Week 1: Understanding Data

Javier Sajuria

Queen Mary University of London

POL252 Political Research

Course Details

Introduction to Statistics

Introduction to Measurement

Descriptive Statistics

Wrap-up

Computer lab today

Lecture overview

Course Details

Introduction to Statistics

Introduction to Measurement

Descriptive Statistics

Wrap-up

Computer lab today

What is this course?

- This is not a course on statistics
  - A statistics course would focus on the theory and derivation of statistical methods
  - We will discuss some theory at a basic level, but will not concern ourselves with the derivation

What is this course?

- This is not a course on statistics
  - A statistics course would focus on the theory and derivation of statistical methods
  - We will discuss some theory at a basic level, but will not concern ourselves with the derivation
- This is a course on applied quantitative research methods
  - Our focus will be on the intuition and application of quantitative methods
  - We will concentrate on how to understand and use these methods to answer political science questions

Why should you take research methods?

- This course (and others on methods) will. . .

Why should you take research methods?

- This course (and others on methods) will. . .
  - . . . provide you with the tools necessary to conduct social scientific research (relevant for writing your dissertations)

Why should you take research methods?

- This course (and others on methods) will. . .
  - . . . provide you with the tools necessary to conduct social scientific research (relevant for writing your dissertations)
  - . . . help you to better understand and evaluate quantitative claims (relevant for evaluating plausibility of current research)

Why should you take research methods?

- This course (and others on methods) will. . .
  - . . . provide you with the tools necessary to conduct social scientific research (relevant for writing your dissertations)
  - . . . help you to better understand and evaluate quantitative claims (relevant for evaluating plausibility of current research)
  - . . . help you to think more critically about evidence-based arguments made in the 'real world' (relevant for being a good human being)

Why should you take quantitative research methods?

- You will learn to apply a wide range of quantitative data to answering your potential research questions

Why should you take quantitative research methods?

- You will learn to apply a wide range of quantitative data to answering your potential research questions
- You will learn the types of questions that can (and cannot) be answered using quantitative analysis

Why should you take quantitative research methods?

- You will learn to apply a wide range of quantitative data to answering your potential research questions
- You will learn the types of questions that can (and cannot) be answered using quantitative analysis
- You will learn to make more persuasive arguments using quantitative data

Why should you take quantitative research methods?

- You will learn to apply a wide range of quantitative data to answering your potential research questions
- You will learn the types of questions that can (and cannot) be answered using quantitative analysis
- You will learn to make more persuasive arguments using quantitative data
- You will learn to interpret and evaluate the quantitative evidence others present in their work

Why should you take quantitative research methods?

- You will learn to apply a wide range of quantitative data to answering your potential research questions
- You will learn the types of questions that can (and cannot) be answered using quantitative analysis
- You will learn to make more persuasive arguments using quantitative data
- You will learn to interpret and evaluate the quantitative evidence others present in their work
- You will learn some transferable skills

Why should you take quantitative research methods?

- You will learn to apply a wide range of quantitative data to answering your potential research questions
- You will learn the types of questions that can (and cannot) be answered using quantitative analysis
- You will learn to make more persuasive arguments using quantitative data
- You will learn to interpret and evaluate the quantitative evidence others present in their work
- You will learn some transferable skills

> *I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?*
>
> *Hal Varian, Chief Economist of Google*

Course Tutors

Javier Sajuria

- E-mail: j.sajuria@qmul.ac.uk
- Office: 2.29, ArtsOne Building
- Office Hours: Mondays 13-14 and Wednesdays 11.00-12.00

Philipp Broniecki

- E-mail: p.broniecki@qmul.ac.uk
- Office: TBA
- Office Hours: TBA

Lecture & Seminar

Lectures

- Lectures are required and run on Mondays from 15-17 at People's Palace PP1
- They will take roughly 2 hours, sometimes less
- We will cover most of the theoretical issues in the lectures and leave the practical exercises for the seminars

