and List (2007) from Module C.2 that reproduced the complete Table 2A. Review this excerpt of Table 2A.

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

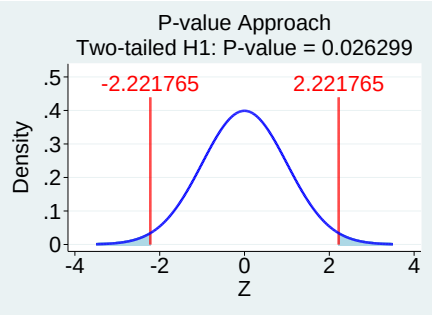
			Match ratio		
	Control	Treatment	1:1	2:1	3:1
Implied price of $1 of public good:	1.00	0.36	0.50	0.33	0.25
Panel A	(1)	(2)	(3)	(4)	(5)
Response rate	0.018	0.022	0.021	0.023	0.023
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Dollars given, unconditional	0.813	0.967	0.937	1.026	0.938
	(0.063)	(0.049)	(0.089)	(0.089)	(0.077)
Dollars given, conditional on giving	45.540	43.872	45.143	45.337	41.252
	(2.397)	(1.549)	(3.099)	(2.725)	(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
Panel B: Blue states					
Response rate	0.020	0.021	0.021	0.022	0.021
	(0.001)	(0.001)	(0.002)	(0.002)	(0.002)

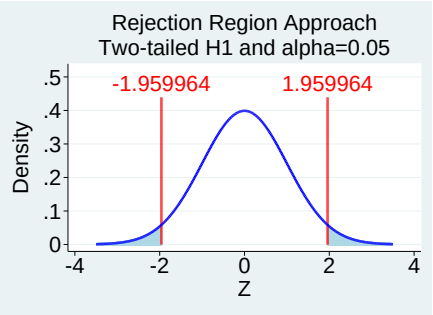
Figure of Table 2A: Karlan and List (2007), p. 1781, Panel A and first row of Panel B only.

- 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? If so, at which conventional significance levels? Before diving into the data analysis to answer, it is useful to quickly refresh yourself on the concepts and background you need to remember.
  - Conventional significance levels are α equals 0.01, 0.05, or 0.10.
  - Translating the question into formal hypotheses and standard notation: $H_0 : (p_T - p_C) = 0$ 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 the null hypothesis says there is no difference between the population proportions and hypothesis testing starts with the presumption that the null is true, it makes sense to

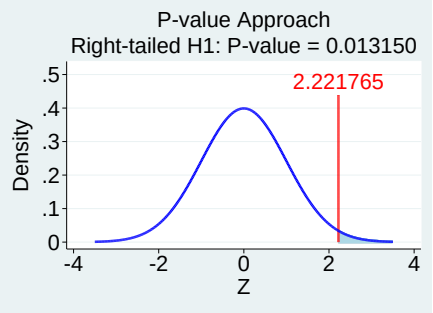
pool the two samples together to get a single estimate ($\bar{P}$) when doing *hypothesis testing*. This argument does *not* translate to confidence interval estimation.

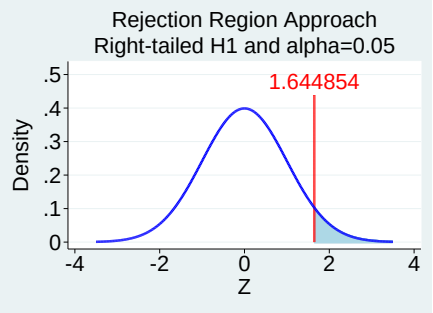
- There are two methods of hypothesis testing: the P-value approach and the rejection region (aka critical value) approach.
  - With the P-value approach, you use the test statistic and the direction of the research hypothesis ($>$, $<$, $\neq$) to properly compute a probability called the P-value. The smaller the P-value, the stronger the evidence in favor of the research hypothesis (i.e. the stronger the evidence *against* the null hypothesis). If the P-value is smaller than α, we have a statistically significant result at that significance level.
  - With the rejection region approach, you use the significance level (α) and the direction of the research hypothesis ($>$, $<$, $\neq$) to compute the appropriate critical value (edge of the rejection region). You then compare your test statistic with the critical value.
- For example, test $H_0 : (p_T - p_C) = 0$ versus $H_1 : (p_T - p_C) \neq 0$ if there are 1,000 people in the treatment group and 80 give a donation and there are 3,000 people in the control group and 180 give a donation. $\bar{P} = \frac{80+180}{1{,}000+3{,}000} = 0.065$ 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 graphs below illustrate both approaches to hypothesis testing. With the P-value approach (left graph) we obtain a P-value of 0.026299, which is quite small and provides good support for our research hypothesis that the proportion donating are different between the control and treatment groups. This P-value is less than 0.05, but bigger than 0.01, which means that the best conventional significance level we meet is $\alpha = 0.05$. With the rejection region approach (right graph) and choosing $\alpha = 0.05$, we obtain critical values of -1.959964 and 1.959964. Our z test statistic of 2.221765 falls in the rejection region and we conclude there is statistically significant difference in the proportion donating between the control and treatment groups at a 5% significance level.





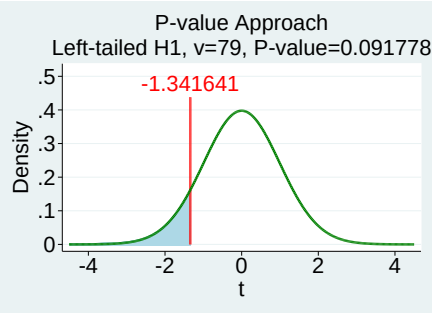
- Continuing the numeric example, what if we wished to test $H_0 : (p_T - p_C) = 0$ versus $H_1 : (p_T - p_C) > 0$. This is a one-tailed test. Recall how this affects both approaches.

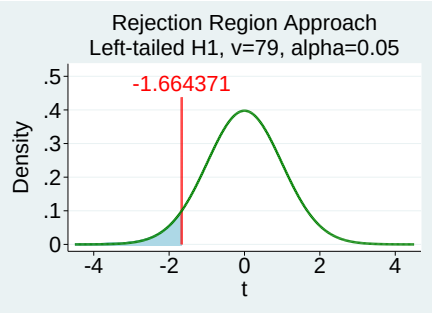




- Review the second and third rows of results in Panel A, Columns (1) and (2) in Table 2A: "Dollars given, unconditional" and "Dollars given, conditional on giving." "Unconditional" means using all observations, including those who gave zero, when computing the mean. "Conditional on giving" means ignoring the zeros when computing the mean (i.e. it is the mean amongst those who actually donated). Note that amount given is an interval (not categorical) variable. Consider an example question: Is the mean dollars given, conditional on giving, for the treatment group statistically significantly lower than \$50? If so, at which conventional significance levels? Again, before diving into the data analysis to answer, it is useful to quickly refresh yourself on the concepts and background you need to remember.

  - Translating the question into formal hypotheses and standard notation: $H_0 : \mu_{T|G} = 50$ 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.
  - Like for inferences about proportions, there are two methods of hypothesis testing for making inferences about means. (These concepts are already reviewed above for proportions.)
    * P-value approach
    * Rejection region (aka critical value) approach
  - For example, test $H_0 : \mu_{T|G} = 50$ versus $H_1 : \mu_{T|G} < 50$ if there are 1,000 people in the treatment group and 80 give a donation and for those 80 the average donation is \$48.50 with a standard deviation of \$10.00. $t = \frac{48.50-50}{10/\sqrt{80}} = -1.341641$ and $\nu = 79$. The graphs below illustrate both approaches to hypothesis testing. With the P-value approach (left graph) we obtain a P-value of 0.091778, which is somewhat small and provides some support for our 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 $\alpha = 0.10$. With the rejection region approach (right graph) and choosing $\alpha = 0.05$, we obtain a critical value of -1.664371. Our t test statistic of -1.341641 does not fall in the rejection region and we conclude that we have insufficient proof of H_1 at a 5% significance level.





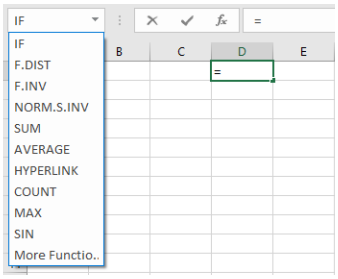
Textbook case studies (extra practice): "Sparton Resources of Toronto" on p. 430

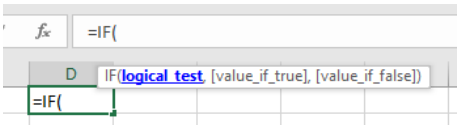
Datasets: For Karlan and List (2007): char_give.xlsx; For Sparton Resources: sparton.xlsx

Interactive tutorial materials:

1. For Karlan and List (2007), use char_give.xlsx. Note that Table 2A slices and dices the data many ways. In char_give.xlsx, some important variables are coded as dummy variables (0's and 1's), which is often useful. However, sometimes it is more useful to have nominal (aka categorical) variables coded as words. Create new variables for treatment, gave, and blue_state that record the category for each observation as a string (text).

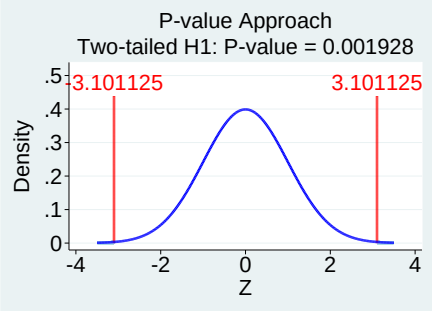
   EXCEL TIPS: Excel has many useful functions. If you type "=" in any cell, you can easily search the functions (see below, left). For this part, the IF function is perfect (see below, right).

