Lecture 6: Tables and Graphs

Yi Chen

ShanghaiTech University

2021

Outline

1. Data Visualization: Tables or Graphs?

2. Making Tables

3. Plotting Graphs

Outline

1. Data Visualization: Tables or Graphs?
2. Making Tables
3. Plotting Graphs

What's new in this lecture?

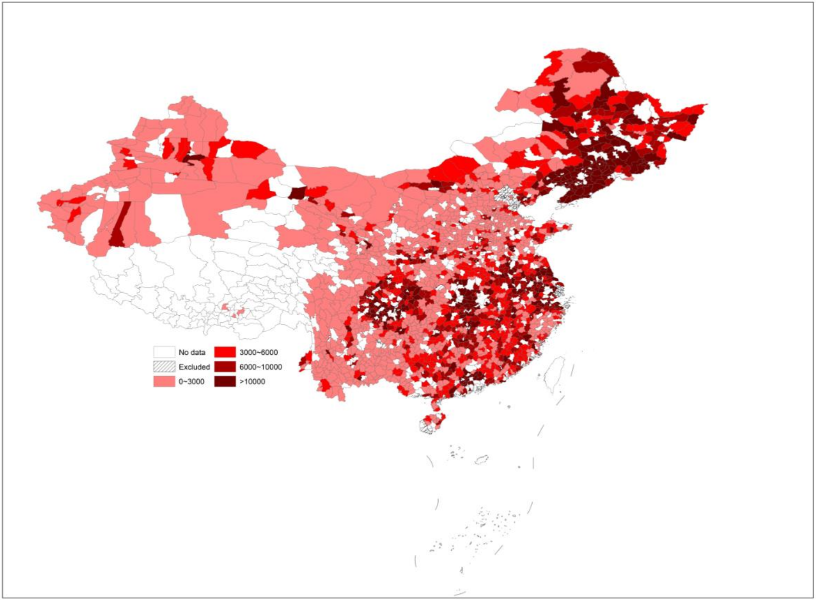
- Understanding the facts ⟶ presenting the facts
- Your own/collaborators' perspectives ⟶ readers' perspective
- A concept of "design"
  1. What should be presented?
  2. How it should be presented?
  3. Why it should be presented?

Tables versus Graphs

In academic papers, tables and graphs are two most frequent ways of visualizing data.

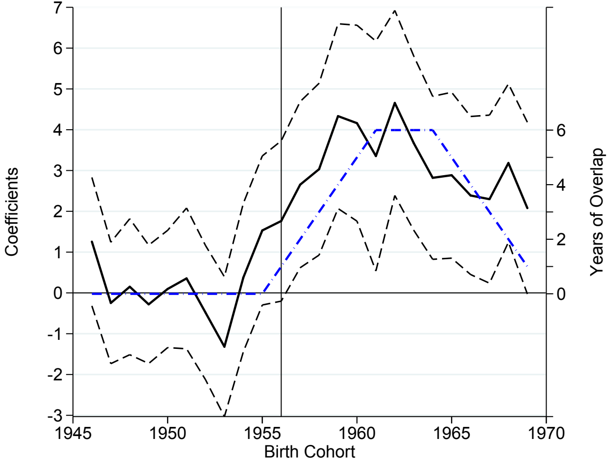
- Accuracy of information
  - Figures often for summary; tables often for regression.
- Complexity of information
  - Tables often have a “model” behind.
- Amount of information
  - It is tremendously difficult for our brains to read lots of numbers and find the regularities.
- Ease of interpreting information
  - Figures often used in presenting trends.

Number of Sent-down-youth Received by Each County



Graphs Not Always for Summary Purpose

$$\text{Y\_Edu}_{i,g,c,p} = \beta_0 + \sum_{\gamma=1946}^{1969} \beta_{1,\gamma} \%\text{SDY}_{c,p} \times I\left(g = \gamma\right) + \beta_2 X_{i,g,c,p} + \lambda_c + \mu_{g,p} + \varepsilon_{i,g,c,p}$$



Outline

1 Data Visualization: Tables or Graphs?

2 Making Tables

3 Plotting Graphs

Before we proceed to the process of making tables, let's think about following questions—

What makes a good table? What makes a bad table?

- Just right the amount of information. No more, no less.

