POLS2044 WEEK 7
Correlation and visualisation

Australian National University
School of Politics & International Relations
Dr. Richard Frank
https://richardwfrank.com/research_design_2022

In Week 7 of POLS2044 we will be focusing on correlation and visualisation. Both topics are fundamental parts of research design. Correlation and the significance testing of correlation coefficients are building blocks of more advanced methods we will cover in this class (and well beyond). Visualisation of data is also part and parcel of descriptive and inferential statistics. It is also a growing profession in its own right as well as part of telling stories in journalism and video.

Today I have two main goals. My first goal is to have you understand what correlation means (and what it doesn't) and how we can tell (and why we would want to) whether a correlation is statistically significant or not. My second goal is to have you understand a few basic types of visualisation and understand why you might want to have some figures and graphs in your final essay.

Reading notes and questions

I do not have very many reading notes and questions this week. This is, in part, due to the short and direct nature of all three readings. Please do the readings in the order in which I discuss them below.

Tacq, Jacques. 2004. "Correlation," in *The SAGE Encyclopedia of Social Science Research Methods*, edited by Michael S. Lewis-Beck, Alan Bryman, and Tim Futing Liao. Thousand Oaks, CA: Sage: 200-204.

This is a short and direct entry in a very large encyclopedia of social science research. It is the most direct source I can find describing the five elements of correlation that does not get bogged down in the math. When reading the article, try and internalise the five main features of a correlation. It is not necessary to spend much time on the standardised coefficients section or on the math. I think the lecture and tutorial activities will give you a more hands-on feeling for how to calculate a correlation coefficient and tell if it is statistically significant or not.

1. What are the five main features of a correlation?

The next two readings are focused on visualisation. They are a bit dated, but they are the shortest and most direct descriptions of good and bad practice when visualising data.

Wainer, Howard. 1992. "Understanding Graphs and Tables." *Educational Researcher* 21(1): 14-23.

This first article focuses on visualising data on one or more feature of our world can aid descriptive and causal understanding. It also highlights a few famous early practitioners of this art, some of whom I will mention in lecture.

2. What types of questions does Wainer (1992, building on Bertin 1973) think graphs can be used to answer.
3. Do you understand the implications for your own work Wainer's (1992: 18) point that "a table is for communication, not data storage."

Wainer, Howard. 1984. "How to Display Data Badly." ***The American Statistician*** **38(2):137–147.**

This article includes several examples discussed in more detail by Tufte (2001) of poor visualisations. Often, I have learned more by peoples' mistakes (including my own) than by people's successes. This reminds me of a famous Michael Jordan quote "I've missed more than 9000 shots in my career. I've lost almost 300 games. 26 times, I've been trusted to take the game winning shot and missed. I've failed over and over and over again in my life. And that is why I succeed."[1]

4. What are the three elements of good data graphics?
5. How do bad data graphics violate these three elements?
6. What are Wainer's (1984) twelve rules of bad data visualisation?

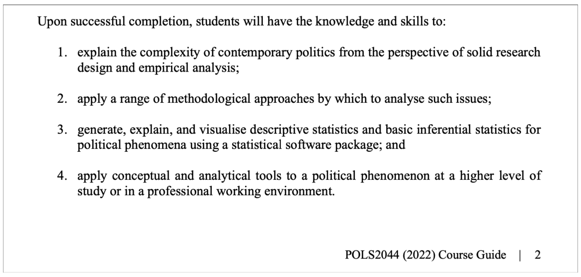
Keep a lookout for visualisations in future readings in this class as well as in the news media. What visualisations stand out for you? How can you use tables and figures in your final paper to help make your final paper more compelling?

LECTURE PART 1: Introduction

Recapping the first six weeks

Week 1: Scientific method
Week 2: Causal theorising
Week 3: Qualitative research approaches
Week 4: Concepts and measurement
Week 5: Surveys and sampling
Week 6: Descriptive statistics

Learning outcomes



Upon successful completion, students will have the knowledge and skills to:

1. explain the complexity of contemporary politics from the perspective of solid research design and empirical analysis;
2. apply a range of methodological approaches by which to analyse such issues;
3. generate, explain, and visualise descriptive statistics and basic inferential statistics for political phenomena using a statistical software package; and
4. apply conceptual and analytical tools to a political phenomenon at a higher level of study or in a professional working environment.

POLS2044 (2022) Course Guide | 2

[1] Quote from Aaditya Krishnamurthy. 2022. "Michael Jordan's Secret To Success: 'I've Missed More Than 9000 Shots In My Career. I've Lost Almost 300 Games... I've Failed Over And Over And Over Again In My Life. And That Is Why I Succeed.'" Fadeaway World. https://fadeawayworld.net/nba-media/michael-jordans-secret-to-success-ive-missed-more-than-9000-shots-in-my-career-ive-lost-almost-300-games-ive-failed-over-and-over-and-over-again-in-my-life-and-that-is-why-i-succeed Accessed 19/09/22.