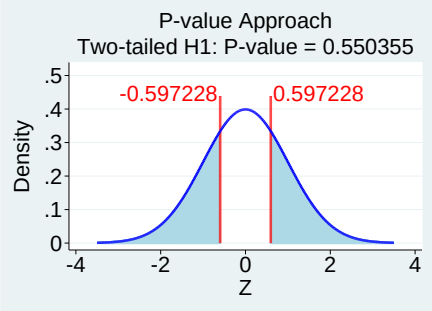




   If Column A has the variable treatment, you can make a new variable named group with the function `=IF(A2=0,"Control","Treatment")`. Similarly, if Column G has the variable gave, make a new variable named give_status with the function `=IF(G2=0,"Did not give","Gave")`. The blue_state variable is tricky 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 Q2 is 1 it evaluates to "Blue State." If Q2 is not 1, it evaluates to "Missing" if Q2 is blank and to "Red State" otherwise. Note the use of the logical function ISBLANK, which returns a value of TRUE if Q2 is blank and FALSE otherwise.

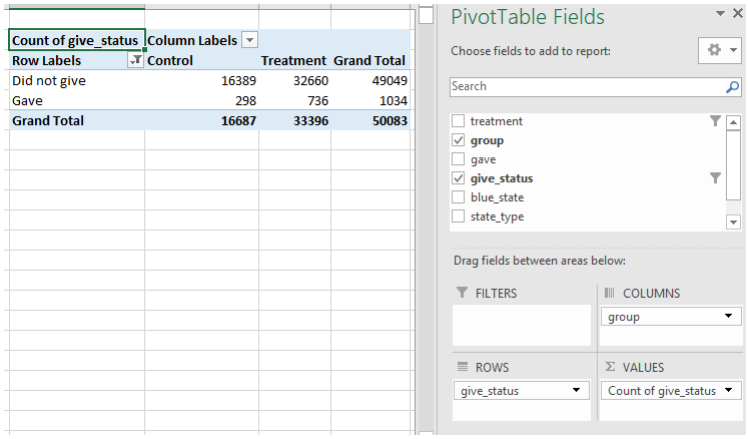
2. Consider the hypothesis test $H_0 : (p_T - p_C) = 0$ versus $H_1 : (p_T - p_C) \neq 0$ associated with the first row of results in Panels A and B, Columns (1) and (2) in Table 2A. These figures visually display the answers to parts 2a and 2b.





   (a) Start with the first results in Panel A. ***Compute*** the P-value for the hypothesis test given at the start of part 1. (Recall the preamble review at the start of this module.) ***Verify*** that you obtain: $\bar{P} = 0.02064573$, $(\hat{P}_T - \hat{P}_C) = 0.00418035$, a SE of the difference presuming the null is true of 0.00134801, a z test statistic of 3.1011251, and a P-value of 0.00192787. (Remember this is a *two-tailed* test.)

EXCEL TIPS: Use the worksheet HT Diff. in Resp. Rates as a template. To obtain the necessary inputs for the proportions, insert a pivot table (see below). Next, use the NORM.S.DIST function. It has two inputs: the z value and a logical value: TRUE or FALSE. We always use TRUE, which returns the cumulative area under the Normal density function. For example, NORM.S.DIST(-1.96, TRUE) returns the value 0.024997895: in other words, $P(Z < -1.96) = 0.024997895$.



Count of give_status	Column Labels		
Row Labels	Control	Treatment	Grand Total
Did not give	16389	32660	49049
Gave	298	736	1034
Grand Total	16687	33396	50083

i. ***For homework, recall*** that a P-value of 0.0019 means this difference is statistically significant at all conventional significance levels, including a 1% level.

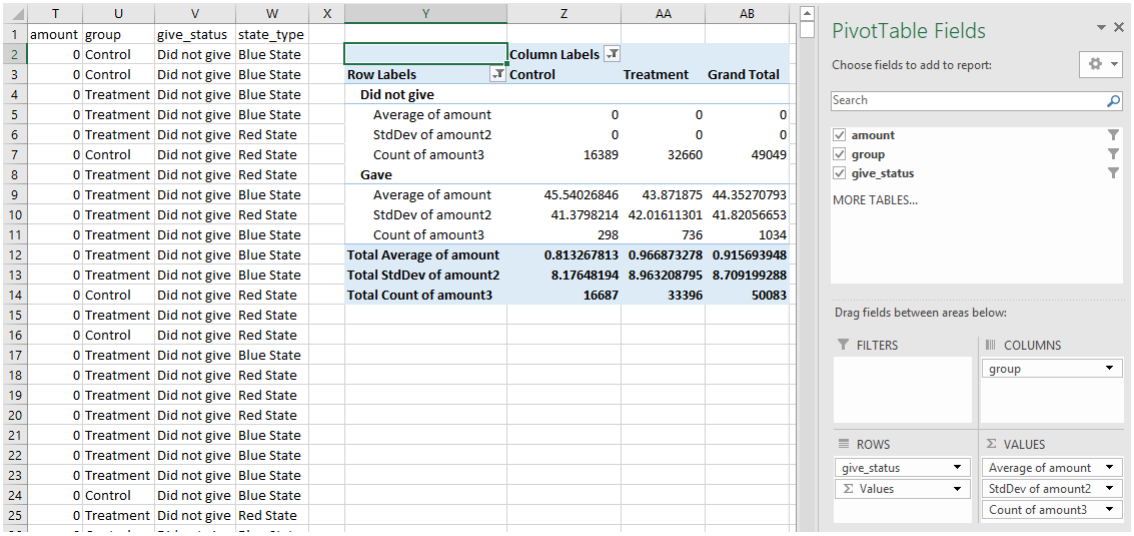
(b) Next, assess the first results in Panel B. In Blue states, is there a statistically significant difference in the response rates between the control and treatment group? If so, at which conventional significance levels? In answering, ***compute*** the P-value. ***Verify*** that you obtain: $\bar{P} = 0.02073408$, $(\hat{P}_T - \hat{P}_C) = 0.00104322$, 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 to answer the question about Blue states. Again, to obtain the necessary inputs for the proportions, insert a pivot table. (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.)

i. ***For homework, recall*** that a P-value of 0.5504 means this difference is not statistically significant at any conventional significance level, including a 10% level.

3. Next, you will replicate parts of Table 2A related to the dollars given: the second and third rows of results in Panel A of Table 2A for the control group and treatment group.

EXCEL TIPS: A single pivot table can do nearly all of the work for this replication. Insert a pivot table 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. In the pivot table fields environment, unselect blanks as there are no missing values for any of these three variables. From the pivot table fields environment, drag a second and third copy of the variable amount to Σ VALUES and change the field settings to show the sample mean for one and the sample standard deviation for the other. To make the chart even more readable, drag the Σ Values field to rows instead of columns. (See the screenshot on the next page that illustrates these steps.)



	T	U	V	W	X	Y	Z	AA	AB
1	amount	group	give_status	state_type					
2	0	Control	Did not give	Blue State			Column Labels		
3	0	Control	Did not give	Blue State		Row Labels	Control	Treatment	Grand Total
4	0	Treatment	Did not give	Blue State		Did not give			
5	0	Treatment	Did not give	Blue State		Average of amount	0	0	0
6	0	Treatment	Did not give	Red State		StdDev of amount2	0	0	0
7	0	Control	Did not give	Red State		Count of amount3	16389	32660	49049
8	0	Treatment	Did not give	Red State		Gave			
9	0	Treatment	Did not give	Blue State		Average of amount	45.54026846	43.871875	44.35270793
10	0	Treatment	Did not give	Red State		StdDev of amount2	41.3798214	42.01611301	41.82056653
11	0	Treatment	Did not give	Blue State		Count of amount3	298	736	1034
12	0	Treatment	Did not give	Blue State		Total Average of amount	0.813267813	0.966873278	0.915693948
13	0	Treatment	Did not give	Blue State		Total StdDev of amount2	8.17648194	8.963208795	8.709199288
14	0	Control	Did not give	Red State		Total Count of amount3	16687	33396	50083
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					
21	0	Treatment	Did not give	Blue State					
22	0	Treatment	Did not give	Blue State					
23	0	Treatment	Did not give	Blue State					
24	0	Control	Did not give	Blue State					
25	0	Treatment	Did not give	Red State					

(a) ***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* (note this does *not* say standard error) for the control group and ***verify*** it is s = \$8.176482.

EXCEL TIPS: You just need to correctly read values from the pivot table you produced.

- ***For homework,*** replicate the standard error of \$0.063 using Excel to program this formula from our aid sheets: $\frac{s}{\sqrt{n}}$.

(b) ***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.

EXCEL TIPS: Again, just find the appropriate values in the pivot table.

- ***For homework,*** replicate the the standard error of \$2.397.

(c) ***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.

- ***For homework,*** replicate the standard error of \$0.049.

(d) ***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.

- ***For homework,*** replicate the standard error of \$1.549.

4. ***Recall*** the "Sparton Resources" case study on p. 430. For a source of coal ash to be economically profitable the concentration of uranium oxide must be **at least 0.32 pounds per tonne**. Use sparton.xlsx for all subparts, which includes the data and a template for your work.

(a) For Location 1, describe the uranium oxide concentrations in those randomly selected batches of coal ash by ***computing*** some basic summary statistics: the sample size (n), the sample mean ($\bar{X}$), and the sample standard deviation (s). ***Verify*** that you obtain 10, 0.325, and 0.204246583, respectively.

EXCEL TIPS: Use the Excel functions COUNT, AVERAGE, and STDEV.S.

(b) For Location 1, ***compute*** the standard error of the mean. ***Verify*** that you obtain 0.064588441.

EXCEL TIPS: Program in the formula $\frac{s}{\sqrt{n}}$ referencing your cells with s and n.