An example of good table—Lundborg et al. (2017) AER 2017

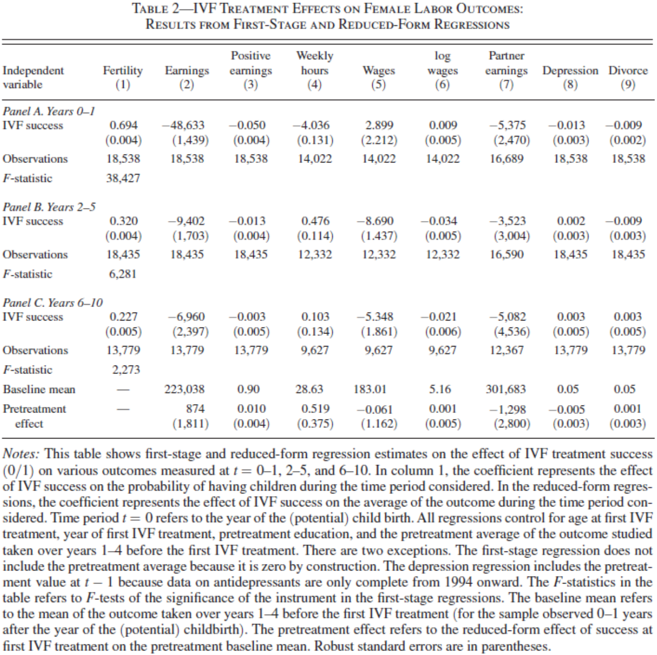


TABLE 2—IVF TREATMENT EFFECTS ON FEMALE LABOR OUTCOMES: RESULTS FROM FIRST-STAGE AND REDUCED-FORM REGRESSIONS

Independent variable	Fertility (1)	Earnings (2)	Positive earnings (3)	Weekly hours (4)	Wages (5)	log wages (6)	Partner earnings (7)	Depression (8)	Divorce (9)
Panel A. Years 0–1									
IVF success	0.694	−48,633	−0.050	−4.036	2.899	0.009	−5,375	−0.013	−0.009
	(0.004)	(1,439)	(0.004)	(0.131)	(2.212)	(0.005)	(2,470)	(0.003)	(0.002)
Observations	18,538	18,538	18,538	14,022	14,022	14,022	16,689	18,538	18,538
F-statistic	38,427								
Panel B. Years 2–5									
IVF success	0.320	−9,402	−0.013	0.476	−8.690	−0.034	−3,523	0.002	−0.009
	(0.004)	(1,703)	(0.004)	(0.114)	(1.437)	(0.005)	(3,004)	(0.003)	(0.003)
Observations	18,435	18,435	18,435	12,332	12,332	12,332	16,590	18,435	18,435
F-statistic	6,281								
Panel C. Years 6–10									
IVF success	0.227	−6,960	−0.003	0.103	−5.348	−0.021	−5,082	0.003	0.003
	(0.005)	(2,397)	(0.005)	(0.134)	(1.861)	(0.006)	(4,536)	(0.005)	(0.005)
Observations	13,779	13,779	13,779	9,627	9,627	9,627	12,367	13,779	13,779
F-statistic	2,273								
Baseline mean	—	223,038	0.90	28.63	183.01	5.16	301,683	0.05	0.05
Pretreatment effect	—	874	0.010	0.519	−0.061	0.001	−1,298	−0.005	0.001
		(1,811)	(0.004)	(0.375)	(1.162)	(0.005)	(2,800)	(0.003)	(0.003)

Notes: This table shows first-stage and reduced-form regression estimates on the effect of IVF treatment success (0/1) on various outcomes measured at $t = 0$–1, 2–5, and 6–10. In column 1, the coefficient represents the effect of IVF success on the probability of having children during the time period considered. In the reduced-form regressions, the coefficient represents the effect of IVF success on the average of the outcome during the time period considered. Time period $t = 0$ refers to the year of the (potential) child birth. All regressions control for age at first IVF treatment, year of first IVF treatment, pretreatment education, and the pretreatment average of the outcome studied taken over years 1–4 before the first IVF treatment. There are two exceptions. The first-stage regression does not include the pretreatment average because it is zero by construction. The depression regression includes the pretreatment value at $t - 1$ because data on antidepressants are only complete from 1994 onward. The *F*-statistics in the table refers to *F*-tests of the significance of the instrument in the first-stage regressions. The baseline mean refers to the mean of the outcome taken over years 1–4 before the first IVF treatment (for the sample observed 0–1 years after the year of the (potential) childbirth). The pretreatment effect refers to the reduced-form effect of success at first IVF treatment on the pretreatment baseline mean. Robust standard errors are in parentheses.

Features of a “Good” Table

- Information is highly condensed.
  - Only the most important numbers are reported (variable of key interest, F-statistics, observations).
  - Different columns provide very different pieces of information.
  - Use different panels instead of separate tables.
- Contain mean values so the readers can quickly have an idea about the economic significance (in addition to statistical significance).
- Very detailed notes to the end of the table.
  - In most cases, readers do not care about the coefficients in front of the control variables. Just tell them what you control for in the notes.
  - The table is “self-contained.”

Ease of Comparison

- Easier to make comparisons of items within columns than within rows.
- Easier to compare items that are in closer proximity in a table than those which are far apart.
- Application: two adjacent columns $>$ non-adjacent columns/different panels $>$ different tables

Which Table Looks Better?

Table 1: The Effect of the Family Planning Leading Group on Other Provincial Outcomes

Dependent Variable:	Sex Ratio at Birth				FPP Funds (Millions RMB)			
Specification:	Linear-Trend		First-Difference		Linear-Trend		First-Difference	
Period:	1969–1978		1971–1978		1969–1978		1971–1978	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Years since FPL Establishment	0.460	0.473	0.390	0.226	78.434*	80.461*	42.703	36.871
	(0.284)	(0.362)	(1.073)	(1.173)	(40.163)	(45.938)	(25.537)	(27.164)
Province Controls	N	Y	N	Y	N	Y	N	Y
R-Squared	0.213	0.255	0.035	0.064	0.881	0.892	0.278	0.328
Observations	280	280	252	252	224	224	196	196

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are clustered at the provincial level using the wild cluster bootstrap method. Year dummies and province dummies are controlled in all specifications. Control variables include: GDP per capita, share of the non-agricultural population, share of primary industry in GDP, and share of secondary industry in GDP.

Which Table Looks Better?

