POLS2044
WEEK 6 2024
Descriptive inference and descriptive statistics

Australian National University
School of Politics & International Relations
Dr. Richard Frank

In Week 6 of POLS2044 we will be focusing on descriptive inference and descriptive statistics. Descriptive inference is what we do when we try and answer "what" questions. What happened in the 2022 election campaign period? What do we mean when we talk about "democracy"? This discussion builds on the concepts and measurement discussion from Week 4. These sorts of questions are related to the "why" questions we ask when we focus on causal descriptions, but as Gerring (2012) stresses they can often involve different challenges and opportunities.

Reading

Even though you are likely to have multiple assignments due in Week 6, I would encourage you to spend the time (strategically and purposively) reading through this week's reading. Why? Chapter 6 of the textbook should give you ideas about how description is distinct from but interrelated with causal inference. This should prove helpful when thinking how your qualitative evidence connects to your causal theory.

LECTURE PART 1: Introduction

Recapping the last few weeks

Week 1: Scientific method
Week 2: Causal theorising
Week 3: Research design
Week 4: Concepts and measurement
Week 5: Surveys and sampling

Learning outcomes

Overview of the course's learning outcomes

Where we are headed

$$Y = \alpha + \beta X + \epsilon + \varepsilon$$

There is often an easier map to quantitative analysis than qualitative analysis.

The basic regression equation

$$Y = \alpha + \beta X + \epsilon + \varepsilon$$

Where:

Y is the outcome you are trying to explain.
X is the main explanatory variable.
(alpha) is the intercept.
(beta) is the estimated relationship between X and Y.
(Epsilon1) is the systematic error.
Epsilon2) is the random error.

We will be coming back to this equation in a few weeks, but first we need to start by learning about our main cause (X) and outcome (Y) variables.

Today's motivating questions

What can descriptive inference tell us that causal inference cannot?
What are the basic descriptive statistics?

Motivating puzzle

Political scientists spend much more time thinking about causal inference and data analysis than they think about conceptualising and describing their causes (X's) and outcomes (Y's).

However, the former is of limited utility without the latter.

LECTURE PART 2: Descriptive inference

Continuing from our measurement week

Most people use real-world data without thinking about how they are generated and whether they capture what they think they do.

Moving from theory to test

Adcock & Collier (2001: 531) conceptualization and measurement levels and tasks

Defining descriptive arguments

"A descriptive argument describes some aspect of the world.

In doing so it aims to answer what questions (e.g. when, whom, out of what, in what manner) about a phenomenon or a set of phenomena."
(Gerring 2012: 722, emphasis added)

Independent variable (a concept) ----------Causal theory-------- > Outcome (also a concept)
| |
| |

Description and causal inference

As Gerring (2012) makes clear, most current political science research focuses on causal inference rather than description.

However, description and causality are intimately related and can often overlap.

First, we need to understand the *what* before we can ask *why.*

That is not to say it is not important or influential.

Example of Cullen Hendrix's most cited article on measuring state capacity in JPR

Comparing causal focus across fields

Graph from Gerring (2012: 731)

Are political science topics just different?

The challenges of description

Concepts—Economic output, population, democracy
Measurement—GDP, Polity, V-Dem
Why is falsifying descriptive arguments so hard?
Describing a concept: What is democracy and how should we measure it?
Causal argument: Does democracy increase the chance of victory in war?

Why is description so hard?

"A description of even the smallest slice of reality can never be exhaustive."
(Max Weber 1905, quoted in Gerring 2012: 738)

"Any phenomenon of significance to social science is likely to call up multiple words, and multiple definitions of those words." (Gerring 2012: 738).

"To describe something is to assert its ultimate value," (Gerring 2012: 740).

Therefore, descriptions include an inherent subjectivity.

Map of the highest mountain on earth with three names

Problematising memory

"Remembrance of things past is not necessarily the remembrance of things as they were."
Marcel Proust. 1922. *In Search of Lost Time: Swann's Way.*

One way of addressing descriptive uncertainty is through robustness checks

We will come back to these techniques in a few weeks.

Most of what we care about are latent concepts

How do we measure latent, unobservable, unmeasurable constructs?
Democracy; Corruption; Conflict; Development; Skill

LECTURE PART 3: Descriptive statistics

Let us get to know our data.

Now that we have some ways of describing our topic, let us look at a few ways that we can measure it.

Remember that we should keep in mind **how the data were generated** so as to not try and take away more than we should from the data.

Measurement metrics

Label: Employment status of survey respondent

Values: "employed" or "unemployed"

Variable type:
(1) categorical/nominal [unemployed, employed]
(2) ordinal [<5 hours, 5-15 hours, 15-35, >35 hours worked per week]
(3) continuous/interval/ratio [time worked last week]

Categorical variables

We can put cases into categories based on their values, but we cannot **rank** or order them.

