

Williams College ECON 460:  
 Women, Work, and the World Economy

**Method Monday 1: How Economists Work with Data**

Professor: Pamela Jakiela

# Outline

- The empirical research pipeline
- Summary statistics
- Visualizing one variable
- Visualizing the relationship between two variables
- Regression tables

# How Economists Work

Finding and analyzing data (data wrangling and statistics, usually in Stata or R)

- Pipeline: finding/getting data, cleaning/wrangling data, analyzing data
- All steps in the process are scripted (e.g. in do files) so as to be replicable
- Final products (e.g. visualizations and regression tables) are polished and self-contained

Communicating the results (writing about your tables/figures, usually in word or L<sup>A</sup>T<sub>E</sub>X)

- Academic writing uses first-person, but in the most matter-of-fact way possible
  - Some (fairly limited) variation in chattiness, citation practices across (e.g.) papers vs. blogs
- References (quotes, figures, summaries/findings) are cited in the text and the bibliography
- The American Economic Association follows the Chicago Manual of Style, so we do too

# How Economists Work: Stata vs. R

All of the empirical analysis in ECON 460 can be done in either Stata or R, but I'll teach Stata

- Most economists use Stata; most replication packages for economics papers are in Stata
  - ▶ R, MATLAB, and Python are also used by many economists, but Stata still dominates
  - ▶ R is increasingly the primary tool for statistical computing in political science, stats, etc.
- ECON 255 is taught in Stata, and Williams ECON majors should all know how to use it
- Stata is easier for beginners, and for relatively straightforward statistical tasks

I use both, and I am happy to accept assignments as either Stata do files or R scripts

- Williams ECON's Stata tutorials: <https://pjakiela.github.io/stata/>

# How Economists Work: Word vs. $\text{\LaTeX}$

$\text{\LaTeX}$  is “a document preparation system used for the communication and publication of scientific documents” i.e. a markup language for technical documents ( $\text{\LaTeX}$ Project 2024)

- .tex files include both content and formatting code that is executed upon compilation
- $\text{\LaTeX}$  is particularly useful for formatting mathematical equations
- $\text{\LaTeX}$  also allows you to link to external (tex) tables and (pdf, png, or jpg) figures/images, so that you automatically include the most recently-updated version each time you compile

It is probably not worth learning  $\text{\LaTeX}$  unless you intend to write a thesis, do a PhD in economics, or otherwise have a career that involves writing mathematically-oriented documents

- If you are wondering why I am so bad at exporting results to word, this is why

# A Summary Statistics Table

Table A1: Summary Statistics on Three Year Olds

	MEAN	S.D.	MEDIAN	MIN.	MAX.	N
Child age (in months)	41.70	3.37	42	36	47	634
Height-for-age z-score	-0.57	1.44	-0.66	-4	4	634
Child is male	0.49	0.50	0	0	1	634
Mother is child's primary caregiver	0.88	0.33	1	0	1	634
Mother's education in years	7.88	2.48	8	0	13	634
Mother is Luo	0.95	0.22	1	0	1	634
Father absent from household	0.14	0.35	0	0	1	625
Father's education in years	8.49	2.67	8	0	13	568
Father is Luo	0.98	0.12	1	0	1	568
Household size	5.80	1.98	6	2	14	634
Older siblings in household	1.40	1.29	1	0	6	634
Asset index (out of 10)	3.43	1.47	3	0	9	634
Distance to school (in km)	0.44	0.17	0.46	0.05	0.75	634
Child is enrolled in school	0.60	0.49	1	0	1	634

Data on 634 children aged 36 to 47 months. Children are from 622 unique households (12 households include two three-year-old children). ASSET INDEX is the sum of indicators for having a cement floor, iron roof, latrine, or connection to the electricity grid, and indicators for owning a motorized vehicle, a bicycle, a television, a mobile phone, a computer, or a radio.