Table 2: The Effect of the Family Planning Leading Group on Other Provincial Outcomes

Specification:	Linear-Trend				First-Difference			
Dependent Variable:	Sex Ratio at Birth		FPP Funds		Sex Ratio at Birth		FPP Funds	
Period:	1969–1978		1971–1978		1969–1978		1971–1978	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Years since FPL Establishment	0.460	0.473	78.434*	80.461*	0.390	0.226	42.703	36.871
	(0.284)	(0.362)	(40.163)	(45.938)	(1.073)	(1.173)	(25.537)	(27.164)
Province Controls	N	Y	N	Y	N	Y	N	Y
R-Squared	0.213	0.255	0.881	0.892	0.035	0.064	0.278	0.328
Observations	280	280	224	224	252	252	196	196

* significant at 10%; ** significant at 5%; *** significant at 1%. Standard errors are clustered at the provincial level using the wild cluster bootstrap method. Year dummies and province dummies are controlled in all specifications. Control variables include: GDP per capita, share of the non-agricultural population, share of primary industry in GDP, and share of secondary industry in GDP.

A Couple of Words before Going to Stata...

- Data cleaning & data analysis: usually *one simple* command for one task
    - Few alternatives
    - Small learning costs
- Table generating: *many complicated* commands for it
    - -outreg2-, -estout-, -asdoc-, -reg2docx-
    - Learning costs are huge
- Mastering *several* commands > "know about" *all* commands
    - Personal experience: -outreg2- at the beginning, later switched to -estout-, hand-generating as the outside option
- A new learning mode
    - Most of the table-generating commands are so complex that it is unlikely to go through them carefully.
    - Examples + revisions

Commands for Generating Tables

- By “hand”. Clearly, any number in a table comes from somewhere. Therefore, you should be able to generate any kind of table as long as you know how to obtain those numbers.
  - Pros: flexible. (comparative advantage in generating non-standard table)
  - Cons: tedious

Commands for Generating Tables

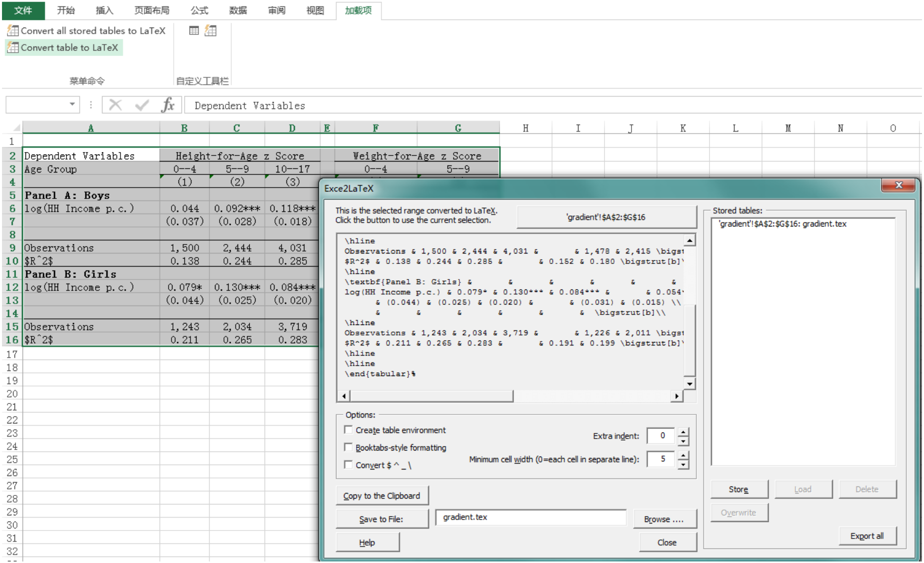
- -outreg2-: easy to learn. But not sufficiently flexible, and does not work well with LaTeX.
  - From the author of -outreg2-: "Philosophically speaking, outreg2 is a research tool to be used DURING research, not AFTER."
  - The readers might only read the table once. But as the researcher, you could read the table for hundreds of times.
- -estout-: arguable the best command out there for table-generating, flexible and work well with LaTeX.
  - The command is more difficult and its full usage requires some knowledge of LaTeX. I'll return to this command afterward.
  - `http://repec.sowi.unibe.ch/stata/estout/esttab.html`

How to Produce a Good-looking Table?

1. Think—The structure of the table. Which variables to report? How to organize the columns? How to organize the panels? Any additional statistics to calculate?
2. Stata—Run regression. Use -outreg2- to store the results. Can combine with -eststo- command from the -estout- package to avoid repeated operations.
    - You can treat each table as independent "inputs" to create a more complicated tables (e.g., multiple panels).
3. Excel—Prepare the headers. Add the lines.
    - Excel is only an intermediary. Latex is our final output.
    - Excel2LaTeX is a very useful Excel plug-in.
    - You can exploit Latex language for special symbols.
    - This step shall be skipped in the end.
4. Latex

Stata Example

Dependent Variables	Height-for-Age z Score			Weight-for-Age z Score	
Age Group	0–4	5–9	10–17	0–4	5–9
	(1)	(2)	(3)	(4)	(5)
Panel A: Boys					
log(HH Income p.c.)	0.044	0.092***	0.118***	0.014	0.033*
	(0.037)	(0.028)	(0.018)	(0.027)	(0.019)
Observations	1,500	2,444	4,031	1,478	2,415
R^2	0.138	0.244	0.285	0.152	0.180
Panel B: Girls					
log(HH Income p.c.)	0.079*	0.130***	0.084***	0.054*	0.040***
	(0.044)	(0.025)	(0.020)	(0.031)	(0.015)
Observations	1,243	2,034	3,719	1,226	2,011
R^2	0.211	0.265	0.283	0.191	0.199



Dependent Variables

	A	B	C	D	E	F	G
1							
2	Dependent Variables	Height-for-Age z Score				Weight-for-Age z Score	
3	Age Group	0--4	5--9	10--17		0--4	5--9
4		(1)	(2)	(3)			
5	**Panel A: Boys**						
6	log(HH Income p.c.)	0.044	0.092***	0.118***			
7		(0.037)	(0.028)	(0.018)			
8							
9	Observations	1,500	2,444	4,031			
10	R^2	0.138	0.244	0.285			
11	**Panel B: Girls**						
12	log(HH Income p.c.)	0.079*	0.130***	0.084***			
13		(0.044)	(0.025)	(0.020)			
14							
15	Observations	1,243	2,034	3,719			
16	R^2	0.211	0.265	0.283			

Exce2LaTeX

This is the selected range converted to LaTeX.
Click the button to use the current selection.