(c) For Location 1, ***fill in*** the value specified in the null hypothesis ($H_0 : \mu = 0.32$), ***compute*** the value of the t test statistic using this formula from our aid sheets: $t = \frac{\bar{X}-\mu_0}{\frac{s}{\sqrt{n}}}$. ***Verify*** that you obtain 0.077413233.

(d) For Location 1, ***consider*** the hypothesis test: $H_0 : \mu_1 = 0.32$ versus $H_1 : \mu_1 > 0.32$, where μ_1 refers to the concentration of uranium oxide for *all* coal ash produced by Location 1. If Sparton Resources wishes only to source from locations that have been *proven profitable*, then it makes sense to set up the research hypothesis as a right-tailed test. ***Compute*** the P-value. ***Verify*** that you obtain 0.46999426.

EXCEL TIPS: Use the Excel function T.DIST. While Excel also has functions T.DIST.RT and T.DIST.2T, in addition to T.DIST, *do not* use those. T.DIST.2T gives an error message if your test statistic is negative and T.DIST.RT is just equal to (1 - T.DIST), which makes it unnecessary.

(e) Tutorial time permitting (otherwise for homework), ***consider***: $H_0 : \mu_1 = 0.32$ versus $H_1 : \mu_1 < 0.32$. If Sparton Resources wishes to source from all locations except those *proven unprofitable*, then it makes sense to set up the research hypothesis as a left-tailed test. ***Compute*** the P-value. ***Verify*** that you obtain 0.53000574.

(f) Tutorial time permitting (otherwise for homework), ***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 for maximizing profits (we care if sources are profitable or unprofitable, not if they are *different* from break-even), it is good practice for you AND 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 obtain 0.93998852.

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

(g) ***For homework,*** copy and paste your work for the other seven locations. Verify that you are able to prove at a 5% significance level that Location 6 and Location 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 are able to 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.

D.2 Module D.2: Hypothesis Testing (Part 2 of 2): Comparing Means

Main course concepts: Inference about how two means differ. Distinguishing independent samples from paired data. Review foundation of inference (with CI estimation or hypothesis testing) about a population and its parameters using a random sample and its statistics.

Source materials (full citations in Section F): We will work with the *population* of *all* Ontario public sector employees making $100K+ using the public disclosures of 2016 and 2015 salaries. These are abbreviated Ontario (2016) and Ontario (2015).

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

- Browse Table D.1. Consider the seeming paradox: at first glance, the mean salary seems (virtually) unchanged between 2015 and 2016 with a tiny $180 mean annual increase, which is just a 0.1% rise. Are the salaries of Ontario 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 reason the first row seems to suggest virtually no increase is 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 new people holds the mean back even if the old people had their salaries rise.

Table D.1: Exploring Why the Mean Salary is Virtually Unchanged

	2015 Salaries (CAN $1,000's)	2016 Salaries (CAN $1,000's)	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

Notes: Shows means with standard deviations in parentheses and number of observations in square brackets. For why 97,600 $\neq$ 97,602, see part 7 on page 98.

- This is the same idea why in annual reports to shareholders, retail firms report revenue growth in **same-store sales**: comparing total sales 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 Ontario public sector salaries, the newly added "low" paid employees each year masks the rising salaries. The reason for reporting **same-employee salaries** or same-store sales is to address 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. Also, by law, *ALL* Ontario 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 will 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).

Textbook case studies (extra practice): "Consumer Spending Patterns" on p. 469

Datasets: For Ontario (2015) and (2016): on_sal_2015.xlsx and on_sal_2016.xlsx, where "on_sal" abbreviates "Ontario salaries." Also, on_sal_16_15.xlsx contains the merged data.

Interactive tutorial materials:

1. Using on_sal_2016.xlsx, ***replicate*** the three numbers summarizing the 2016 salaries of the same-employee subset in Table D.1, which are marked in boldface.

   EXCEL TIPS: As usual, to separately describe subsets of data, use a pivot table. Create a new worksheet and copy the variables salary and disc2015 to it. Insert a pivot table. 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 detect whether the degrees of freedom correction has been done). Drag the field Σ Values to ROWS to make the pivot table look more like Table D.1.

2. ***Browse*** Table D.2, *including* the note below it.

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

	2015 Salaries			2016 Salaries			Difference 2016-2015	
	mean	s.d.	s.e.	mean	s.d.	s.e.	mean	s.e.
$n_{2015} = 100$, $n_{2016} = 100$	**132.841**	**56.364**	**5.636**	**134.752**	**35.812**	**3.581**	1.911	6.779
$n_{2015} = 4{,}000$, $n_{2016} = 4{,}000$	129.377	37.893	0.599	132.195	39.156	0.619	2.818	0.862

Notes: To replicate, sample from the subset of employees whose 2015 and 2016 salaries are *both* 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: Copy fullname, salary, disc2016, and random3 to a new worksheet. Sort the worksheet by disc2016 in *descending* order and random3 in *ascending* order. Select the first 100 observations (101 rows, with variable labels) and compute summary statistics using descriptive statistics in data analysis.

   (b) Using on_sal_2016.xlsx, ***replicate*** the second three numbers in boldface. Use the variable disc2015 to restrict to "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. (While the variable name random3 is the same in both files, it is an entirely independent random number sequence.)

(d) ***For homework,*** replicate the remaining numbers in Table D.2. If you are confused about the first two columns of standard errors, review Module C.2. For the last column of standard errors, use this formula from your aid sheets: $\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$.

(e) ***For homework,*** consider the question: are the numbers in Table D.2 statistics or parameters? (If you have trouble answering, recall that Table D.2 shows numbers summarizing random samples.)

(f) ***For homework,*** notice that Table D.2 reports point estimates and standard errors, a format researcher often choose to present their results. Formal inference via hypothesis testing or confidence interval estimation would require computing the degrees of freedom: recall $\nu = \dfrac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2}$. However, in this case, we do not need to bother with formal inference because it is obvious, under conventional significance levels, that the confidence intervals would include the correct answers (the parameter in Table D.1: 2.754) and that we would be unable to reject a null hypothesis $H_0 : \mu_{2016} - \mu_{2015} = 2.754$ for both rows of results. (If these conclusions are not obvious to you, conduct formal inference via CI estimation and HT testing using appropriate formulas and statistical tables.)

3. The independent sample approach in Table D.2 is a very inefficient way to make an inference about how "same-employee" salaries have changed. Let's try a paired data approach instead. ***Browse*** Table D.3, *including* the note below it.

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

	2015 Salaries			2016 Salaries			Difference 2016-2015		
	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 of results for $n = 100$ in Table D.3, which are in boldface.

EXCEL TIPS: Copy name, salary16, salary15, and random3 to a new worksheet, naming it Replicate Table D.3. Create a new variable named difference that is salary16 minus salary 15. Sort by random3 (in ascending order). Select the first 100 observations (101 rows, with variable labels) of the variable difference and compute summary statistics using descriptive statistics in data analysis.

Notes: To link what Excel is doing to your aid sheets, recall that for paired data you construct a new variable d, which is the difference, and find $\bar{d}$ and s_d. For the s.e., use this formula from your aid sheets $\sqrt{\frac{s_d}{n}}$, where the numerator is just the regular s.d. of the difference variable.

(b) Tutorial time permitting (otherwise for homework), ***replicate*** the other six numbers for the $n = 100$ in Table D.3.

EXCEL TIPS: Continuing with the worksheet Replicate Table D.3, simply compute summary statistics using descriptive statistics in data analysis a couple of more times:

once for salary16 and once for salary15.

(c) ***For homework,*** replicate the results for $n = 4,000$ in Table D.3.

4. ***Notice*** how much smaller the s.e.'s of the difference in means are in Table D.3 compared to Table D.2, even though the samples sizes are the same.

   (a) ***Consider*** the reason for the smaller s.e.'s from the paired data approach, which means it is a more efficient method of inference. The fundamental reason is on your aid sheets: $V[a + bX + cY] = b^2V[X] + c^2V[Y] + 2bc * SD[X] * SD[Y] * CORR[X, Y]$, which for a difference is simply $V[X - Y] = V[X] + V[Y] - 2 * SD[X] * SD[Y] * CORR[X, Y]$. 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 $2 * SD[X] * SD[Y] * CORR[X, Y]$.

   (b) Using on_sal_16_15.xlsx, ***verify*** that the population coefficient of correlation ρ between the 2016 and 2015 salaries is 0.9304. Hence, $V[Sal_{2016} - Sal_{2015}] < (V[Sal_{2016}] + V[Sal_{2015}])$.

5. Tutorial time permitting (otherwise for homework), let's illustrate how paired data and independent samples differ. We are going to contrast two approaches to assessing how salaries changed from 2015 to 2016. In a paired data approach, we look at the *same employees* both years: compare each person with her/himself. In an independent samples approach, we take a random sample of employees in 2016 and take another random sample of employees in 2015.

   (a) Start with the paired data. From the worksheet Replicate Table D.3, ***copy and paste*** the first 100 observations of salary15, salary16, random3, and difference to a new worksheet, naming it Ind. Samples vs. Paired Data.

   (b) Next, add in an *independent random sample* for 2015 (which we can compare with the sample for 2016). From the original data, copy name, salary15, and random4 to another new worksheet in on_sal_16_15.xlsx, and ***sort*** by the variable random4. ***Rename the variable*** salary15 to Xsalary15. ***Copy and paste*** the first 100 observations of Xsalary15 and random4 to Ind. Samples vs. Paired Data. ***Create*** a variable Xdifference that is salary16 minus Xsalary15.

   (c) Before continuing, ***verify*** that the top of your worksheet Ind. Samples vs. Paired Data looks like the following. Also, ***note*** that this worksheet contain both paired data (salary16 vs. salary15) and independent samples (salary16 vs. Xsalary15).