Source: Jakiela, Ozier, Fernald, and Knauer (2024)

## What's Wrong With This Picture?

Statistic	N	Mean	St. Dev.	Min	Max
Female	812	1.49	0.50	1	2
Age	812	35.72	22.70	-99	60
Education	812	3.00	1.41	1	5
Married	812	0.84	0.37	0	1
Income	683	55.98	26.42	20.18	180.58

# Why a Summary Statistics Table Is Important

A summary statistics table allows you and your readers to check for data preparation errors

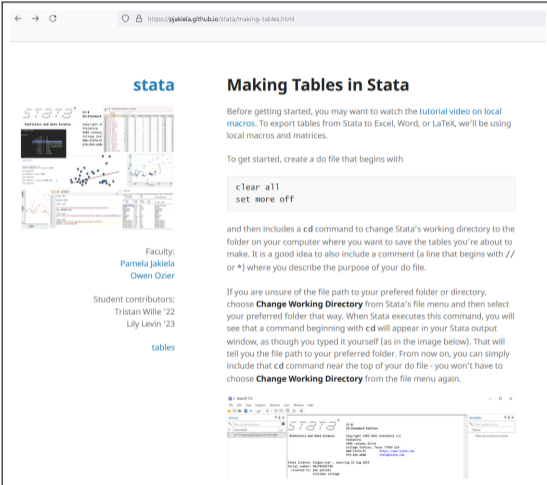
- Should include: mean, SD, min, max,  $N$  for your analysis sample
- May also include: median, additional percentiles, sub-group means, tests of equality
- Variables and columns must be labeled clearly (no varnames), table notes must explain data sources and unclear aspects of variable construction, table must be self-contained

Things to look out for:

- Varying sample sizes, unexplained missing values
- Unreasonable minimum and maximum values
- Categorical variables treated as numeric (e.g. marital status, state ID)



# Making Tables in Stata: Williams ECON'S Online Tutorials



← → ↻ <https://pjakiel.github.io/stata/making-tables.html>

## stata

### Making Tables in Stata

Before getting started, you may want to watch the [tutorial video on local macros](#). To export tables from Stata to Excel, Word, or LaTeX, we'll be using local macros and matrices.

To get started, create a do file that begins with

```
clear all
set more off
```


and then includes a `cd` command to change Stata's working directory to the folder on your computer where you want to save the tables you're about to make. It is a good idea to also include a comment (a line that begins with `//` or `*`) where you describe the purpose of your do file.

If you're unsure of the file path to your preferred folder or directory, choose **Change Working Directory** from Stata's file menu and then select your preferred folder that way. When Stata executes this command, you will see that a command beginning with `cd` will appear in your Stata output window, as though you typed it yourself (as in the image below). That will tell you the file path to your preferred folder. From now on, you can simply include that `cd` command near the top of your do file - you won't have to choose **Change Working Directory** from the file menu again.

Faculty:  
Pamela Jakiela  
Owen Ozier

Student contributors:  
Tristan Wille '22  
Lily Levin '23

tables



# Stata Commands for Making Summary Statistics Tables

Useful commands for making reproducible summary statistics tables in Stata

- `summarize` calculates summary statistics, the `, detail` option expands the list
  - ▶ Use `return list` after `summarize` to see which statistics are saved as locals
- `estpost summarize` stores estimated summary statistics
- `esttab` makes a table of stored estimates

Remember:

- Use `help [command]` to access Stata help files
- <https://pjakiela.github.io/stata/making-tables.html> walks you through the steps you'll need to follow to make a professional-looking table of summary statistics
- Write a do file that makes each table so that you can easily replicate/modify the process

## Making a Summary Statistics Table with esttab

```
summarize b_* // look at summary statistics for variables starting with b_  
estpost summarize b_* // store summary statistics for variables starting with b_  
esttab, cells("mean sd min max count") // make a summary statistics table
```

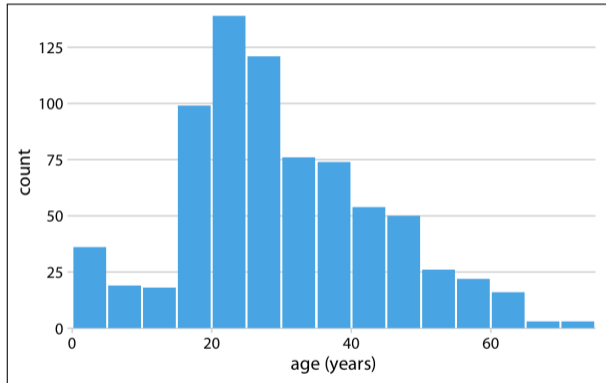
	mean	sd	min	max	count
b_h_edu	5.300699	3.940068	0	16	572
b_knowledg~t	.5391304	.4989005	0	1	575
b_hh_size	5.466087	2.48658	1	14	575
b_acres	2.270779	2.697221	0	34	462
b_dist_km	1.667154	.9307161	.0332327	3.982576	574
b_h_age_im~d	39.2537	15.3422	17	88	575
b_h_age_mi~g	.0504348	.2190309	0	1	575
N	575				