'gradient'!$A$2:$G$16

```
\hline
Observations & 1,500 & 2,444 & 4,031 &       & 1,478 & 2,415 \bigst
$R^2$ & 0.138 & 0.244 & 0.285 &       & 0.152 & 0.180 \bigstrut[b]\
\hline
\textbf{Panel B: Girls} &       &       &       &       &       &
log(HH Income p.c.) & 0.079* & 0.130*** & 0.084*** &       & 0.054
      & (0.044) & (0.025) & (0.020) &       & (0.031) & (0.015) \\
      &       &       &       &       &       & \bigstrut[b]\\
\hline
Observations & 1,243 & 2,034 & 3,719 &       & 1,226 & 2,011 \bigst
$R^2$ & 0.211 & 0.265 & 0.283 &       & 0.191 & 0.199 \bigstrut[b]\
\hline
\hline
\end{tabular}%
```

Options:

- ☐ Create table environment
- ☐ Booktabs-style formatting
- ☐ Convert $ ^ _ \

Extra indent: 0

Minimum cell width (0=each cell in separate line): 5

Save to File: gradient.tex

Stored tables:

'gradient'!$A$2:$G$16: gradient.tex

Outline

1. Data Visualization: Tables or Graphs?
2. Making Tables
3. Plotting Graphs

Figures in academic papers are very different from those in policy reports or in industry presentations.

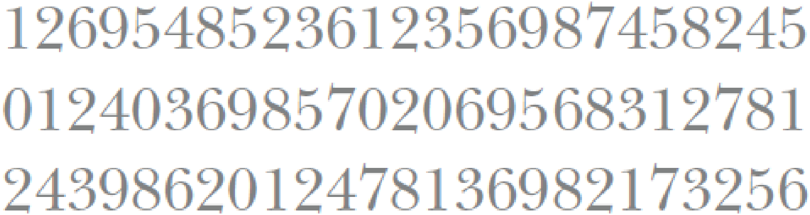
- You do not see those beautiful graphs with little information.
- Such as pie graph.

What's the difference between graphs and tables?

- A picture was worth a thousand words.
- An effective graph should tap into the brain's "pre-attentive visual processing" (Few 2004; Healey and Enns 2012). Pre-attentive processing allows the reader to perceive multiple basic visual elements simultaneously.

Reading Material 6.1 "An Economist's Guide to Visualizing Data" (Schwabish 2014)

Find Number 3



1269548523612356987458245
0124036985702069568312781
2439862012478136982173256

Find Number 3

12695485 2**3**612**3**56987458245

01240**3**6985702069568**3**12781

24**3**986201247 81**3**698217**3**256

Again, before we proceed to drawing graphs, let's think about following questions —

What makes a good graph? What makes a bad graph?

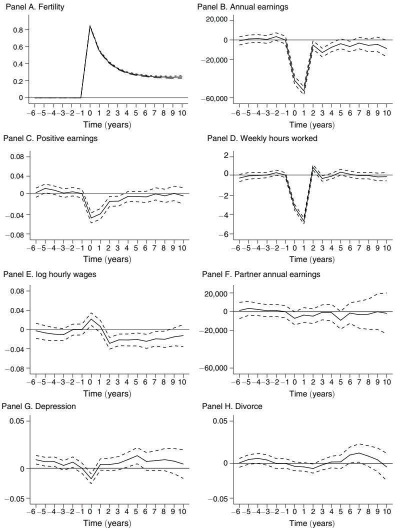


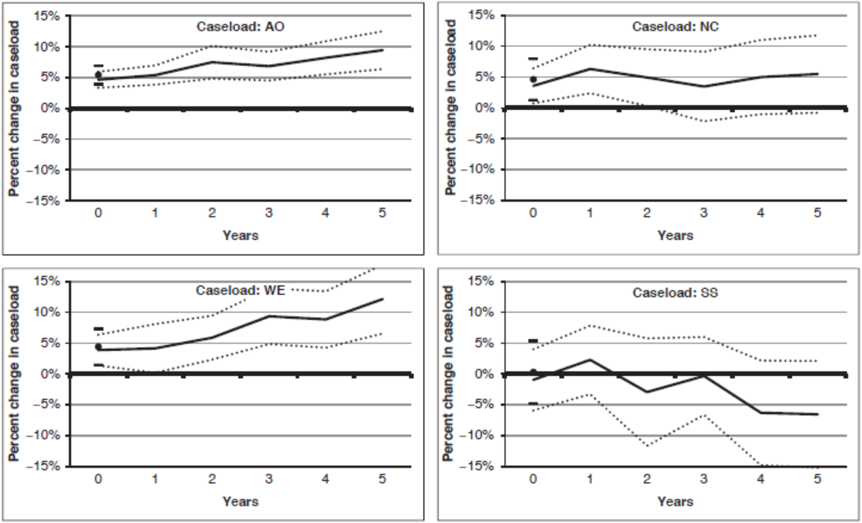
Figure 2. Event Study Graphs of IVF Treatment Effects on Various Outcomes

Notes: The figures plot coefficients from an event-study analysis. Event time is defined as years before and after (potential) childbirth. The coefficients in period −1 are normalized to 0. The models are estimated for the sample of IVF treated women who had their first IVF treatment between 1995 and 2005. The coefficients in panel F are estimated for the partners of the IVF treated women. For details on the model, see the text.