	A	B	C	D	E	F	G
1	salary16	salary15	random3	difference	Xsalary15	random4	Xdifference
2	108.7407	106.6779	1	2.0628	135.0675	1	-26.3268
3	129.3803	125.5766	2	3.8037	103.536	2	25.8443
4	134.7101	128.9692	3	5.7409	103.5756	3	31.1345

   (d) In the worksheet Ind. Samples vs. Paired Data, ***verify*** that the coefficient of correlation between salary15 and salary16 is 0.965341. ***Verify*** that the mean and s.d. of the difference are 2.929 and 9.295, respectively.

   EXCEL TIPS: In data analysis use two tools: correlation and descriptive statistics.

i. ***For homework,*** verify that $9.295 = \sqrt{sd_{2016}^2 + sd_{2015}^2 - 2 * r * sd_{2016} * sd_{2015}} = \sqrt{1248.640 + 1244.171 - 2 * 0.965341 * 35.3361 * 35.2728} < 49.928 = \sqrt{sd_{2016}^2 + sd_{2015}^2} = \sqrt{1248.640 + 1244.171}$.

(e) ***Verify*** that the coefficient of correlation between Xsalary15 and salary16 is -0.0190. ***Verify*** that the mean and s.d. of Xdifference are 6.437 and 49.563, respectively.

i. ***For homework,*** verify that $49.563 \approx \sqrt{sd_{2016}^2 + sd_{X2015}^2} = \sqrt{1248.640 + 1161.979} = 49.098$. Because the correlation in salaries across two independent samples is virtually zero, it makes little difference if we ignore it.

EXCEL TIPS: You can copy the variable salary16 and insert it next to Xsalary15 to enable using correlation in data analysis.

6. ***For homework,*** consider these questions:

- Why is the variance of Xdifference (salary16 - Xsalary15) MUCH BIGGER than difference (salary16 - salary15)? (If you have trouble answering, review part 5 of today's tutorial. It directly addresses this question.)
- Why are the standard errors in the last column of Table D.2 much bigger than the standard errors in the last column of Table D.3? Illustrate the answer by referencing the work in part 5 of today's tutorial. The s.d. of Xdifference is way bigger: 49.563 versus 9.295. Because the standard error is the s.d. divided by the square root of the sample size, a bigger s.d. means a bigger standard error. We broke the positive correlation when we pulled two *independent* samples of 100 salaries. (The variables random3 and random4 are each random and hence give us two independent samples.) The magic of the paired data approach is caused by the positive correlation that naturally occurs with real paired data: people who make a lot of money in 2015 tend to make a lot of money in 2016. Another way to think about it is: 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.

7. **OPTIONAL: Issues in Merging Data.** Each year, Ontario publishes the required salary disclosure. However, if we wish to use a paired data approach, we must merge 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 files included a unique employee number (e.g. the employee's SIN), this merge would be very 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 from one year to the next: as examples, sometimes the name includes the middle initial and sometimes not or a person may change their name with a change in marital status or legal reasons. Also, some names are very 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. All of the complexities combined mean that 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.

D.3 Practice test questions for Module D

QUESTIONS:

Q1. 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?

Q2. 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.

(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.

Q3. 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.)

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.

Q4. 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.

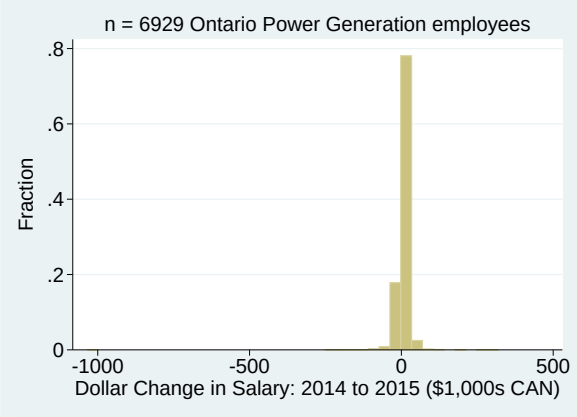
> Amongst *all* ON public sector employees of the Toronto Police Service making at least $100K in 2016 who also made at least $100K in 2015, the mean salary in 2016 is _____ dollars [answer in dollars] and the standard deviation of salary is _____ dollars [answer in dollars]. Amongst *all* ON public 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.

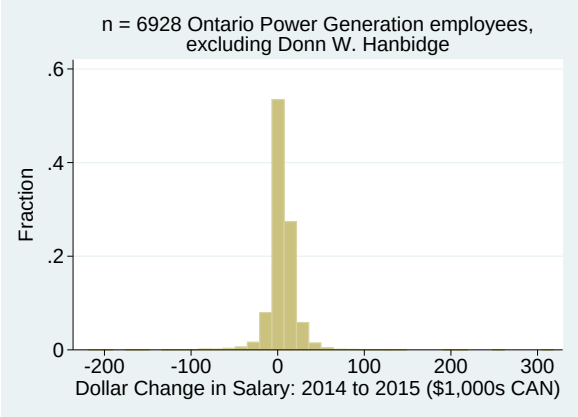
(b) Consider an independent samples (unequal variances) method of inference about the difference in mean salaries from 2015 to 2016 conditional on the employee having her/his salary disclosed in both 2015 and 2016. To make inference necessary, suppose that you did not have access to *all* relevant salaries 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 $H_0 : \mu_{samepolice16} - \mu_{samepolice15} = 0$ versus $H_1 : \mu_{samepolice16} - \mu_{samepolice15} > 0$. What is the P-value?

Q5. 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) Start by looking at the population of all 6,929 employees. Consider the histograms below. There is an obvious outlier in the population: Donn W. Hanbidge, Senior Vice President & Chief Financial Officer, who had his salary drop by over a million dollars between 2014 and 2015 (http://www.ontariosunshinelist.com/people/zkrdtf). The second histogram excludes him. Is this population (without Hanbidge) Normally distributed? To answer, fill in the blanks below. Follow the answer guides given in square brackets.

> The Empirical Rule holds for a Normal Distribution. 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. Also, use the correct Excel function for computing the *sample* standard deviation.]





(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.

Q6. 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 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.)

ANSWERS:

A1. (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)

A2. (a) There is not a statistically significant difference at a 5% significance level, but there is at a 10% significance level. The difference in percent female is 0.8 percentage points (which is very small). The standard error of this differences is 0.43 percentage points. The z test statistic is 1.7583496. The P-value is 0.07868805.

(b) There is not a statistically significant difference at any conventional significance level. The difference in highest previous contribution is 64 cents (which is very small). The standard error of this differences is 65.65 cents. The t test statistic is -0.97043004. The degrees of freedom are 35913.884. The P-value is 0.33183872.

(c) There is not a statistically significant difference at any conventional significance level. The difference in the number of previous donations is 0.54 donations (which is very small). The standard error of this differences is 0.4801. The t test statistic is 1.1296069. The degrees of freedom are 1837.0323. The P-value is 0.2587894.

A3. 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 (39.613) [16,373]	147.332 (41.486) [17,063]	1.839
Conditional on employee having both her/his 2014 salary and 2015 salary disclosed: "same-employee" comparison	146.849 (39.539) [15,201]	151.121 (41.447) [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 7 on page 98.

A4. (a) Amongst *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 18,058 dollars. Amongst *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 18,580 dollars. These four numbers are parameters. In comparing 2016 with 2015, this paragraph uses a conditional ("same-police") approach.

(b) Remember that the relevant population subset for 2016 are observations with employer="Toronto Police Services" and disc2015=1 and that the relevant population 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.)

A5. (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 mu_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.

A6. (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.

E Module E: Interactive Tutorial Materials & Test Prep

E.1 Module E.1: Multiple Regression in Applied Research

Main course concepts: Distinction between multiple (multivariate) regression and simple (bivariate) regression. How correlation does *not* relate to a slope coefficient in a multiple regression like in simple regression. Linking course concepts about multiple regression to real research and data.

Source materials (full citations in Section F): You will replicate parts of the analysis from an academic journal article "Toward an Understanding of Learning by Doing: Evidence from an Automobile Assembly Plant," abbreviated Levitt et al. (2013).

Most relevant required readings: Sections 20.1 - 20.3 and "Logarithms in Regression Analysis with Asiaphoria". 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.

- Consider Figure 1. "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)

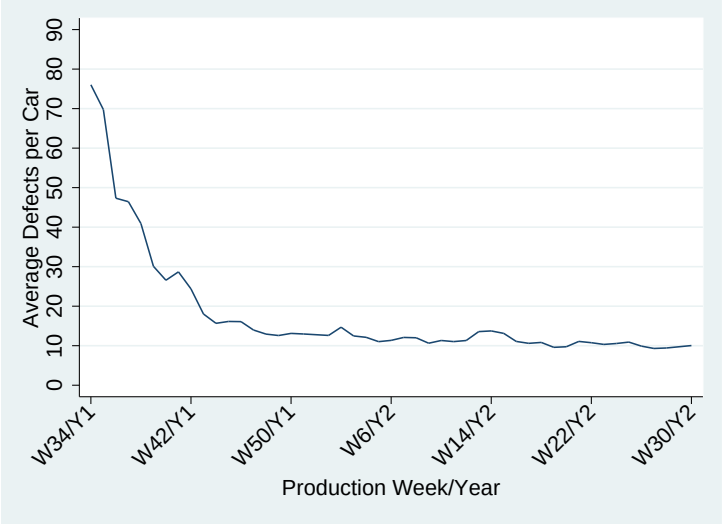


Figure 1: Levitt et al (2013), p. 654.