Seminars

- Seminars are mandatory and will be run by Dr Broniecki. They will take 50 minutes.
- They will take place on Tuesdays, check your timetable to find out the right time and place
- The tutor will be both present to help you through the activities in R. It is **encouraged** to bring your own laptops
- **During the first weeks, there will be a mismatch between lectures and seminars**
- At the end of each seminar there are exercises. Solutions are posted on the following Monday

QMPlus

- QMPlus access is essential for this course
- Students should be automatically enrolled, but please let me know if you have problems accessing it
  - Lecture slides will be available on QMPlus only minutes before each lecture
  - All substantive questions should be asked via the forum at QMPlus. The tutors will respond as soon as possible.
  - All administrative questions should be asked either before/after lecture or in office hours
  - When possible, avoid e-mails. There is a forum on QMPlus for substantive questions.

GitHub

- There is a special website designed to contain all relevant information for your tutorials:

  `https://qmul-spir.github.io/POL252/`

- The instructions and solutions will be posted there, not in QMPlus.

Assessment

- 40% of the course mark is based on midterm assessment
- 60% of the course mark is based on a final take-home research project
- The two assignments will require you to:
  1. understand the theoretical concepts
  2. answer applied questions
  3. work with R
- Details will follow during the term

Readings

- Required readings:
  - Diez et al. 2013. OpenIntro Statistics, 2nd Edition. http://www.openintro.org/stat/textbook.php
  - Lane et al. Online Edition. Introduction to Statistics. http://onlinestatbook.com/Online_Statistics_Education.pdf
- You are expected to complete the required readings prior to attending lecture
- Further readings are optional

Advice on the Readings

- Statistical readings can be intimidating
- Here is some advice:
  1. Do the required readings before class
  2. Do not expect to understand everything on the first pass
  3. If overwhelmed, focus on the text, not the equations
  4. After lecture, re-read to maximise understanding

R

- R is statistical software
- Why R?
  - More powerful than some alternatives – e.g. Excel, SPSS
  - Easier to learn than others – e.g. Matlab, SAS
  - Many scholars and practitioners use R.
  - R is free free free!
  - Public and private sector organisations increasingly switch to free (open source) statistical software like R and Python.
- Learning to use R is essential to do well in this course
- Don't worry if you have trouble the first few weeks!

Lecture overview

Course Details

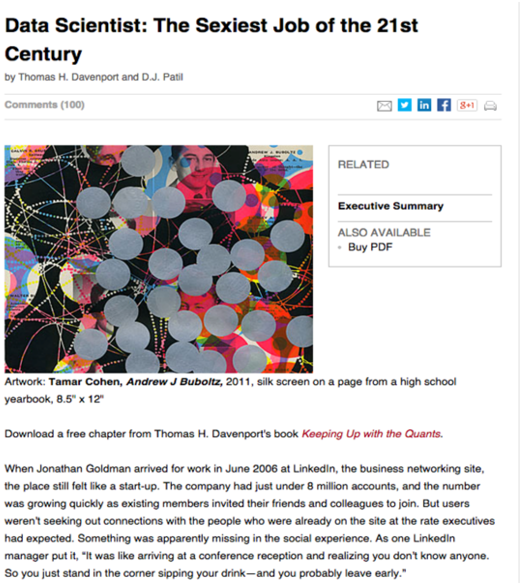
Introduction to Statistics

Introduction to Measurement

Descriptive Statistics

Wrap-up

Computer lab today



Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Comments (100)

RELATED

Executive Summary

ALSO AVAILABLE

- Buy PDF

Artwork: **Tamar Cohen, *Andrew J Buboltz,*** 2011, silk screen on a page from a high school yearbook, 8.5" x 12"

Download a free chapter from Thomas H. Davenport's book *Keeping Up with the Quants*.