Principles of “Good” Figures (Schwabish 2014)

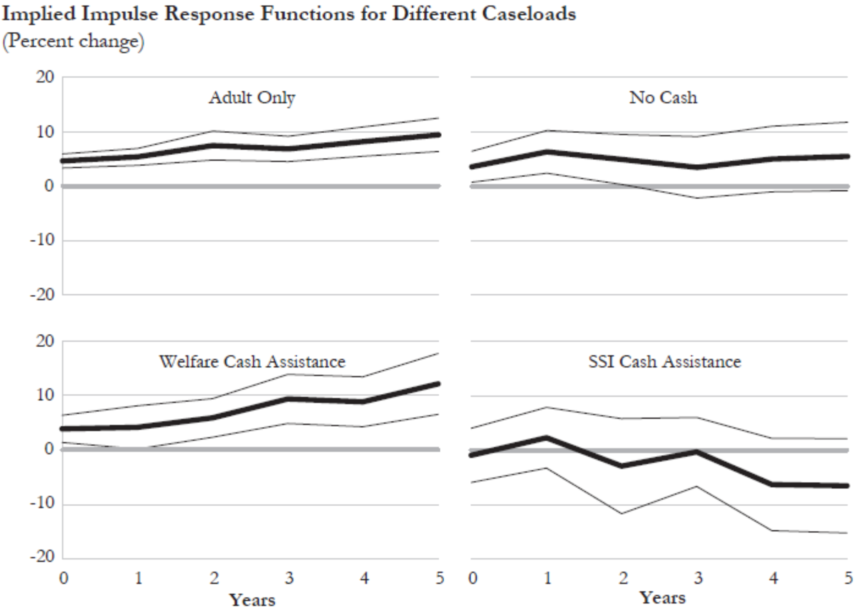
1. Show the data. Data should be presented in the *clearest* way possible.
2. Reduce the clutter. Clutter—unnecessary or distracting visual elements.
3. Integrate the text and the graph. *Complement the text* and at the same time to contain enough information to *stand alone*.

Application 1, Original Graph

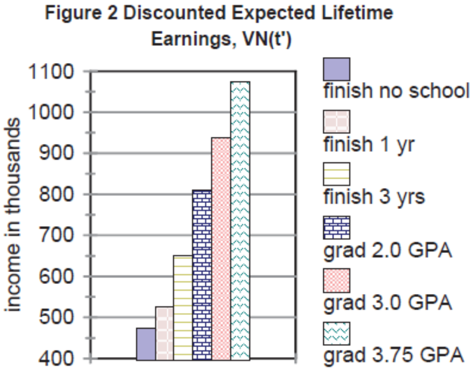


1. Emphasize the data. Not the 0 percent gridline.
2. Clutter. y axis labels and percentage sign.
3. Integrate the text and the graph. What do AO, NC, WE, and SS mean in the figure?

Application 1, Revised Graph



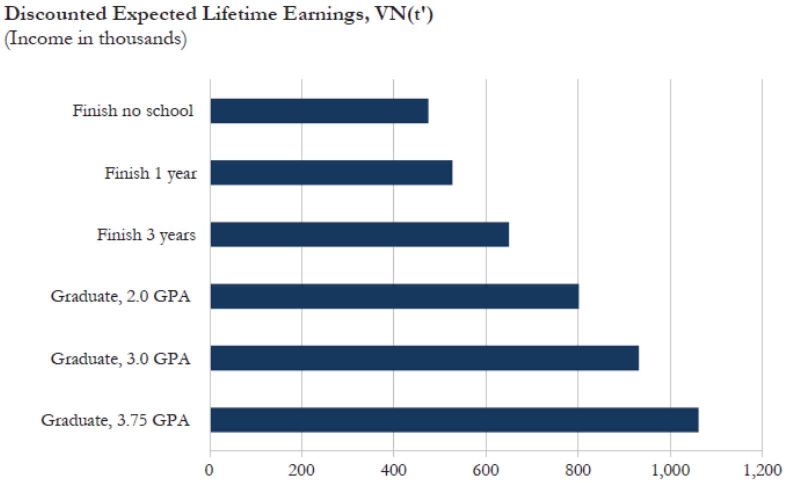
Application 2, Original Graph



Source: Stinebrickner and Stinebrickner (2013).

1. Show the data.
   - When it comes to bar and column charts, a first rule is to start the chart at zero.
   - Otherwise, the differences between the columns are overemphasized.
2. Clutter.
   - Too often graphs use textured or filled gradients when simple shades of a color could accomplish the same task.
3. Integrate the text and the graph.
   - Readers need to find out the bar and figure out the corresponding legend.

Application 2, Revised Graph



Source: Author's calculations using numbers inferred from text in Stinebrickner and Stinebrickner (2013).

How to Plot a Graph?