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 this simple empirical model of the learning process:

$$ln(D_t) = \alpha + \beta ln(E_t) + \varepsilon_t \tag{1}$$

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 up to that point (cumulative production). The researchers use

the natural logarithm (which they abbreviate either as log or ln) to transform the original data. They also explore an alternative model that includes a time trend:

$$ln(D_t) = \alpha + \beta ln(E_t) + \gamma * t + \varepsilon_t \quad (2)$$

where t is a variable measuring the number of time periods since the start of production.

- Consider Table 1. "Table 1 shows the results of estimating these specifications with our sample. Panel A contains the results from specifications using weekly data (average defect rates over the week and production experience at the week's outset); Panel B shows results obtained 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: Supplement to the ECO220Y1Y April 2017 final exam, p. 3, which is a clarified version of Table 1 on p. 655 of Levitt et al (2013).

- Wait one minute. Doesn't Figure 1 show a strong and clear *negative* association between average defects per car and a time trend? How can the coefficient on the time trend in Table 1, Panel A, Column (2) be *positive*? It has nothing to do with the natural log transformation: the simple correlation between ln(average defects per car) and the time trend is also strongly negative while the *multiple regression* coefficient is positive. (You may check your answers against the examApril 2017 final exam solutions for Question (3).)

- Finally, consider Figure 2. "The simple empirical model fits the data very well at both frequencies, 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)

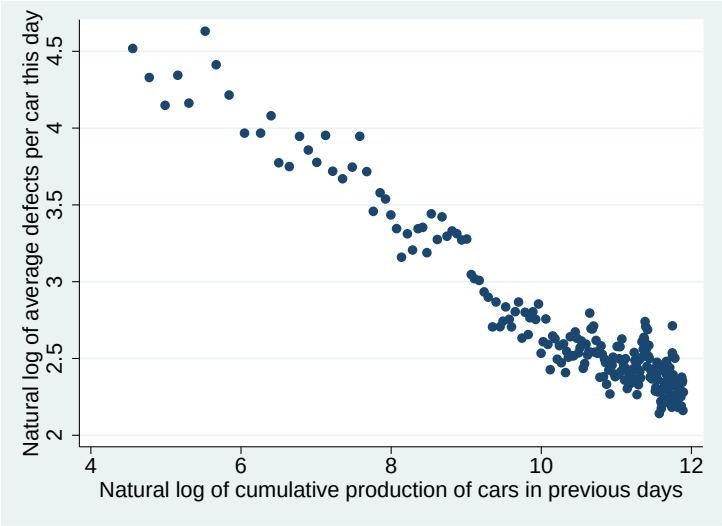


Figure 2: Levitt et al (2013), p. 656.

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.

- Also, you will need some background information about the production of automobiles.
  - 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" for Toyota, 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 and later production is ramped up to full capacity by adding a second shift.

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 tutorial materials:

1. Consider Levitt et al (2013) and use learn_do_weekly.xlsx for all subparts.

(a) ***Browse*** the data and variable definitions.

(b) ***Replicate*** Figure 1. ***Create*** the y-axis variable with the appropriate two variables. Name the new variable. Like the authors, use only weeks with production of at least 100 cars. (Do *not* worry about the x-axis tic values in Figure 1: just use the week number.)

EXCEL TIPS: One straightforward way to restrict your graph to weeks with production of at least 100 cars is to selectively clear cells of your new analysis variable for those weeks. For a more sophisticated approach, use the `IF` function when creating the new variable.

(c) ***Replicate*** the simple regression in Table 1, Panel A, Column (1). ***Create*** a variable measuring the *cumulative* production in *previous* weeks and then use a natural log transformation. To create the y variable, apply a natural log to the y variable in Figure 1. Remember to graph only weeks when at least 100 cars are produced: your number of observations should match the table. To help you visualize the replication, see Figure E2. Excel cannot compute robust standard errors: you will obtain (0.009) instead of (0.007).

EXCEL TIPS: To create the cumulative *previous* production variable, use the function `=SUM($C$2:C2)` in *row 3* of the new variable. The sum is up to but not including row 3. Copy and paste to other rows. Check row 51 has `=SUM($C$2:C50)` and evaluates to 143074.0461. Remember to clear your new analysis variable for weeks with production below 100 cars. (Cumulative production includes *all* weeks, including those with production below 100 cars. In other words, do *not* clear the original variable with each week's production.) Finally, copy and paste the analysis data to a new worksheet (starting with the first non-missing values). Insert a first row with variable names for the y and x variables and select the labels option in the regression tool. It is helpful to see the variable names in the regression output.

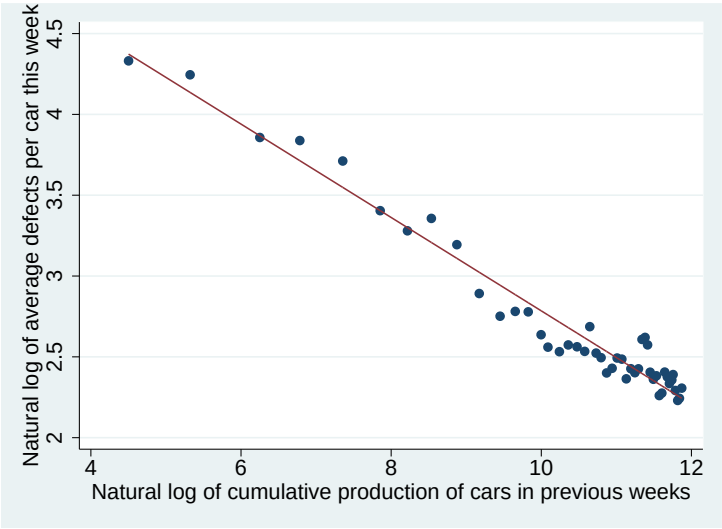


Figure E2: A figure like Figure 2 on p. 656 of Levitt et al (2013), *except that* it shows the weekly data (instead of the daily data). This figure corresponds exactly to Table 1, Panel A, Column (1).

(d) ***Replicate*** the multiple regression results in Table 1, Panel A, Column (2). Note that you will obtain (0.016) and instead of (0.017). For (0.002), the regular and robust standard errors differ only a little: rounded they are equal.

EXCEL TIPS: Use the regression tool in data analysis. Because multiple regression involves more than one x variable and Excel can only accept adjacent x variables, you will need to copy the time trend variable and paste it into a column next to the other x variable.

(e) ***Construct a correlation matrix*** with the simple pairwise correlations between each of the three variables: ln(average defects per car), ln(cumulative production), and the time-trend variable using the same data you used for the multiple regression in the previous part. ***Check*** your work against the output below. ***Note*** the negative correlation between the time trend and natural log of defects per car: this does *not* contradict the positive coefficient on the time trend in the multiple regression in the previous part.

	ln_def_car	*ln_cum_prod*	*time_trend*
ln_def_car	1.0000		
ln_cum_prod	-0.9802	1.0000	
time_trend	-0.8091	0.8703	1.0000

- ***For homework,*** explain how the *correlation* between ln(average defects per car) and the time trend can be strongly negative while the *multiple regression* coefficient is positive. (Check your answer against April 2017 exam solutions, Question (3).)

(f) Table 1, Panel A, Column (1) shows a *simple regression* which means there is an uncomplicated relationship between correlation and the slope (and they never do crazy things like have different signs). To illustrate, ***standardize*** each variable in the Column (1) regression (i.e. standardize the ln(average defects per car) and standardize the ln(cumulative production)) and ***rerun the simple regression***. Recall that standardizing a variable means transforming it by subtracting its mean and dividing by its standard deviation: $z_X = \frac{X-\bar{X}}{s_X}$. ***Verify*** that the slope of the standardized regression equals the coefficient of correlation you obtained in the previous part (-0.9802).

EXCEL TIPS: Use the functions `AVERAGE` and `STDEV.S` (yes, the sample standard deviation). Create new variables named something like s_∗ and remember to use the $ to anchor to the cells containing the sample mean and sample standard deviation.

(g) Tutorial time permitting (otherwise, for homework), let's double check your understanding of multiple versus simple regression and correlation. What if you standardized all three variables in the multiple regression in Table 1, Panel A, Column (2)? ***Write down a sentence*** predicting the sign (negative or positive) of each multiple regression coefficient if each variable is standardized prior to running the multiple regression.

i. Next, let's see if your prediction holds true. ***Standardize*** each variable in the Column (2) regression (i.e. standardize the ln(average defects per car), standardize the ln(cumulative production), and standardize the time trend) and ***rerun the multiple regression***. Now ***verify*** that standardization has no effect on the sign of the multiple regression coefficients. It is still the case that the multiple regression coefficient on the time trend variable is positive even though the correlation between the time trend variable and the variable measuring defects is negative. Also, note that standardizing all variables will result in a zero (to machine precision) for the intercept. This is because a regression always passes through the mean and all of the variable means are zero after standardization.

