Intro to Metrics & Data Analysis with R

ECON 490

Taylor Mackay || Email: tmackay@fullerton.edu

Overview

Stats material covered in these slides:

- Reviewing basic definitions from statistics
- Defining and describing distributions
- Defining outcome and explanatory variables

Programming material:

- Getting started with R + RStudio
- Setting up Swirl activities
- Introduction to R

Working with Data

In broad terms, what is the goal of working with data?

- Lots of potential answers…
- Trying to understand the world, make predictions, etc.

For this class (and economics broadly), we're interested in describing ***relationships***

- What is the effect of this policy on employment?
- How did this marketing program affect sales?
- How do market conditions impact user retention?

Basic Terminology

Whenever we refer to data in this class, we mean something we can observe

- Sounds obvious...
- But we'll see why this matters later

Think of data in a spreadsheet format:

- Rows of your data are ***observations***
- Columns of your data are ***variables***

	state_name	age	employed	hhincome
1	california	45	1	102000
2	california	48	1	254000
3	california	2	0	360000
4	california	50	1	335300
5	california	25	1	133800
6	california	40	1	210000
7	california	5	0	157000
8	california	62	1	121600

Different Types of Variables

From the last slide, we have demographic data on a sample of people in CA

- Variables included employment status, age, and household income
- Let's characterize these variables by the values they can take

Continuous variables can take on any value (possibly within some range)

- Income is continuous – it can be $1,000 or $1,001, or $10,642.10…

Discrete variables take on a limited number of values

- They might represent count data or qualitative data
- From the last slide, Employed is discrete – it's either 0 or 1

Discrete Variables

In this class (and data analysis more generally), discrete data is everywhere

- How we handle discrete data depends on what information it contains

Factor variables contain qualitative information

- Different ***levels*** of a factor variable correspond to different characteristics
- E.g., education variable with 1 = non-HS grad, 2 = HS-grad, 3 = college-grad

Binary variables are factor variables with exactly two levels (think, “yes” or “no”)

- For this class, binary variables will always equal 0 or 1
- Depending on context, might refer to them as dummy or indicator variables

Distributions

From the last slide, we have demographic data on a sample of people in CA

- Describes characteristics like employment status, age, and household income
- Not surprisingly, these factors can differ a lot across people!

The ***distribution*** of a variable tells you how often that variable takes on a given value

Whether a variable is discrete or continuous determines how we visualize it

Discrete Distributions

education	257448	
... No HS	77414	30%
... HS Grad	112262	44%
... College Grad	67772	26%

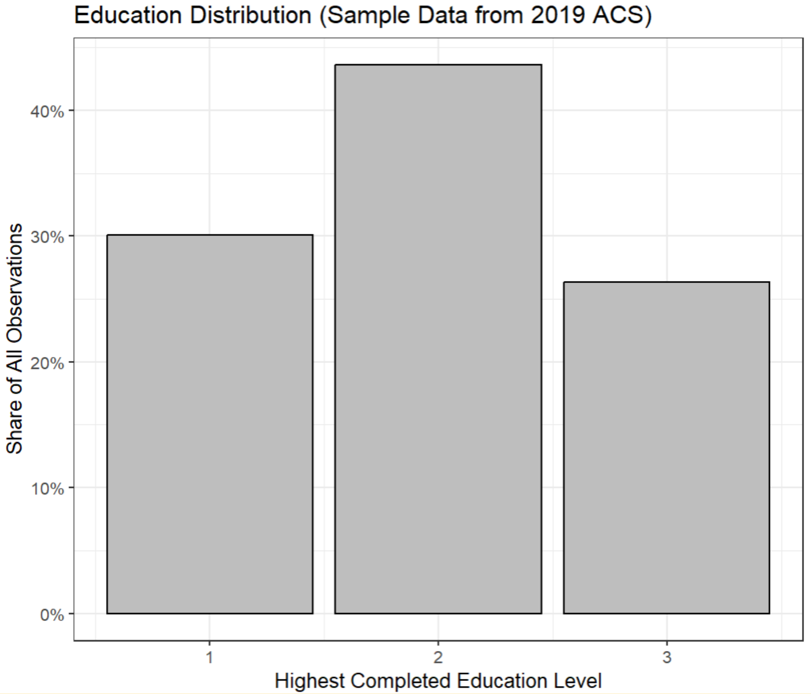
Plotting discrete distributions is easy

- Just calculate the percentage of observations that take on each value

Education is a factor variable with 3 levels

Two options to show this distribution:

1. Show the percentages in a table
2. Create a bar graph



Continuous Distributions

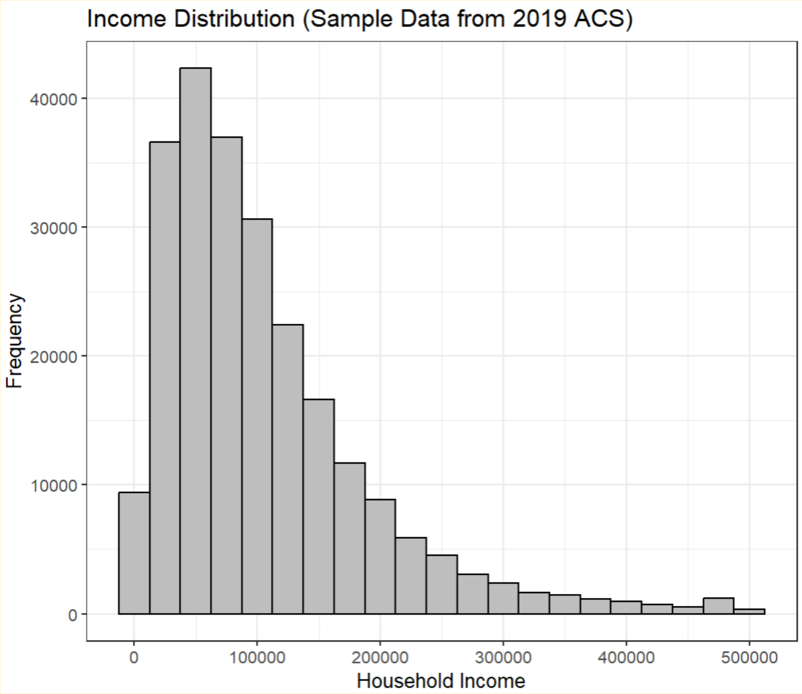
Plotting continuous distributions is a bit trickier

- Income could be 100.00, or 100.01, or …
- How do we create a table for each value?

Use ***histograms*** to plot continuous distributions

Divide values into ***bins***, then show proportion of observations falling into each bin

- In effect, make the variable discrete to facilitate plotting (just like on the last slide)



Histogram with bins equal to $25,000 increments

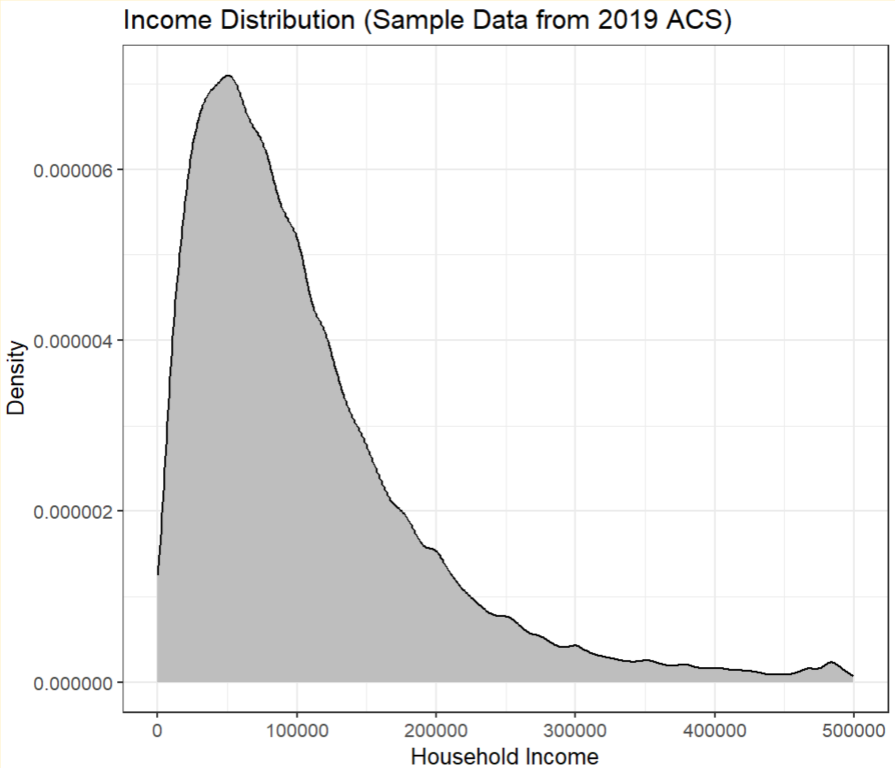
Density Plots for Continuous Distributions