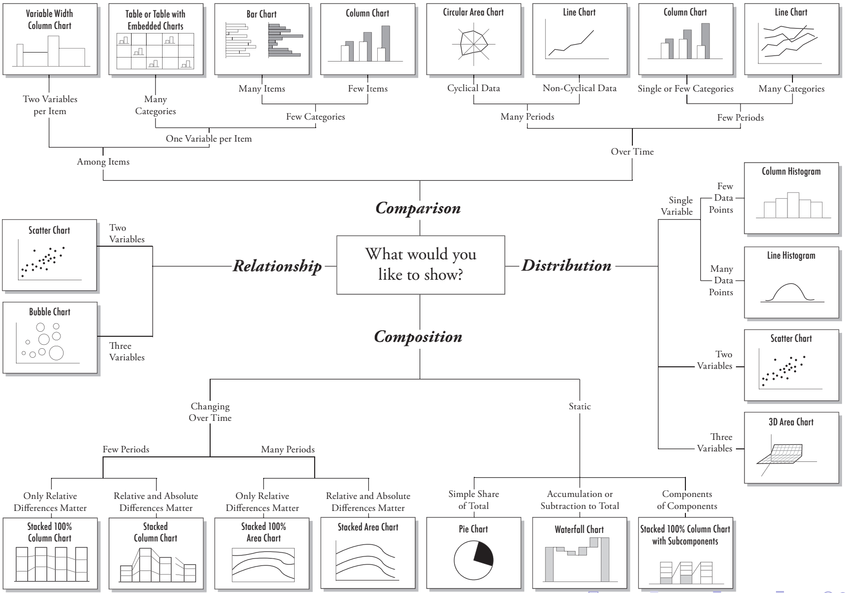
There are numerous types of graph and we are not going to show them one by one

- Read Mitchell (2016)—a 700 pages book on Stata graphics!

Instead, we will work through general steps about plotting a graph.

- Decide type of graph
- Adjust the details
- Combine multiple graphs

Chart Suggestions—A Thought-Starter



What would you like to show?

Comparison

Among Items

Two Variables per Item

Variable Width Column Chart

One Variable per Item

Many Categories

Table or Table with Embedded Charts

Few Categories

Many Items

Bar Chart

Few Items

Column Chart

Over Time

Many Periods

Cyclical Data

Circular Area Chart

Non-Cyclical Data

Line Chart

Few Periods

Single or Few Categories

Column Chart

Many Categories

Line Chart

Relationship

Two Variables

Scatter Chart

Three Variables

Bubble Chart

Distribution

Single Variable

Few Data Points

Column Histogram

Many Data Points

Line Histogram

Two Variables

Scatter Chart

Three Variables

3D Area Chart

Composition

Changing Over Time

Few Periods

Only Relative Differences Matter

Stacked 100% Column Chart

Relative and Absolute Differences Matter

Stacked Column Chart

Many Periods

Only Relative Differences Matter

Stacked 100% Area Chart

Relative and Absolute Differences Matter

Stacked Area Chart

Static

Simple Share of Total

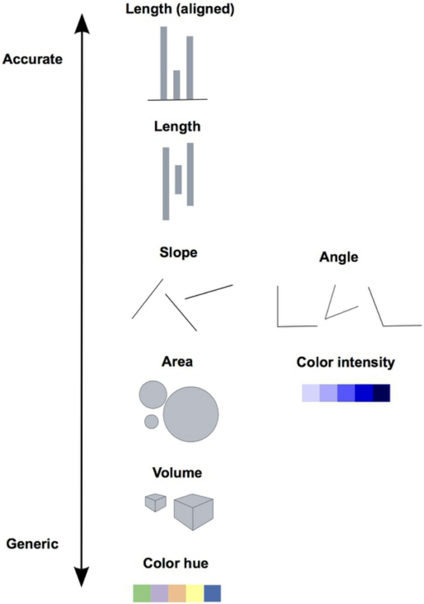
Pie Chart

Accumulation or Subtraction to Total

Waterfall Chart

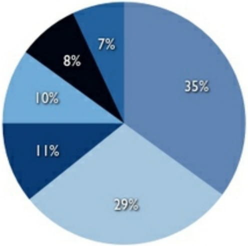
Components of Components

Stacked 100% Column Chart with Subcomponents

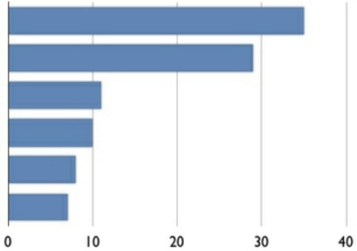


- Q: should we always use bar plots?
- Try to balance readers' effort in interpreting each graph.
  - Accurate information requires more effect.
  - Use generic graphs when you have lots of potential numbers in a graph (e.g., map).

Grudge Match



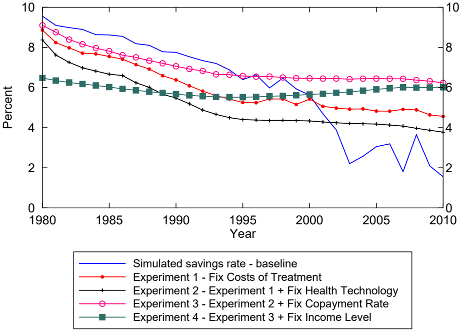
vs.



Different Types of Graphs

- Stata offers a huge variety of different graphs. Type -help graph- and -help twoway- for details.
- Two useful websites that give you an idea how each type of graph looks like.
    - https://www.stata.com/support/faqs/graphics/gph/stata-graphs/
    - https://www.surveydesign.com.au/tipsgraphs.html
- In this lecture, we will focus on lines. -twoway line- and -twoway connected-

Make Your Graphs “Clear”—in Color and in Black & White

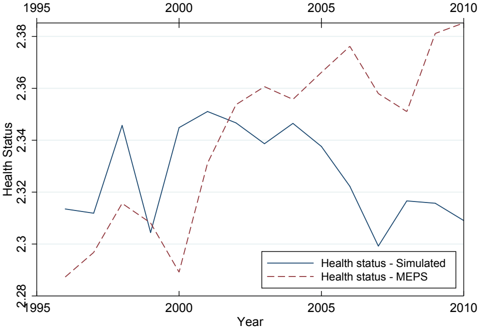


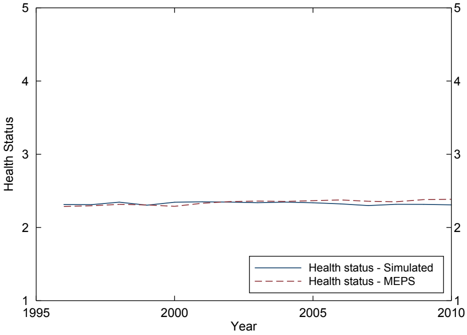
The readers should be able to distinguish between different lines if they print the paper in black and white.