E.2 Module E.2: Multiple Regression & Inference

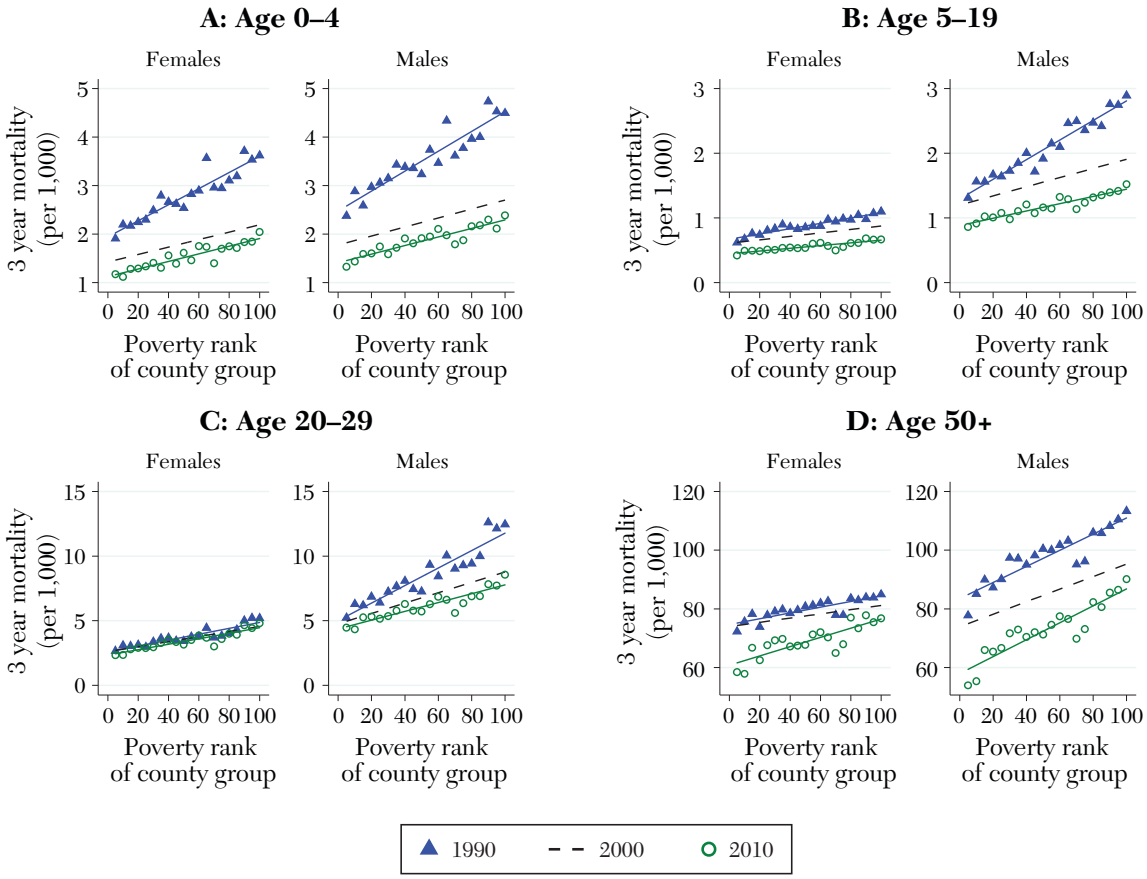
Main course concepts: Inference with simple and multiple regression. Assessing a multiple regression model overall.

Source materials (full citations in Section F): We revisit percent body fat from "Just Checking" on p. 695 from "Fitting Percentage of Body Fat to Simple Body Measurements," abbreviated Johnson (1996). We replicate parts of the analysis from an academic journal article "Mortality Inequality: The Good News from a County-Level Approach," abbreviated Currie and Schwandt (2016).

Most relevant required readings: Sections 20.1 - 20.6. *Everything* about Currie and Schwandt (2016) in Module B.1, including how they addressed composition effects (aka Simpson's paradox).

- For Currie and Schwandt (2016), carefully review Figure 3, reproduced below.

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**."

- 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

- A quick refresher on statistical inference with simple and multiple regression:
  - There are two methods of statistical inference: confidence interval estimation and hypothesis testing (via P-value or rejection region (aka critical value) approach). Conventional significance levels are α equals 0.01, 0.05, or 0.10. You also can see others, like $\alpha = 0.001$.
  - For inferences about *individual regression coefficients*, use t test statistics and the Student t distribution.
    - The confidence interval estimate of a slope coefficient is: $b \pm t_{\alpha/2} s_b$ with $\nu = n - k - 1$ where k is the number of x variables and s_b is the standard error of the slope coefficient.
    - In testing a regression coefficient, there are three cases:
      - Two-tailed test: $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$
      - Right-tailed test: $H_0 : \beta = \beta_0$ versus $H_1 : \beta > \beta_0$
      - Left-tailed test: $H_0 : \beta = \beta_0$ versus $H_1 : \beta < \beta_0$
    - The most common test – automatically run by Excel – is the classic test of statistical significance of a coefficient: $H_0 : \beta = 0$ versus $H_1 : \beta \neq 0$, which is two-tailed.
    - For these hypothesis tests you use a t test statistic given by $t = \frac{b - \beta_0}{s_b}$ with $\nu = n - k - 1$.

- For inferences about the *overall statistical significance of a regression*, use the F test statistic and the F distribution.
  - Only one case: $H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$ versus H_1 : *Not all the slopes are zero.*
  - You use a F test statistic given by any of these equivalent formulas: $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$, $F = \frac{(SST-SSE)/k}{SSE/(n-k-1)}$, $F = \frac{SSR/k}{SSE/(n-k-1)}$, or $F = \frac{MSR}{MSE}$, with numerator degrees of freedom $\nu_1 = k$ and denominator degrees of freedom $\nu_2 = n - k - 1$.
- Note that in the special case of a simple regression (which means that $k = 1$), the F test and t are the exact same test: $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$. Regardless of whether you do a t test or an F test, you will reach the same conclusions with the same P-value. However, for multiple regression (which means that $k > 1$), you must use the F test if you wish to test the overall statistical significance. To test an individual coefficient, you use a t test, regardless of whether it is simple or multiple regression.
  - 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. Recalling that for a simple regression the R^2 is the coefficient of correlation squared, it is convenient to use this test statistic formula: $F = \frac{R^2/k}{(1-R^2)/(n-k-1)}$.

Textbook case studies (extra practice): "Golf Success" on p. 713

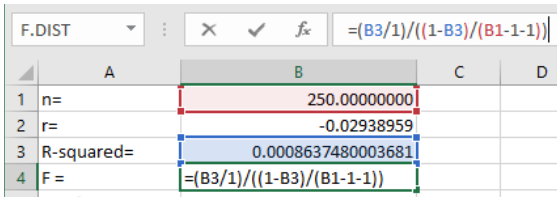
Datasets: For Johnson (1996): pct_body_fat.xlsx, where "pct_body_fat" abbreviates "Percentage of Body Fat" from the title. For Currie and Schwandt (2016): mort_in_figure_3_table_a3.xlsx, where "mort_in" abbreviates "Mortality Inequality" from the title and the rest tells that these are the data needed to replicate Figure 3 and Table A3 in that paper.

Specific topics covered in this tutorial:

1. Consider Johnson (1996) and use pct_body_fat.xlsx:

   (a) ***Replicate*** the multiple regression results shown in the "Just Checking" box. Note that you will need to convert height from inches to centimeters (1 inch = 2.54 cm) and weight from pounds to kilograms (1 pound = 0.45359237 kg). Also, the coefficient on weight will be 0.5592256, which doesn't perfectly match the textbook result of 0.55805 because the textbook did not precisely convert pounds to kilograms.

   (b) ***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:** Be careful putting the F statistic formula in Excel. You must use parentheses to preserve the proper order of operations.

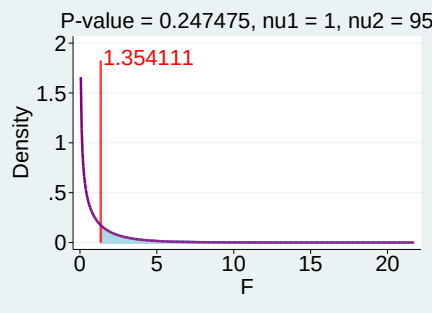


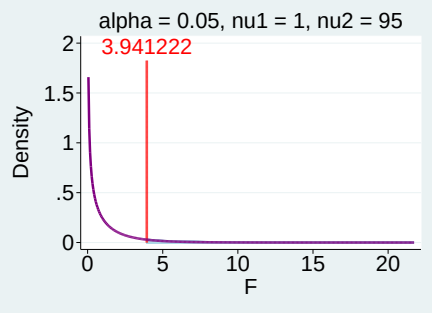
(c) ***Verify*** the value of the F statistic by ***running a simple regression*** where the y-variable is percent body fat and the x-variable is height. Using your simple regression output, ***identify*** the P-value to test if there is a statistically significant correlation between percent body fat 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.

- ***For homework,*** 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.)

(d) Suppose that for a different sample with 184 males, the correlation between percent 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, which means that this correlation is not statistically different from zero at any conventional significance level.

EXCEL TIPS: Use the F.DIST function. It returns the cumulative area under the F distribution to the left of the F test statistic, given the numerator and denominator degrees of freedom. Like other such functions in Excel, set the logical value to TRUE (to say you want the area below that value, not the height of the density function). Remember that the P-value is the area *above* the F test statistic, which means you need (1-F.DIST()). For example, `=1-F.DIST(5.84893192,2,20,TRUE)` returns 0.01, which is consistent with 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 1d. The figure on the right visually illustrates the answer to part 1e.

(e) Suppose that for yet another sample with 97 males, you wish to assess the correlation between percent 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, which means that you must obtain an F test statistic of at least 3.941222 for the correlation to be statistically significant at a 5% significance level.

EXCEL TIPS: Use the F.INV function. It returns the critical value given a cumulative area below it and the degrees of freedom. For example, `=F.INV(0.99,2,20)` returns 5.84893192, which matches the critical value provided by the F table in the Aid Sheets for $\alpha = 0.01$, $\nu_1 = 2$ and $\nu_2 = 20$.

(f) ***Re-run the multiple regression*** from part 1a using only the first 10 observations. Using your regression output, ***identify*** the P-value to test the overall statistical significance of the model. ***Check*** your work against the Excel output below.

Regression Statistics	
Multiple R	0.7733
R Square	0.5980
Adjusted R Square	0.3969
Standard Error	6.2889
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	352.9296	117.6432	2.9746	0.1186
Residual	6	237.2994	39.5499		
Total	9	590.2290			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	261.9057	85.9843	3.0460	0.0226	51.5096	472.3017
height_cm	-1.9510	0.6595	-2.9584	0.0253	-3.5647	-0.3373
weight_kg	1.4810	0.5860	2.5273	0.0448	0.0471	2.9150
age	-0.5931	1.5250	-0.3889	0.7108	-4.3247	3.1384

- ***For homework,*** explain why the P-value is higher than in the previous part.