On the last slide, created bins with intervals of $25,000

- What if we made smaller bins?
- E.g., $10,000, or $1,000, or…

We can think about taking the limit of this thinking… where do we wind up?

The result is a ***density plot***



Summarizing Distributions

Distributions contain a lot of information – how can we summarize them?

Two common methods are the ***mean*** (or average) and ***median***

- Descriptions of central tendency = ways of picking a “representative” value

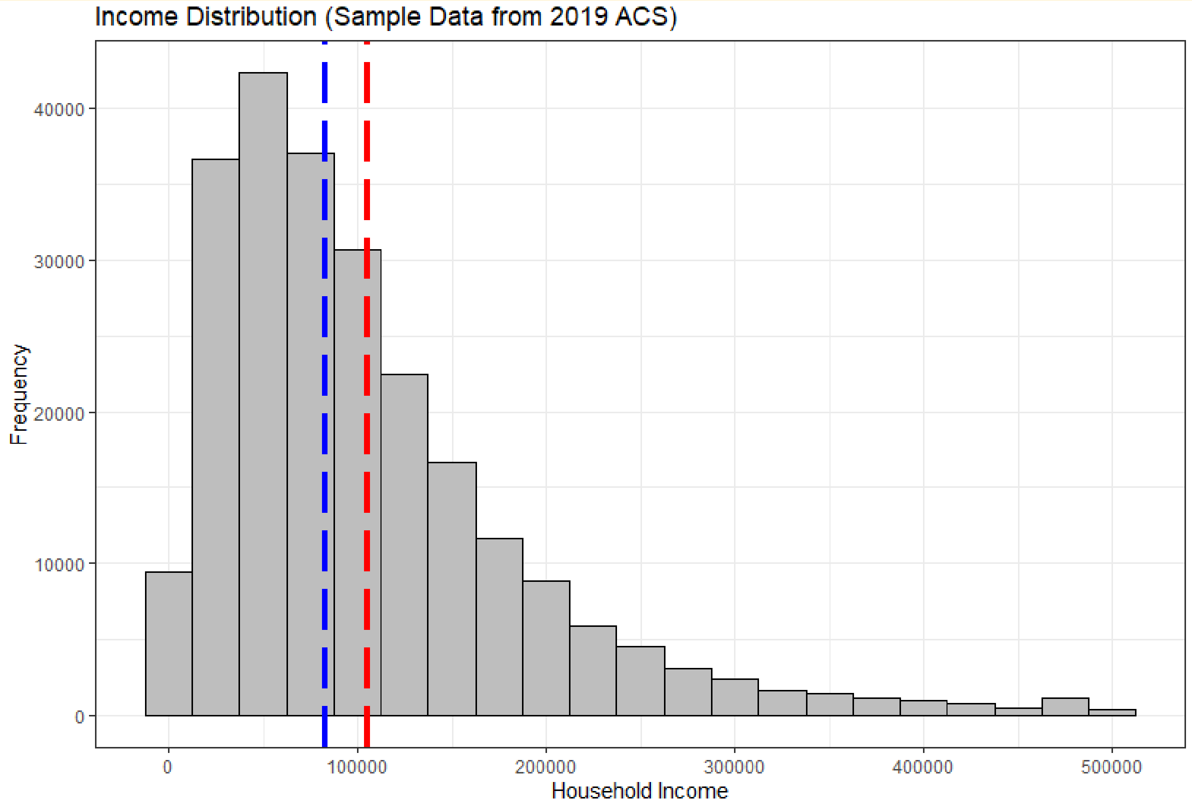
Percentiles are another useful way of characterizing distributions