And the graph should be even more clear in a presentation.

Graphs can be More Misleading than Tables

Even if the numbers are the same, the impression upon people can be very different if presented in different ways.





Several principles of choosing the appropriate scale:

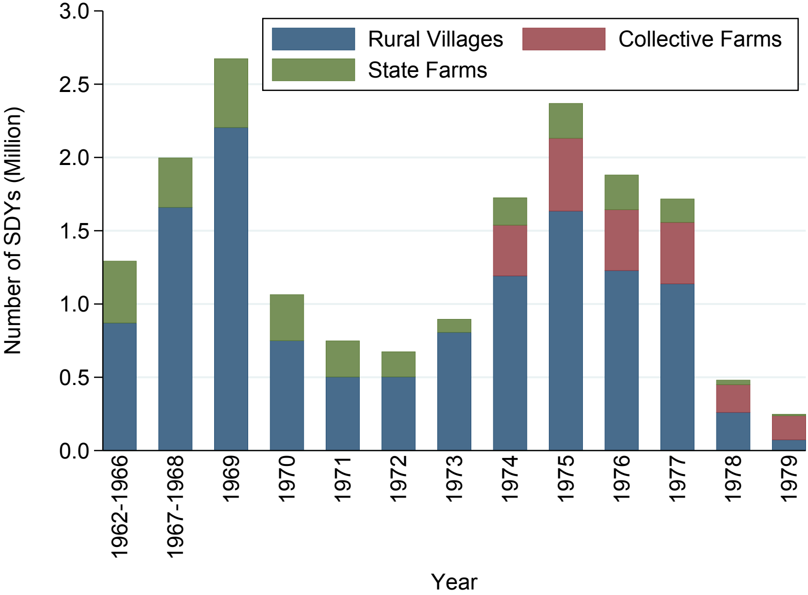
1. Min–Max (could be problematic if the variation in the y variable is small)
2. Natural scale (e.g., 0 to 1 may be a natural scale for probabilities)
3. Presenting the S.D./S.E. is also helpful.

Several Useful Tips

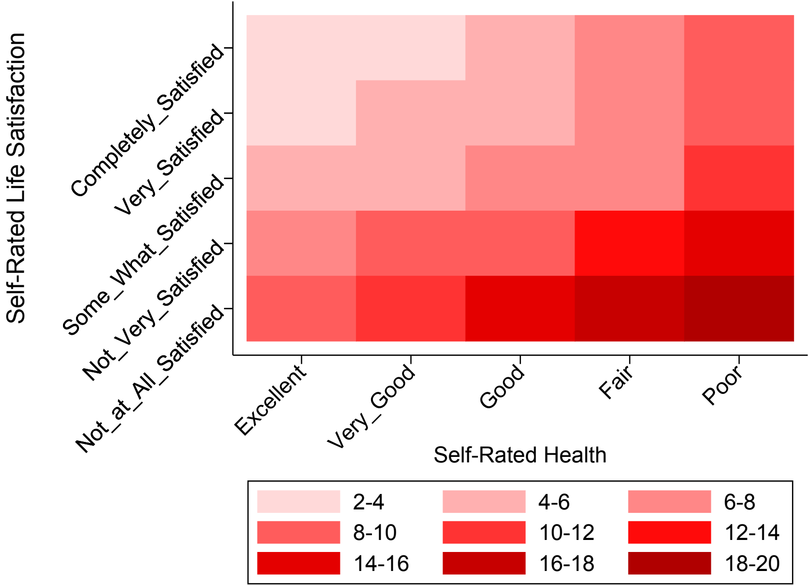
- Multiple lines
    - Add external information to the graph
- Adjust the details
- Two y-axis
- Graph by group
- Combine multiple graphs