2. 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 111). ***Verify*** that you get y-hat = 1.9404 + 0.0166*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.

   EXCEL TIPS: First, add a variable measuring the adjusted mortality to the original data. Next, recall the Filter tool for selecting the subset from the original data. After applying a filter to the original data, copy the subset of the data to a new worksheet (carefully named) and put the regression output in the same worksheet as the subset. (It is important not to run regressions directly on the original data with the filter because this creates issues. Instead, follow the suggestion to copy-and-paste the filtered data to a new worksheet.) Creating a new worksheet also helps document which subset the regression results are for, even if you have not named your worksheet extremely precisely. Also, copying the subset for the regression to a new worksheet allows you to select the labels option in regression (and it is helpful to have variable names in the regression output).

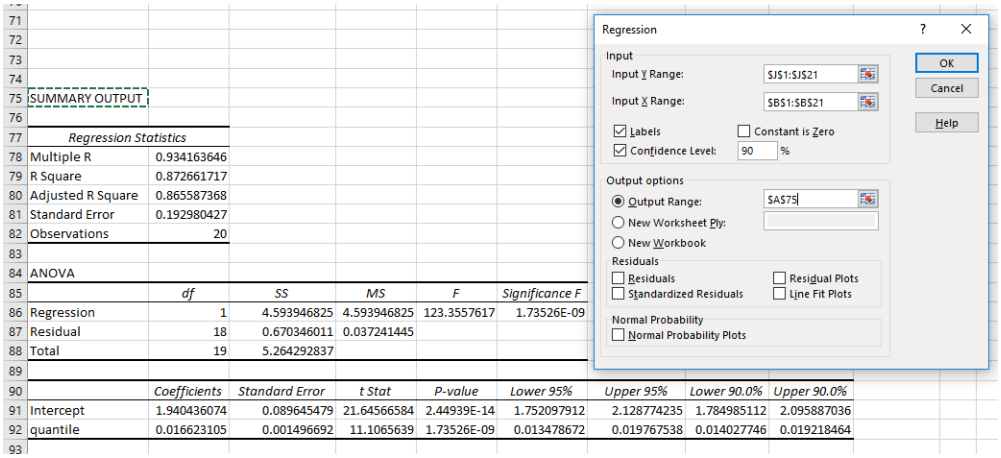
      i. ***Assess*** whether we can conclude that the slope is less than 0.02.[5] If so, at which significance levels? ***Note*** that this requires testing $H_0 : \beta = 0.02$ versus $H_1 : \beta < 0.02$. ***Verify*** that you obtain a t test statistic of -2.256239 and a P-value of 0.018366, which means 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.

      EXCEL TIPS: To compute the P-value, recall the T.DIST function.

      ii. ***Find*** the 90% confidence interval estimate of the slope coefficient. ***Verify*** that you obtain [0.0140277, 0.0192185].

[5] Why 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 111, 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: $slope = \frac{\Delta y}{\Delta x} \approx \frac{4-2}{100-0} = 0.02$.

EXCEL TIPS: There are two ways to do this. Make sure you can do it *both* ways. 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.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.934163646
R Square	0.872661717
Adjusted R Square	0.865587368
Standard Error	0.192980427
Observations	20

ANOVA

	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

(b) In Module B.1 we studied how adjusted mortality rates control for changes in age composition over time so we can focus on real changes in mortality (i.e. not get confounded by demographic changes). How much does the adjustment affect the key regression results? ***Run TWO simple regressions*** for a specific subset, males aged 50+ in 2010:

i. First, use the **adjusted mortality rate** (i.e. replicating the green line for males in Panel D of Figure 3 on page 111). ***Verify*** that you get y-hat = 57.941 + 0.2879*x, n = 20. (Note: Currie and Schwandt (2016) tried various age groupings, including 50+ and 65+, which have some overlap. The 50+ group is everyone 50 years or older (including over 65).)

EXCEL TIPS: First, add a new variable measuring the unadjusted mortality rate to the original data. Next, recall the Filter tool for selecting the subset from the original data. After applying a filter to the original data, copy the subset of the data to a new worksheet (carefully named) and put the regression output in the same worksheet as the subset. (It is important not to run regressions directly on the original data with the filter because this creates issues. Instead, follow the suggestion to copy-and-paste the filtered data to a new worksheet.) Creating a new worksheet also helps document which subset the regression results are for, even if you have not named your worksheet extremely precisely. Also, copying the subset for the regression to a new worksheet allows you to select the labels option in regression (and it is helpful to have variable names in the regression output).

ii. Second, use the **unadjusted mortality rate** for the same subset: males aged 50+ in 2010. ***Verify*** that you obtain y-hat = 60.58724 + 0.2127863*x, n = 20.

- ***For homework,*** compare and contrast the point estimates of the slopes and the 95% confidence interval estimates of the slopes. Which method leads us to detect to greater inequality across county groups? (You may double-check your interpretation of these results by reviewing the April 2017 final exam and solutions.)

(c) Look at Figure 3 on page 111 for males in 1990 (the blue triangles) aged 20-49 (Panel C)

versus aged 50+ (Panel D).

i. ***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+?

ii. ***Check*** your answers by running the two simple regressions that correspond to each. ***Verify*** that you obtain y-hat = 5.018749 + 0.0675931*x, n = 20, s_e = 0.73538 for males in 1990 aged 20-49. ***Verify*** that you obtain y-hat = 83.54337 + 0.2735487*x, n = 20, s_e = 3.7934 for males in 1990 aged 50+.

- ***For homework,*** discuss why the regression results are consistent with a careful visual inspection of the graphs in Figure 3, which notices the much higher mortality rate in the older age group and hence differing scales on the y-axes.

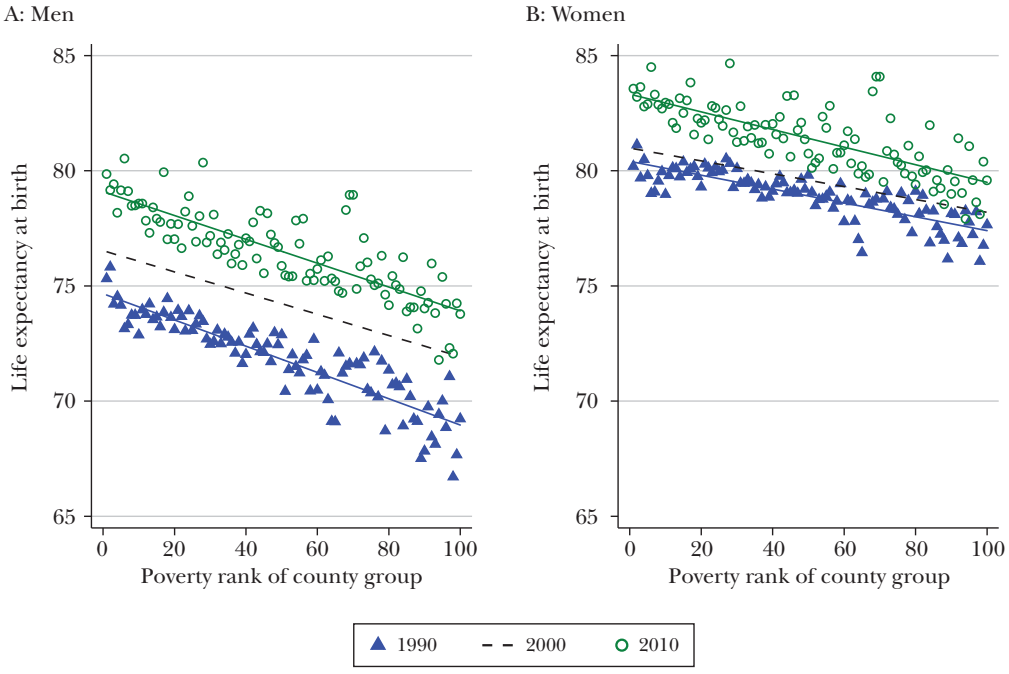
E.3 Module E.3: Dummy Variables & Interaction Terms

Main course concepts: Dummy variables and interactions when relationships differ across groups.

Source materials (full citations in Section F): Continue with Currie and Schwandt (2016).

Most relevant required readings: Sections 21.1 - 21.2 in the textbook. *Everything* about Currie and Schwandt (2016) in Modules B.1 and E.1. Review Figure 2, including the note and the units of measurement of the x and y variables, and Table A2 on page 119 (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.

Additional readings (not required, but strongly recommended): Currie and Schwandt (2016) discuss and interpret the key figures and tables. For some particularly helpful discussion/interpretation, see sections "Life Expectancy at Birth" on pp. 38-40 and "Age-specific Mortality" on pp. 40-41. (Note: While these readings are not required, you *are* required to interpret the figures and tables.)

Textbook case studies (extra practice): "Canadian Snow Birds" on p. 757

Datasets: For Currie and Schwandt (2016): mort_in_figure_2_table_a2.xlsx[6] and mort_in_figure_3_table_a3.xlsx, where "mort_in" abbreviates "Mortality Inequality" from the title and the rest either tells which figures/tables the data can be used to replicate.

Interactive tutorial materials:

1. Consider Currie and Schwandt (2016) and use mort_in_figure_2_table_a2.xlsx.

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)
(a) LE at birth by poverty ranking of county group								
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
(b) Slope of fitted regression line								
	Slope 1990		Slope 2010		Slope 1990		Slope 2010	
	-0.0570		-0.0518		-0.0301		-0.0383	
(c) p-value of test Slope1990=Slope2010								
	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 118. Panel (c) tests if those slopes differ between 1990 and 2010.