# Making a Summary Statistics Table with esttab

```
esttab, cells("mean(fmt(2)) sd min max count(fmt(%9.0g))") ///  
label noobs nonum nomtitle varwidth(28) ///  
collabels("Mean" "S.D." "Min." "Max." "N") ///  
title(Summary Statistics Table)
```

	Mean	S.D.	Min.	Max.	N
Education	5.30	3.94	0.00	16.00	572
Malaria Knowledge	0.54	0.50	0.00	1.00	575
Household Size	5.47	2.49	1.00	14.00	575
Acres of Land	2.27	2.70	0.00	34.00	462
Distance to Health Center	1.67	0.93	0.03	3.98	574
Age	39.25	15.34	17.00	88.00	575
Age Data Missing	0.05	0.22	0.00	1.00	575

# Histograms



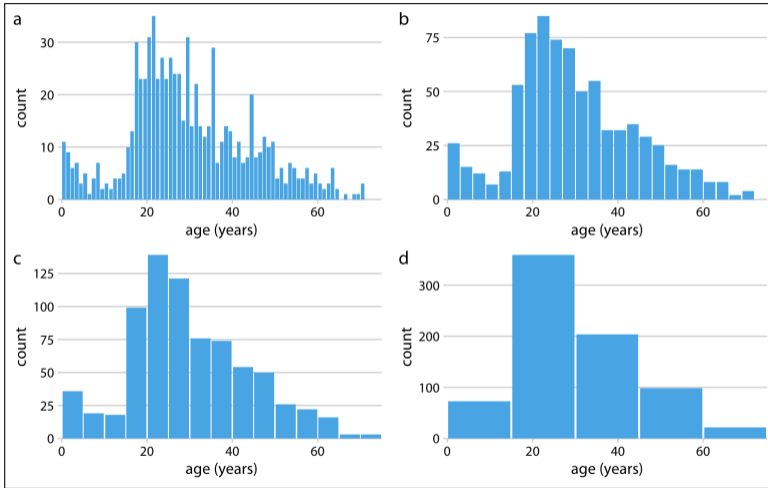
*Source: data Wilke (2019) on the ages of Titanic passengers*

# Stata's histogram Command Makes Histogram

A **histogram** is a bar graph that summarizes the distribution of a variable by dividing its range into equally-sized bins, counting the observations in each bin, and plotting the resulting counts

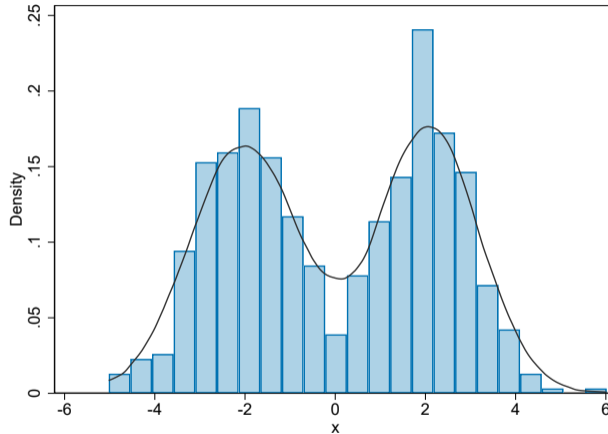
- The option `discrete` can be used for variables that take on a small set of discrete values
- Options for the scale of the  $y$ -axis:
  - ▶ `frequency`: bar heights are counts of observations in each bin
  - ▶ `density`: bar heights are scaled so that their total area sums to one (default)
  - ▶ `fraction`: bar heights are scaled so that their heights sum to none
- Control the number of bins using the `bins(#)` or `width(#)` options
  - ▶ Under-smoothing leaves distracting (excess) variability in bar heights
  - ▶ Over-smoothing can obscure important information about the distribution

# Choosing an Appropriate Bin Width



Source: data Wilke (2019) on the ages of Titanic passengers

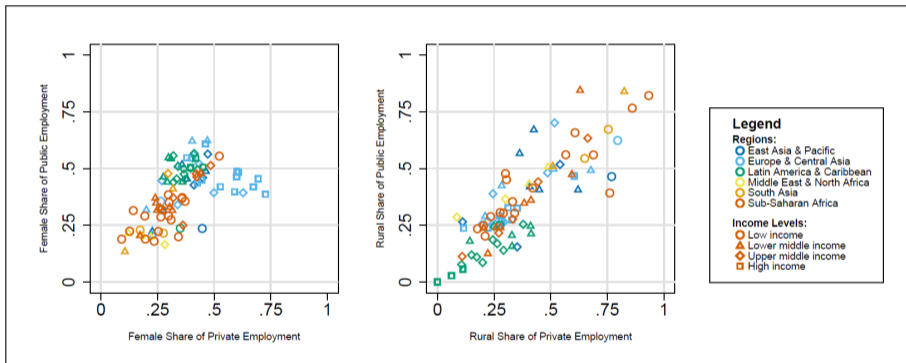
# Adding a Kernel Density Estimate



```
histogram x, width(0.48) kdensity color(sea) fcolor(sea*0.32) xlabel(-6(2)6)
```