Four Stata example

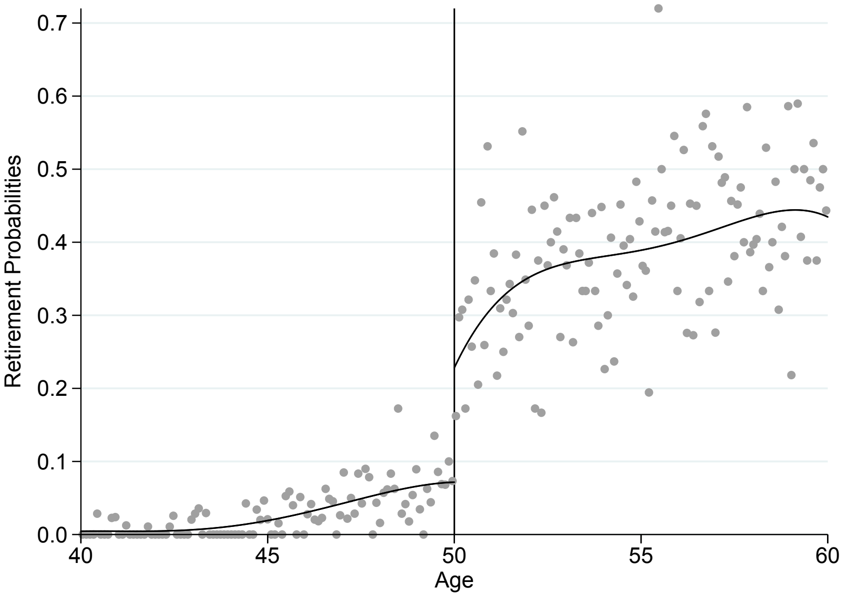
More Examples—Bar Graph



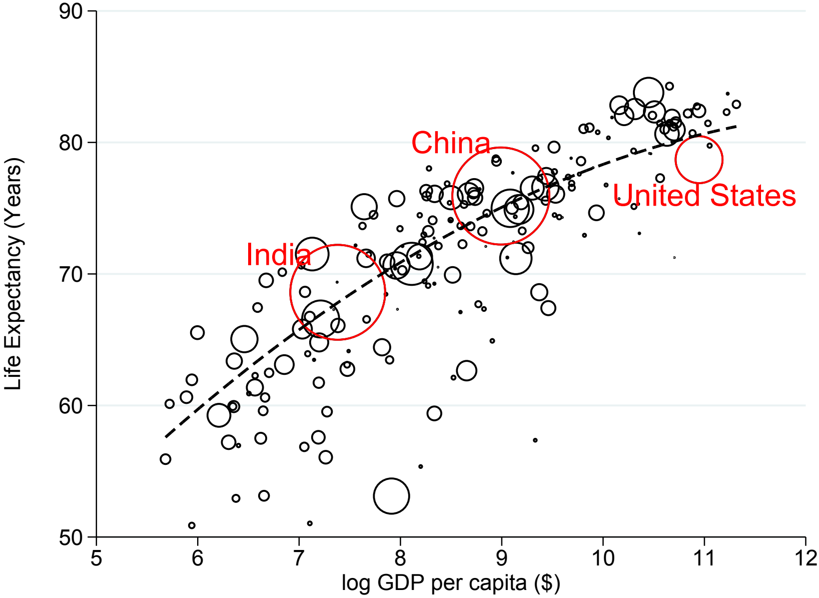
More Examples—Matrix Graph



More Examples—Regression Discontinuity



More Examples—Bubble Plot



Plotting Graphs with Softwares Other than Stata

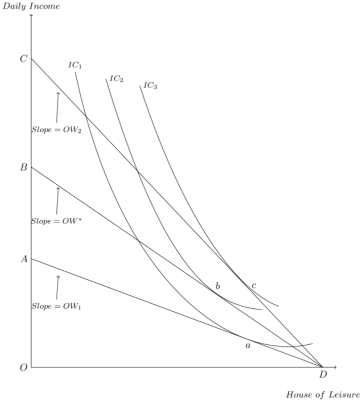
Strictly speaking, all previous graphs are “statistical” graphs—they visualize some statistics.

But sometimes we need other types of graphs as well.

- Indifference curve
- Flow chart (e.g., sampling process)

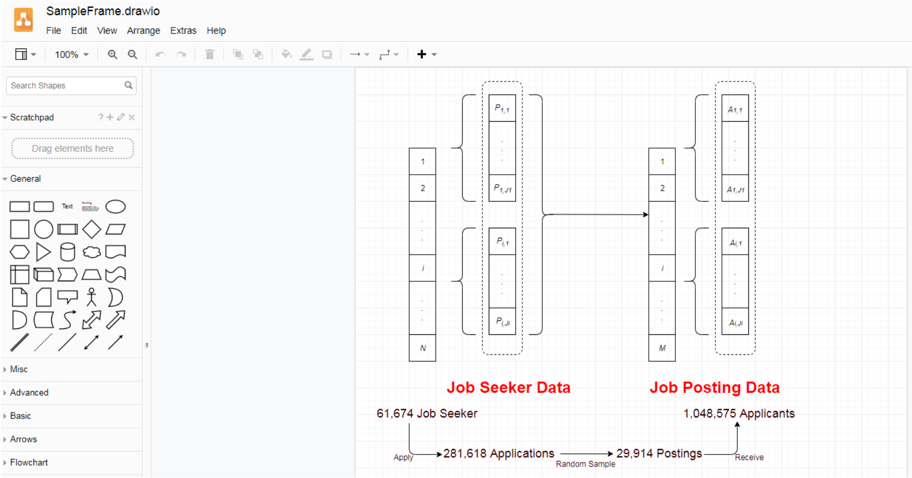
Drawing Indifference Curves: TikZ Package

- Graphs can not only be used to visualize data, but also to illustrate theory or idea. e.g., labor supply decision



- TikZ package, large learning costs (check out Prof. Chiu Yu Ko's website `https://sites.google.com/site/kochiuyu/Tikz`)

Flow Chart, https://www.draw.io/


SampleFrame.drawio
File
Edit
View
Arrange
Extras
Help
100%
Search Shapes
Scratchpad
Drag elements here
General
Text
Misc
Advanced
Basic
Arrows
Flowchart
1
2
i
N
$P_{1,1}$
$P_{1,J1}$
$P_{i,1}$
$P_{i,Ji}$
1
2
i
M
$A_{1,1}$
$A_{1,J1}$
$A_{i,1}$
$A_{i,Ji}$
Job Seeker Data
Job Posting Data
61,674 Job Seeker
1,048,575 Applicants
Apply
281,618 Applications
Random Sample
29,914 Postings
Receive