(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 on page 116 (Module E.2, part 2(b)i).

[6]These 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.

(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.

- ***For homework,*** study how your OLS lines in parts 1a (1990) and 1b (2010) match the blue and green lines, respectively, in Figure 2, Panel A (males).

(c) ***Create*** the required dummy variables and interaction terms and ***run the multiple regression*** corresponding to Panel A (males) in Figure 2 for *both* years: 1990 and 2010.

i. Create a dummy variable either for year 1990 or for year 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. Create an interaction term between the year dummy and the x variable (quantile).

iii. Do *not* include the year 2000 data in your regression. In total, your multiple regression should have $k = 3$ and $n = 199$. ***Check*** your output against that shown below.
EXCEL TIPS: There are several ways to exclude the year 2000 data. A recommended way is to handle this the same way you would any subset of 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

OR

	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

(d) ***For homework,*** review Panels (b) and (c) in Table A2 to verify that the results for males match up with your multiple regression results in part 1c. Note: There is a tiny typo in the original table: the P-value should be 0.2748, not 0.2749.

2. Consider Currie and Schwandt (2016) and use mort_in_figure_3_table_a3.xlsx. Recall Figure 3 on page 111. Like Figure 2 and Table A2 go together, Figure 3 and Table A3, go together.

(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).

- ***For homework,*** study how your OLS lines in parts 2a (1990) and 2b (2010) match the blue and green lines, respectively, in Figure 3, Panel A (aged 0-4 years) for females.

(c) To check if the slopes in parts 2a (1990) and 2b (2010) differ in a statistically significant way requires a multiple regression (not two separate simple regressions as we've just done). ***Create*** the required dummy variables and interaction terms and ***run the multiple regression*** corresponding to Panel A in Figure 3 for females for *both* years: 1990 and 2010. ***Verify*** the P-value on the interaction term matches Column (11) of Table A3.

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. Do *not* include the year 2000 data in your regression. In total, your multiple regression should have $k = 3$ and $n = 40$.

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 *lowest* poverty rate				counties with *highest* poverty rate						
	1990		2010		1990		2010				p-value of
	rate	std. err.	rate	std. err.	rate	std. err.	rate	std. err.	1990	2010	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.

(d) ***For homework,*** review Table A3 and make sure you understand what is being reported in Columns (9), (10), and (11) and how these results are obtained (which is what we just did in tutorial). Also, study 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. We just did that to help you better grasp the multiple regression, which combines them and allows you to statistically test for differences.)

E.4 Practice test questions for Module E

QUESTIONS:

Q1. 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.

Q2. 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.

Q3. 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.

(a) What is the t test statistic for the hypothesis test $H_0 : \beta_3 = 0$ versus $H_1 : \beta_3 \neq 0$?

(b) What is the P-value for the hypothesis test $H_0 : \beta_3 = 1$ versus $H_1 : \beta_3 > 1$. Conclusion?

(c) What is the 99% confidence interval estimate of β_3?

(d) What is the critical t value for the hypothesis test $H_0 : \beta_3 = 2$ versus $H_1 : \beta_3 < 2$ given $\alpha = 0.01$?

Q4. 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 $H_0 : \beta_1 = \beta_2 = ... = \beta_{34} = 0$ versus H_1 : *Not all β's are zero.*)

Q5. 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.

(d) 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? 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).

Q6. Recalling Levitt et al (2013), use learn_do_daily.xlsx for all subparts. Like the authors, *use only days where at least 20 cars are produced.*

(a) Replicate the simple regression results in Table 1, Panel B, Column (1).

- 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).

- 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.

(c) Indicate which level of data – weekly or daily – yields a higher R-squared value in Table 1, Panel A, Column (1). Do the concepts discussed in Section 19.4 of the textbook "Working with Summary Values" apply in this context?

Q7. Levitt et al (2013) check if the results in Table 1 hold up to scrutiny by exploring many variations on the analysis. In addition to re-running everything for weekly and daily data, they consider these variations: 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 of those variations: 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*	−.346*	−.154*	−.088*
	(.010)	(.008)	(.011)	(.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). 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). Like the authors, use only weeks where at least 100 cars are produced during the second shift.

- 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.

Q8. Recalling Currie and Schwandt (2016), use mort_in_figure_3_table_a3.xlsx

(a) 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?

(b) 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?

(c) In 2010, how is mortality related to poverty for males aged 20-49 versus males aged 50+? First, note that we are making a comparison of 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. So, what can we do? We can take the natural log of the y-variable, which will allow us to interpret 100*"the slope coefficient" as the percentage change in mortality given a one unit increase in the poverty ranking. However, 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+?

Q9. 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.?

Q10. 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 instead of comparing 1990 with 2010: in other words, suppose that Column (9) corresponded to the year 2000 instead of 1990. Conduct an appropriate analysis to find the values for Columns (9), (10) and (11) for females aged 5-19 in this scenario.

Q11. 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).

(a) 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.[7] (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?

Q12. 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. Consider the question: 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).

(a) 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?

[7] You 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).

ANSWERS:

A1. With the given information we can do an F test. We compute an F test statistic of 3.68070610 and a P-value of 0.06754125. Hence, we can conclude at a 10% significance level that this correlation is statistically significant, 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.

A2. With the given information we can do an F test. We compute an F test statistic of 8.80846052 and a P-value of 0.00306638. Hence, we can conclude at a 1% significance level that the slope coefficient is statistically significant.

A3. (a) The t test statistic is 7.75136451.

(b) The t test statistic is 2.29316057 and the P-value is 0.01212493. We can conclude at a 5% significance level that β_3 is larger than 1, but we have not met a 1% burden of proof.

(c) For a 99% confidence level, LCL is 0.93764127 and the UCL is 1.90262069.

(d) The critical value is -2.36997678, which means that we must obtain a t test statistic below -2.36997678 to conclude that β_3 is less than 2 at a 1% significance level.

A4. The critical F value is 1.94212641, which means you 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.

A5. (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.

(d) 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.

A6. (a) Verify that you obtain same coefficient, s.e., number of observations, and R-squared values as reported in Table 1, Panel B, Column (1) of Table 1 in Levitt et al (2013).

(b) Verify that you obtain same coefficients, s.e.'s, number of observations, and R-squared values as reported in Table 1, Panel B, Column (2) of Table 1 in Levitt et al (2013)

EXCEPT that you will obtain 0.011 as the s.e. on the cumulative production coefficient (instead of 0.014, which is the robust s.e.).

(c) 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.

A7. (a) Check your "slope" coefficient, sample size and R-squared against Table 2.

(b) Check your "slope" coefficient, sample size and R-squared against Table 2. 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: including it or not does not make much difference.)

A8. (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).

A9. (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.

A10. (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 estimates being: adjusted_mortality_hat = 2.479075 + 0.0204704*quantile - 1.071161*yr2010 - 0.011643*quantileXyr2010.)

(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: adjusted_mortality_hat = 1.407914 + 0.0088274*quantile + 1.071161*yr1990 + 0.011643*quantileXyr1990.)

(c) This involves running a multiple regression with $n = 40$ and $k = 3$ (with the x variables including 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.

A11. (a) This involves running a multiple regression with $n = 40$ and $k = 3$ (with the x variables including 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-squared 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 (newborn up to a year) 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.

A12. (a) This involves running a multiple regression with $n = 40$ and $k = 3$ (with the x variables including 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

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.

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

Clements, Kenneth W., and Jiawei Si. 2017. "Simplifying the Big Mac Index." *Journal of International Financial Management & Accounting*, 28(1): 86-99. DOI: 10.1111/jifm.12058. http://onlinelibrary.wiley.com/doi/10.1111/jifm.12058/abstract

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

Data & Statistical Services at Princeton University (online). "Descriptive Statistics Using Excel and Stata." http://www.princeton.edu/~otorres/Excel/

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. http://www.rug.nl/ggdc/productivity/pwt/pwt-releases/pwt8.0
- Penn World Table 8.1, Released Apr. 13, 2015. DOI: 10.15141/S5NP4S. http://www.rug.nl/ggdc/productivity/pwt/pwt-releases/pwt8.1
- Penn World Table 9.0, Released Jun. 9, 2016. DOI: 10.15141/S5J01T. http://www.rug.nl/ggdc/productivity/pwt/

Google (online). "Google Finance: Stock market quotes, news, currency conversions & more." https://www.google.ca/finance

- "S&P/TSX Composite index (INDEXTSI: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. 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-disclosure-act-disclosures-2014. Data downloaded on May 25, 2017.

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

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

O'Brien, Thomas J., and Santiago Ruiz de Vargas. 2017. "The Adjusted Big Mac Methodology: A Clarification." *Journal of International Financial Management & Accounting*, 28(1): 70-85. DOI: 10.1111/jifm.12054. http://onlinelibrary.wiley.com/doi/10.1111/jifm.12054/abstract

U.S. Department of Energy. 2017. "Fuel Economy Guide: 2017 Datafile." Retrieved from https://www.fueleconomy.gov/feg/download.shtml on Jun. 9, 2017.

Organisation for Economic Co-operation and Development (online). "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.

Picker, Les. 2015. "Digest: Asiaphoria Meets Regression to the Mean." 2015. *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

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

G Appendages: Some other required supplements to the textbook

In the hardcopy version of this DACM Handbook, some other required supplements to the textbook are appended after this page. This is merely for the convenience of having paper copies of these: all supplements (including the DACM Handbook) are available free-of-charge in electronic (.pdf) format. Here is the order of the appendages:

1. "Logarithms in Regression Analysis with Asiaphoria"
2. Aid sheets (for entire course): includes formulas and statistical tables
3. A duplicate copy of the Summer 2018 Calendar of DACM Events from Section 2.3 on page 7 (for easy reference on the back cover)