When Jonathan Goldman arrived for work in June 2006 at LinkedIn, the business networking site, the place still felt like a start-up. The company had just under 8 million accounts, and the number was growing quickly as existing members invited their friends and colleagues to join. But users weren't seeking out connections with the people who were already on the site at the rate executives had expected. Something was apparently missing in the social experience. As one LinkedIn manager put it, "It was like arriving at a conference reception and realizing you don't know anyone. So you just stand in the corner sipping your drink—and you probably leave early."

Why?

Big Data

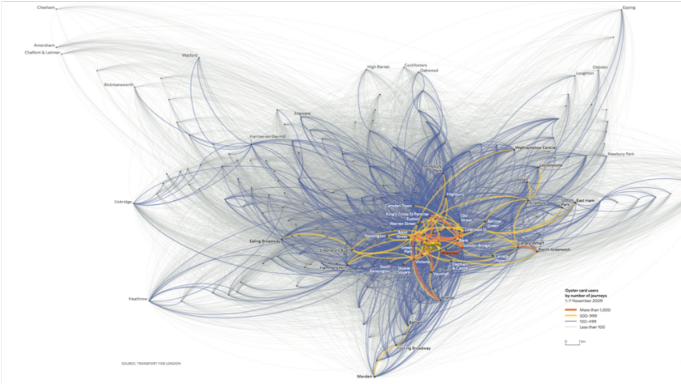
- Wikipedia says
  - “Big data represents the information assets characterised by such a high volume, velocity and variety to require specific technology and analytical methods for its transformation into value”
- In social science the term is often used to mean “bigger than an Excel spreadsheet”
- It has become a buzzword, but one that has awakened people to the need for people to clean, manipulate and analyse data

Some examples:

- UK Census (hundreds of variables across 255,000 small areas)
- Credit card records
- Healthcare records (NHS)
- Social Media data
- Text sources

Big Data and “Big” Analysis

- Pairwise comparisons
- Machine Learning
- Network Analysis



SOURCE: TRANSPORT FOR LONDON

To think big, start small

- This module is about high quality data analysis
    - “Small data” can be just as powerful
    - “Big data” contains a lot of noise
- Many of the methods are *scalable*. The is, that they can be applied to datasets small and large
- All analyses should start with the same basic checks and descriptions of the data

Research Questions

- A brief **question** that clearly identifies the problem or puzzle one seeks to answer
- All scientific research starts with a research question!
- How will you know if you have identified a good research question?
    1. Do you find the question interesting?
    2. Do your advisor and your friends find the question interesting?
    3. Can it be answered through empirical research?
    4. Is discovering the answer feasible?
    5. Is the literature unable to answer the question?
    6. Does the question have broad applicability or implications?

An Example from Comparative Politics

- After World War II, there was a general consensus that Western European countries were generally more democratic than countries in other parts of the world.
- Western European countries also tended to be more economically developed than countries in other parts of the world
- What research question might you arrive at after making these two observations?

Theories

- “[A] tentative conjecture about the causes of some phenomenon of interest” (Kellstedt and Whitten 2013, p.3)
- “[T]he aim of theory should be to construct a collection of models that is sufficiently small to be remembered and used, and covers a sufficiently large portion of the spectrum of facts.” (Dixit, 2004)
- Theories explain how or why something happens
- Some Properties of “Good” Theories
  1. Answer interesting research questions
  2. Generalizable – i.e. applicable in a broad range of contexts
  3. Parsimonious – i.e. simple
  4. Causal

Hypotheses

- “[A] theory-based statement about a relationship that we expect to observe” (Kellstedt and Whitten 2013, p.3)
- Properties of Hypotheses:
  - An educated guess about the cause(s) of something
  - Derived from one’s theory
  - Must be falsifiable

Modernization Theory and Its Hypotheses

- Early modernization theory – economically developed countries are more likely to be democratic because they have a large middle-class that moderates political conflicts (Lipset 1959, p. 83; Moore 1966)
- What hypotheses can you derive from modernization theory?