- The Xth percentile of a variable tells us the value for which X percent of observations are less
- The median is the 50^{th} percentile – 50% of observations are < the median

Mean vs. Median

Mean income (in red) ~ $105,000

Median income (in blue) ~ $83,200



Variance and Standard Deviation

We can summarize the variability or “spread” of a variable using ***variance***

Suppose we have a variable Y (a column of data) with n observations (rows in our data) and average value of $\bar{Y}$ → we can define the variance of Y as:

$$Var(Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

The ***standard deviation*** of Y is the square root of its variance → $SD(Y) = \sqrt{Var(Y)}$

Thinking about Relationships

Up to this point, we've talked about characterizing single variables

In economics (and most jobs), what we're interested in is ***relationships***

- How is one variable related to another?
- How is changing one variable likely to impact another variable?

Next week, we'll talk about how to answer these questions

- Tonight, we'll just introduce several important definitions

Characterizing Relationships

Covariance is one way of describing the relationship between two variables

- Given 2 columns of data, Y and X, with n rows and avg. values $\bar{Y}$ and $\bar{X}$:

$$Cov(Y,X) = \sum_{i=1}^{n} (Y_i - \bar{Y})(X_i - \bar{X})$$

When the covariance between Y and X is:

- ***Positive*** it means when Y is ***higher*** than average, X tends to be ***higher*** as well
- ***Negative*** it means when Y is ***higher*** than average, X tends to be ***lower***

Correlation

Covariance depends on the scale of X and Y – makes interpretation tricky

- ***Correlation*** is a way of measuring relationships without scale or units
- Correlations range between -1 and 1

Don't need to know the formula for correlation – in R, use `corr()`

- Key point is interpretation → 0 implies no (linear) relationship
- Values closer to -1 or 1 imply a stronger relationship between Y and X

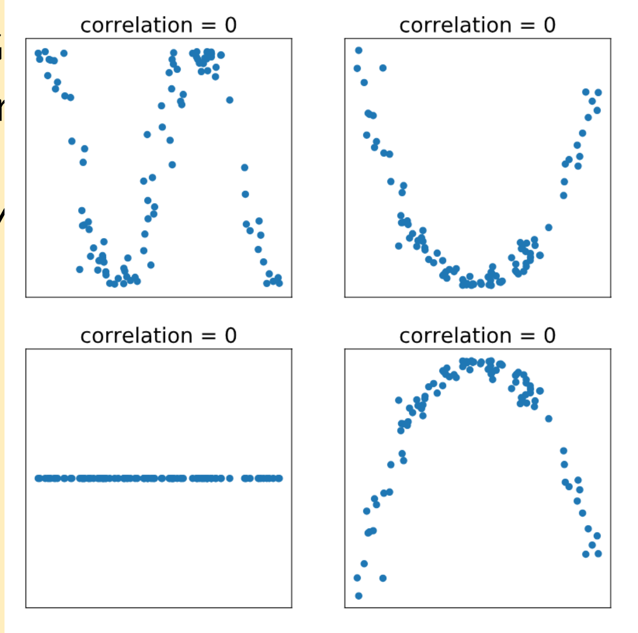
NOTE: Correlation ***doesn't*** tell us ***why*** variables are related (correlation $\neq$ causation)