A categorical (and count/continuous) variable example

Magpie swooping attack percentages across Australia by state in 2022

Ordinal variables

Variables for which cases have values where we can make universal ranking distinctions.

If we treat an ordinal variable like a categorical variable, we are acting as if we have less information than we really do.

Two examples include rating Domino's ordering experience and age groups

Continuous variables

Sometimes called interval variables or ratio variables (if they have a meaningful 0). They have equal unit differences.

Example of elevation and sea level

Describing categorical variables

Usually, we focus on the frequency distribution of categorical variables with a table, pie charts, or bar graphs.

The only central tendency statistic is the mode (the most frequent value).

Quantiles (including percentiles) are also used. They are a measure of position within a distribution.

Example of quantiles/percentiles is the (originally named) Standard Aptitude Test (SAT), created in 1926.

Describing continuous variables

We are primarily interested in the central tendency and the distribution of values around this central tendency.

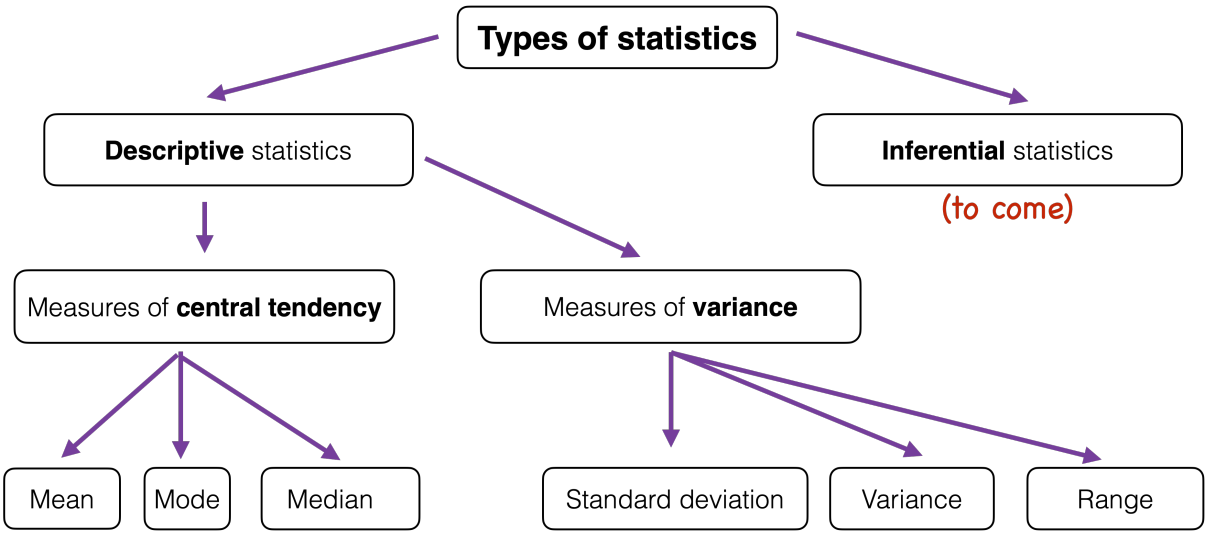
We are also interested in outliers.

The midpoint value is the median.

The average value is the mean.

The dispersion around the mean is described by the standard deviation.

My way of thinking about types of statistics



Finding the mean

Mean= sum of observations / number of observations

Desirable properties are that the derivations from the mean are zero-sum and the least squares property (to be discussed later)

Standard deviation

This is basically a way of telling the reader how the data are scattered around the average value.

A sample's standard deviation (*sd*) is given by the square root of the variance (the average distance away from the average value over the number of observations minus one (the degree of freedom, more on that later).

Or more concretely:

$$sd = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

Where:
x bar is your variable's mean.
x_i is an individual value.
n is the sample size.

The normal distribution

With only the mean and standard deviation we can tell a lot about our observations if they approximate the normal distribution, which is at the heart of probability theory.

A descriptive distribution example (magpie swooping)

The normal (Gaussian) distribution example

Height values together form something close to a normal distribution

Outliers happen.

Example of a person struck by lightning seven times.

An example of descriptive statistics and graphing in Excel is on Wattle

Using a graduate outcomes survey from 2022

https://www.qilt.edu.au/surveys/graduate-outcomes-survey-(gos)

An additional example using F1 data if I have time.

Motivating puzzle

Political scientists spend much more time thinking about causal inference and data analysis than they think about conceptualising and describing their causes (X's) and outcomes (Y's).

However, the former is of limited utility without the latter.