Modernization Theory and Its Hypotheses

- Early modernization theory – economically developed countries are more likely to be democratic because they have a large middle-class that moderates political conflicts (Lipset 1959, p. 83; Moore 1966)
- What hypotheses can you derive from modernization theory?
  - Economically developed countries are more likely to be democratic
  - Economically developed countries are likely to have a larger middle-class
  - Countries with a large middle-class are more likely to be democratic
  - Moderate political parties are more likely to hold positions of political power in countries with a large middle-class

Modernization Theory and Its Hypotheses

- Early modernization theory – economically developed countries are more likely to be democratic because they have a large middle-class that moderates political conflicts (Lipset 1959, p. 83; Moore 1966)
- What hypotheses can you derive from modernization theory?
  - Economically developed countries are more likely to be democratic
  - Economically developed countries are likely to have a larger middle-class
  - Countries with a large middle-class are more likely to be democratic
  - Moderate political parties are more likely to hold positions of political power in countries with a large middle-class
- "To make sure a theory is falsifiable, choose one that is capable of generating as many observable implications as possible." (King, Keohane and Verba, 1994, p. 19)

Null Hypotheses

- For every hypothesis there is a corresponding null hypothesis (H_0)
- Null hypotheses state what we would observe if a hypothesis is false
- Statistical tests always evaluate the likelihood that some null hypothesis is false
- Based on that likelihood, we either reject or fail to reject the null hypothesis

Data Collection

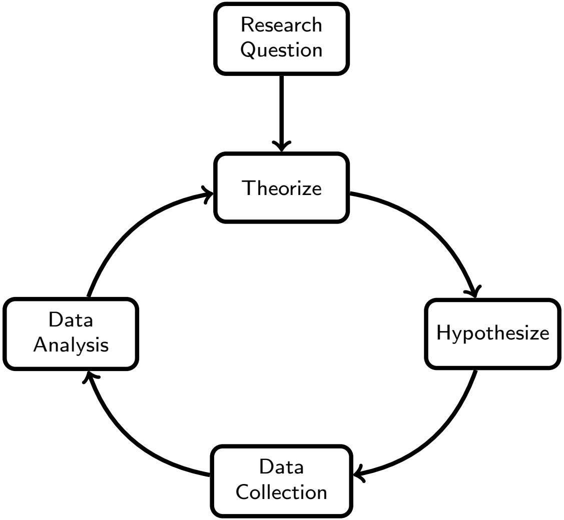
- Process of systematically gathering and measuring information on variables of interest
- Observations or subjects
  - Unit of analysis – the “what” or “who” being studied
- Variables – anything that we can measure about the subjects in our sample; should vary across our subjects
  - Dependent variable (Y) – variable to be explained or the outcome variable
  - Independent variable(s) (X) – determinant(s) of the outcome variable; also called the explanatory or predictor variable(s)
- Hypotheses should be able to be restated in terms of variables:

$$X \to Y$$

Data Analysis

- Evaluate the null hypothesis
- Does the data you collected suggest that you should reject or fail to reject the null hypothesis?
- You will learn some statistical tests that you can use to answer this question

The Scientific Method



How can we evaluate scientific claims?

- Internal validity – are the causal claims made by a researcher warranted?
- External validity – can the causal claims be generalized to the population one is studying?

How can we evaluate scientific claims?

- Internal validity – are the causal claims made by a researcher warranted?
- External validity – can the causal claims be generalized to the population one is studying?

- There is often a trade-off between internal and external validity:
  - Experimental research tends to have high internal validity but low external validity
  - Observational research tends to have high external validity but low internal validity

What is statistics?

- Study of the collection, analysis, interpretation and presentation of data
- Statistics help us solve a problem: we usually are not able to study the population of interest directly
  - Population – entire set of items or subjects one wishes to study (e.g. countries or UK citizens)
  - Sample – subset of a population chosen for study
    - Random samples are best

What is statistics?