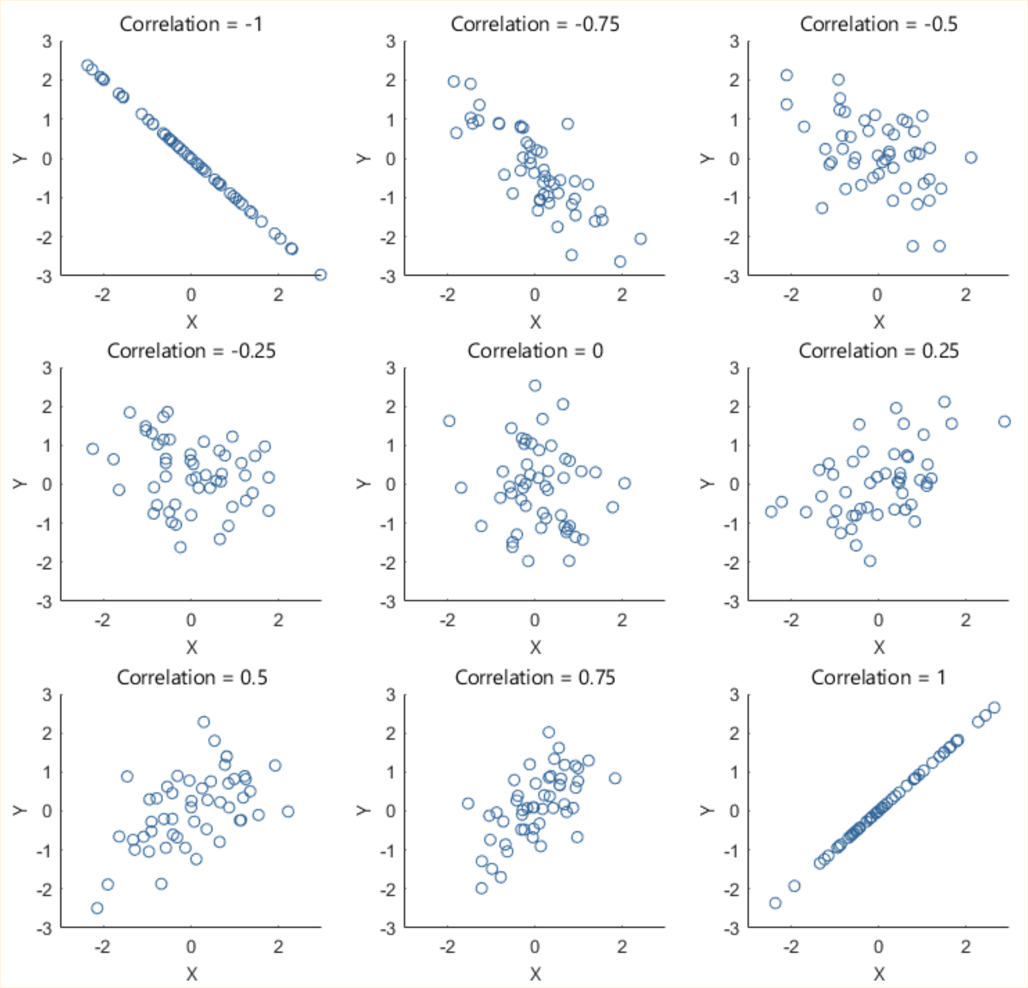
Visualizing Correlation

G ,

gr

M





Two Important Terms

We're often specifically interested in how one variable impacts another

- For example, how does education affect income?
- This question implies "directionality" – education explains variation in income

In these situations, we'll use the following terms:

- Our ***outcome*** variable is what we're trying to explain (here, income)
- Our ***explanatory*** variable drives variation in our outcome (here, education)

A Couple of Examples

Distinction between outcome and explanatory variables is important

- Imposes clarity about the goals of data analysis
- Let’s cover a couple more examples

How does a company’s marketing expenditures affect their sales?

- Outcome variable is sales
- Explanatory variable is the company’s spending on marketing

How does incarceration affect the employment of criminal offenders?