# Scatter Plots



Source: Jakiela (2018) Center for Global Development blog post on World Bank's Worldwide Bureaucracy Indicators Database

# Stata's `twoway` Command Is a Blank Canvas for Two-Variable Graphs

Stata's `twoway` allows you to overlay (multiple) scatter plots, line graphs, and regression fits

- Syntax: `twoway ([plottype] y1 x1, options) ([plottype] y2 x2, options)`
- Common `twoway` plots:
  - ▶ `scatter`: scatter plot of the  $(x, y)$  points in the data
  - ▶ `connected`: line graph connecting (ordered)  $(x, y)$  points
  - ▶ `lfit/lfitci`: linear regression fit of  $y$  on  $x$
  - ▶ `lpoly/lpolyci`: locally-weighted polynomial regression fit of  $y$  on  $x$
- Important `twoway` options
  - ▶ Size, color, and symbols used in scatter plots can convey additional attributes of data

# Regression Tables

Table A16: OLS Regressions of Gender Differences in Labor Force Participation in Africa

<i>Dependent variable:</i>	IN LABOR FORCE	
	OLS (1)	OLS (2)
<i>Specification:</i>		
Female × gender language	-0.17 (0.05) [0.001]	-0.15 (0.03) [ $p < 0.001$ ]
Native language is a gender language	-0.08 (0.02) [ $p < 0.001$ ]	-0.04 (0.02) [0.011]
Female	-0.10 (0.01) [ $p < 0.001$ ]	-0.15 (0.03) [ $p < 0.001$ ]
Country-Wave Fixed Effects	No	Yes
Individual Controls	No	Yes
Ethnography Controls	No	Yes
Observations	26328	26328
$R^2$	0.04	0.12

Robust standard errors clustered at the language level. The dependent variable is an indicator for being in the labor force (either working for a wage, self-employed, or actively seeking employment). Data is from Afrobarometer Rounds 2 through 5. The analysis includes data from Kenya, Niger, Nigeria, and Uganda; Niger was only added to the Afrobarometer in Round 5, while the other countries appear in all four rounds. Individual controls are age and age-squared and indicators for being identifying as Muslim, Catholic, Protestant, or another religion, plus interactions between these controls and the female dummy. Ethnographic controls are characteristics of pre-industrial societies identified by lasso as predictors of the use of gender languages (use of horses and/or camels, use of the plough, and regular milking of domestic animals).

Source: Jakiela and Ozier (2021)

# “Publication-Ready” Regression Tables

Requirements for a “publication-ready” regression table:

- Variables labeled, labels are self-explanatory, notes explain variable construction
- Columns numbered, labeled with info about specification and/or differences between them
- Number of decimal places is reasonable, consistent (and significant digits make sense)
- Standard errors in parentheses, either p-values in square brackets or significance stars
- Sample size is constant across table
- Title indicates content of table
- Dependent variable(s) in title or table header
- Controls indicated in body of table or table notes

## “Publication-Ready” Regression Tables with `eststo` and `esttab`

It is very, very easy to make replicable regression tables in Stata with `eststo` and `esttab`:

```
eststo clear
eststo:  reg y x1
eststo:  reg y x1 x2
esttab
```

Options to improve the appearance of your tables:

- `label`: use variable labels instead of variable names
- `mtitles()`: provide a list of titles for the columns/specifications
- `se(#)`: report standard errors rather than t-statistics, # of digits to report
- `indicate(Text = vars)`: omit some coefficients from table (e.g. fixed effects)
- `varwidth(#)`, `modelwidth(#)`: set width of columns

# Summary

Almost every economics paper will include:

- A summary statistics table
- A histogram or scatter plot (or both) showing the distribution of key variables
- A regression table

Practice transparent, replicable social science by creating do files that load, clean, merge, and analyze your data – and allow you to quickly repeat and/or modify the entire process as needed