- Study of the collection, analysis, interpretation and presentation of data
- Statistics help us solve a problem: we usually are not able to study the population of interest directly
    - Population – entire set of items or subjects one wishes to study (e.g. countries or UK citizens)
    - Sample – subset of a population chosen for study
        - Random samples are best
- We use statistics to make predictions (or inferences) about a population based on data from a sample of that population
    - Parameter – numerical summary of some quantity of interest from a population
    - Statistic – numerical summary of some quantity of interest from a sample

Types of statistical analysis

- We will primarily focus on statistical analysis
- The statistics used for analysis can be broken into two types:
  - Descriptive statistics – summarize data
  - Inferential statistics – make predictions about wider population

Lecture overview

Course Details

Introduction to Statistics

Introduction to Measurement

Descriptive Statistics

Wrap-up

Computer lab today

Introduction to measurement

- What is measurement?
  - Assignment of numbers to objects or events
  - Measurement is essential for quantitative research
- Three steps to measuring social science concepts:
  1. Selection of indicators
  2. Selection of measurement level
     - Nominal
     - Ordinal
     - Interval
  3. Aggregation of indicators

Levels of measurement

- Categorical/Nominal
  - Lowest level of measurement
  - Values indicate different, mutually exclusive categories
  - Examples: gender, race, party identification

Levels of measurement

- Categorical/Nominal
  - Lowest level of measurement
  - Values indicate different, mutually exclusive categories
  - Examples: gender, race, party identification
- Ordinal
  - Values indicate relative differences between categories
  - Imply a ranking but precise difference between categories is not uniform and often unknown
  - Examples: Likert scales, education

Levels of measurement

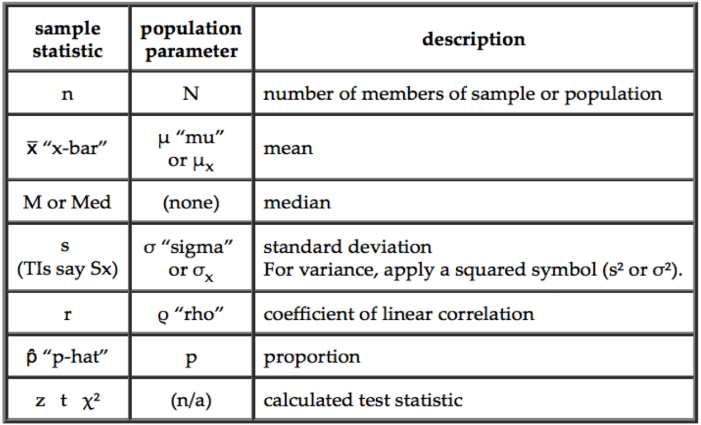
- Categorical/Nominal
  - Lowest level of measurement
  - Values indicate different, mutually exclusive categories
  - Examples: gender, race, party identification
- Ordinal
  - Values indicate relative differences between categories
  - Imply a ranking but precise difference between categories is not uniform and often unknown
  - Examples: Likert scales, education
- Continuous/Interval
  - Highest level of measurement
  - Values indicate precise differences between categories (i.e. equal-unit differences)
  - Examples: age, income, casualities

Statistical Notation

Statistical notation is like learning a language - certain symbols have certain definitions

- $\pm$ means to add and also subtract
- $\surd$ is the square root symbol. What number multiplied by itself results in the number under the radical
- $\leqslant$ less tan or equal to
- $\geqslant$ greater than or equal to
- $\neq$ is not equal to
- Σ the Greek letter Sigma, meaning "to sum" what comes after it
- Π the Greek letter (capital) Pi, meaning "to multiply" what comes after it

Statistical Notation



sample statistic	population parameter	description
n	N	number of members of sample or population
$\bar{x}$ "x-bar"	μ "mu" or μ_x	mean
M or Med	(none)	median
s (TIs say Sx)	σ "sigma" or σ_x	standard deviation For variance, apply a squared symbol (s^2 or σ^2).
r	ρ "rho"	coefficient of linear correlation
$\hat{p}$ "p-hat"	p	proportion
z t χ^2	(n/a)	calculated test statistic

Statistical Notation

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + \cdots + X_n$$

The "$i = 1$" at the base of Σ means "start at your first x-value". This would be X_1. The "n" at the top of Σ means "end at n". In statistics, n is the number of items in the data set. So what this summation is asking you to do is "add up all of your x-values from the first to the last."