Important Week 6 terms

Variable label
Variable values
Variable types
    Categorical/nominal
    Ordinal
    Continuous/interval/ratio
Equal unit difference
Central tendency
Mode
Quantiles
Outliers
Mean
Median
Variance
Standard deviation

III. WEEK 6 WORKSHOP

In this week's workshop, we will be discussing the strengths and weaknesses of descriptive inference when applied to a particular concept, giving our Microsoft Excel analytical superpowers, and taking our first steps in generating descriptive statistics.

PART 1: Install Excel's Data Analysis ToolPak (individually)

We will be using Microsoft Excel repeatedly over the remainder of the semester. As I mentioned in the first week, almost all modern workplaces have computers with access to Microsoft Excel including through Microsoft 365. Therefore, instead of having you use a more advanced (and often quite cool) but unnecessary software like RStudio or Stata, we will be doing our data analysis in Excel. All ANU students can download Microsoft 365 including Excel if you have not downloaded it yet.[1] You can then use Excel online or on your desktop. If you are all in on Google Sheets, there is a comparable analytics add-on for Google Sheets called XLMiner Analysis ToolPak.

Below are instructions for installing add-ons to your Excel or Google Sheets that we will be using extensively. If you have not done so yet please install your analysis tool pack now as a group so we can all be on the same page moving forward.

Excel instructions

Load and activate the Analysis ToolPak.[2]

1. Click the **File** tab, click **Options**, and then click the **Add-Ins** category.
2. In the **Manage** box, select **Excel Add-ins** and then click **Go**.
   If you're using Excel for Mac, in the file menu go to **Tools** > **Excel Add-ins.**
3. In the **Add-Ins** box, check the "**Analysis ToolPak"** check box, and then click **OK**.
   a. If **Analysis ToolPak** is not listed in the **Add-Ins available** box, click **Browse** to locate it. Do not install any other analysis tool pack that may be similarly names.
   b. If you are prompted that the Analysis ToolPak is not currently installed on your computer, click **Yes** to install it.

   If needed, here is a video introduction to installing Analysis ToolPak on a PC. https://youtu.be/V60-IFnih3Q .

Google Sheets instructions

1. Open a blank spreadsheet.
2. Click the **Extensions** tab, click **Add-ons**, click **Get Add-ons**.
3. In the Search apps box, type **Analysis ToolPak**.
4. Click on the relevant result (downloaded almost 3 million times).

[1] Download information here: https://services.anu.edu.au/information-technology/software-systems/microsoft-office-365.
[2] Steps taken from at https://support.microsoft.com/en-us/office/load-the-analysis-toolpak-in-excel-6a63e598-cd6d-42e3-9317-6b40ba1a66b4

5. Click **Install**.

PART 2: Descriptive inference (as groups of 3-4 students)

1. Choose one of the following theoretical phenomena: "corruption", "democracy", or "urban population".

2. Describe the most important characteristics of your chosen outcome. When doing so, think about the following questions:

   a. What is the level of analysis of your descriptive argument?

   b. What actors or institutions are included in your description?

   c. What is the spatial (across space) and temporal (across time) domain considerations?

3. Can your group agree or disagree on the most important elements of your description?

4. Do you think that other groups would reach the exact description you have reached? Why or why not?

Paste your completed responses to the four questions above in Wattle/POLS3033/Week 6/Workshop/Item 6.1.

PART 3: Descriptive statistics (as groups of 3-4 students)

Using the methods I demonstrated in the video on Wattle and in lecture, find and report descriptive statistics (mean, median, mode, standard deviation, minimum, and maximum) for one variable in the three datasets listed below using the Analysis ToolPak in Excel or Google Sheets. It is up to your group which dataset you chose to analyse. All datasets are in a folder in Wattle/Week 6/Workshop. These are data excerpted from much larger and more complex datasets to make them a bit more manageable to work with. The focus here is on practicing basic descriptive analysis using real data.

If you need help with the process, Wattle will have a recorded version of my Week 6 lecture hands on with the Analysis ToolPak.

Table 1. Dataset choices

Dataset	Source	Description	Variable to analyse
cpi_2021	Transparency International	Annual corruption score	`CPI score 2021`
Polity_2018	Polity IV Project	Annual democracy score	`polity2`
nyc_squirrels	https://www.thesquirrelcensus.com/	A 2020 census	`squirrel_number`

Table 2. Summary statistics and interpretation

Variable name	Mean	Median	Mode	Standard deviation	Minimum	Maximum
Below describe any descriptive findings you find particularly interesting about the distribution of your variable. *Any substantive descriptive conclusions you can reach given these descriptive statistics?*						

Paste your completed version of Table 2 in Wattle/POLS3033/Week 6/Workshop/Item 6.2.