- Outcome variable is employment status (employed vs. unemployed)
- Explanatory variable is incarceration (either yes / no or length of sentence)

Quick Note on Terminology

You might've used the following terms in other classes:

- ***Dependent variable*** = Y = outcome variable from prior slides
- ***Independent variable*** = X = explanatory variable from prior slides
- There's nothing "wrong" with these terms... but they're not very clarifying

Two benefits of the outcome vs. explanatory variable distinction:

1. Directly connects with regression equations
2. Gives you clues about which is which (less likely to mix them up!)

A bit more context for point (2) above while you're studying :

- Equations like $Y = mX + b$ or $Y = \beta_0 + \beta_1 X$ are ways of saying $Y = f(X)$ = "Y is a function of X"
- In other words, $Y = f(X)$ is saying, let's use X to ***explain*** Y, meaning that X is our ***explanatory*** variable

Key Concepts for Quiz Next Week

Everything in slides is fair game, but the following concepts are important:

1. Defining different types of variables (continuous, factor, binary, etc.)
2. Showing the distribution of discrete and continuous variables
3. Distinguishing between outcome vs. explanatory variables

For covariance and correlation, ***don't*** need to memorize formulas, but you should be able to identify positively and negatively correlated variables (given either a correlation value or a scatter plot).

Why Learn R?

Some students worry a lot about the “perfect” language to learn

- As econ majors, you’ll generally be applying for “generalist” roles
- If you want a programming-specific role, this might matter more

General skills developed with one language are highly portable to another:

- Thinking rigorously about inputs and desired outputs
- Asking clearly defined questions and using documentation

Why not Excel?

- Overly-forgiving for sloppy inputs & very labor-intensive to achieve real proficiency
- Learning R will make you a much more careful user of Excel, Python, etc.

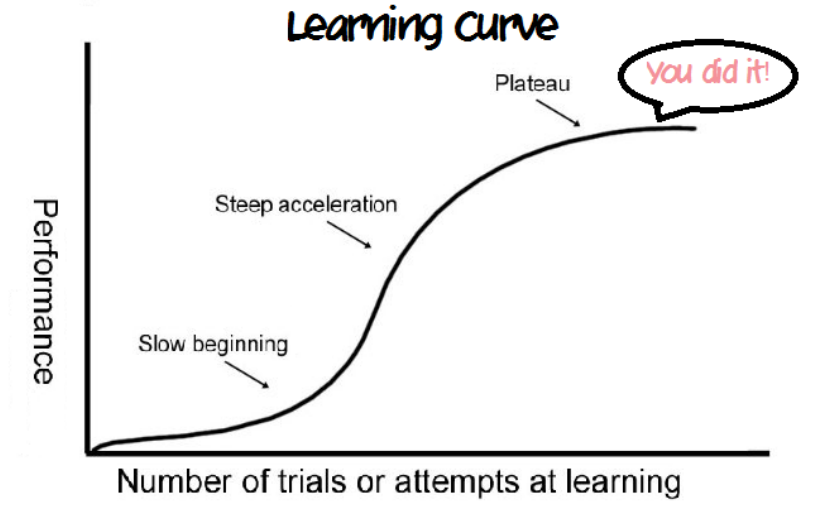
Learning Curves

Getting proficient with any skill takes time

With R, slow beginning is 10-20 hours of frustration – “nothing works!”

- Packages won’t load
- Error messages will be mystifying

The faster you can get through that slow beginning, the easier things will be later



Getting Started with R

Here, we'll review the "Learning to Speak R" slides

- Handout and slides are available on Canvas Week 2 Overview page
- Use these to help complete first coding activity!

Completing the First Swirl Activity

Install R + RStudio (if you haven't already) and open RStudio

Start by installing Swirl and loading the course material into R:

1. Use `install.packages("swirl")` to install Swirl package
2. Run `library("swirl")`
3. Run `install_course("R Programming")`

To access the first activity:

1. Run `swirl()` in the command line and follow the prompts
2. For tonight, we want to complete Lesson 1: Basic Building Blocks

General Tips for Swirl Activities

Resist the temptation to speed through activities!

1. Before you run a line of code, ask yourself, “what do I expect to happen?”
2. After you run your code, check for any surprises
3. Always identify ***where*** output is going (e.g., environment pane, console, etc.)

Want to close out of a Swirl activity? Type `bye()` in command line