Note: If you see a number above Σ, instead of n, it means to add up to a certain point. For example, a 3 above the Σ means to sum up the the third item (X_3) in the set.

What does this mean?

$$\prod_{i=1}^{n} X_i$$

What does this mean?

$$\prod_{i=1}^{n} X_i$$

- The multiplication of each element i until the element n

Error is inherent in measurement

$$x_i = z_i + e_i$$

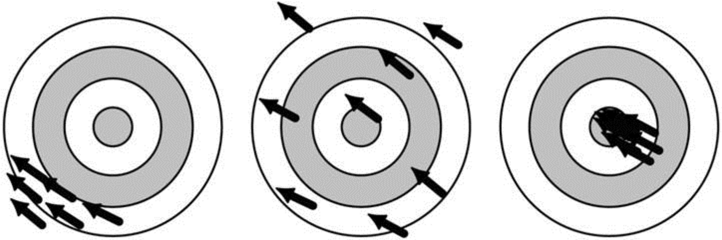
where

- z_i is the true value of some measure
- x_i is the observed value of some measure
- e_i is error

Evaluating measurement error

- Validity – am I measuring what I purport to measure?
- Reliability– is my measure consistent?

FIGURE 5.3 **Comparing the Reliability and Validity of Three Measures**



Reliable, but Not Valid

Valid, but Not Reliable

Valid and Reliable

Source: Clark, Golder, and Golder (2013)

Lecture overview

Course Details

Introduction to Statistics

Introduction to Measurement

Descriptive Statistics

Wrap-up

Computer lab today

Introduction to descriptive statistics

- Descriptive statistics are a good way to get a feel for the data
- The first step to data analysis
- Most interested in two qualities of the variables we are working with:
    - Central tendency
    - Dispersion

What are Descriptive Statistics?

- Univariate
    - Analysis of only one variable on some characteristic
        - Frequency Distributions – essentially a count or distribution of values on some single variable
        - Other descriptive statistics – some summary measure that describes the data in a way not obvious by looking at the frequency distribution
- Bivariate
    - Analysis of two variables – can be simple scatter plots or cross-tabulations
- Multivariate
    - Analysis of more than three variables

Descriptive Statistics

- Are not used to make inferences to the population
- Descriptive statistics (including frequency distributions) are a good way to get a first glimpse at the data
- Not very powerful in terms of analysis
- Are a good place to start when you come across a new set of data
- Best used as a complements to other forms of analysis
  - Can be used for some qualitative work like case studies or to look at a larger pattern when investigating a micro-level phenomenon
  - Can be used with other forms of quantitative analysis to provide introductory remarks or support the forms of analysis being used

Lecture overview

Course Details

Introduction to Statistics

Introduction to Measurement

Descriptive Statistics

Wrap-up

Computer lab today

Advice

- Keep a logbook or similar to keep track of ideas and concepts you are unsure of
  - Could include a glossary of terms, software functions, practice code, flow diagrams, etc.
- Be patient. Things will crash or you will get lost and need to start again
- Relax. We have time.

Summary

- We have covered a lot!
- Quantitative methods are important
- Go back over terminology / notation to acquaint yourselves with it
- Never has there been more data or tools to analyse data

Lecture overview

Course Details

Introduction to Statistics

Introduction to Measurement

Descriptive Statistics

Wrap-up

Computer lab today

Today's lab session

In the lab sessions today, you'll learn how to load data into R, look at measures of central tendency, and how to graphically display data.