POLS2044's three commandments (so far)

1. Know thy "what" and "why".
2. Know thy data.
3. Write it down.

Where we are headed: Ordinary least squares regression

Y=a + BX + e1 + e2

Where:

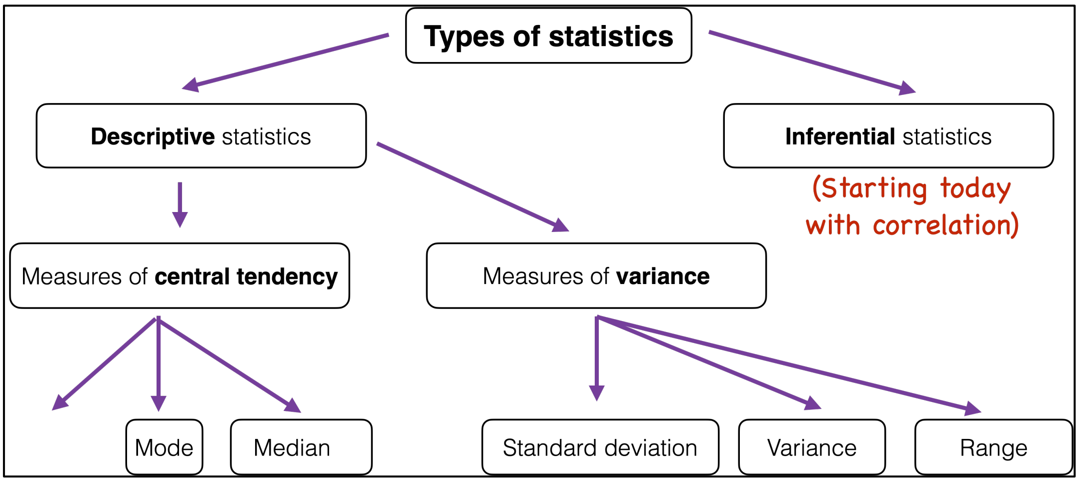
Y is the outcome you are trying to explain.
X is the main explanatory variable.
a(alpha) is the intercept.
B (beta) is the estimated relationship between X and Y.
e1 is the systematic error.
e2 is the random error.

Today's motivating questions

How can correlation measures help our descriptive and causal inference?
What is actually going on under the statistical hood?
How can data visualisations help with descriptive or causal inference?

Motivating puzzle

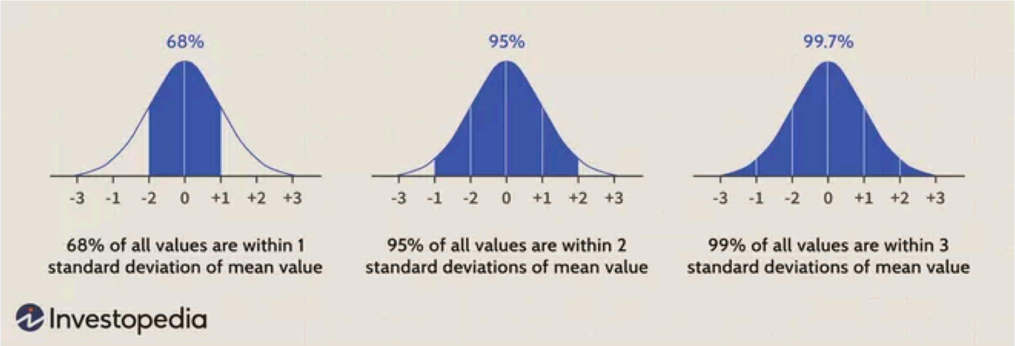
Most political scientists use research methods without thinking about (1) where they came from and (2) what their underlying assumptions and methods are doing to their analysis.
However, context and underlying assumptions can shape what conclusions we reach and the likelihood of these conclusions holding up.
Therefore, we want to try and avoid easy mistakes and learn from previous mistakes.


Types of statistics
Descriptive statistics
Inferential statistics
(Starting today with correlation)
Measures of **central tendency**
Measures of **variance**
Mode
Median
Standard deviation
Variance
Range

Finding the standard deviation

The normal distribution

With only the mean and standard deviation we can tell a lot about our observations if they approximate the normal distribution.



LECTURE PART 2: Correlation

What is a correlation?

"It is the statistical association between two variables of interval or ratio measurement level." (Tacq 2004: 2)

London's John Snow's pub

BBC article about Syria's cholera outbreak

Drawing of London's Soho, August 1854

Miasma theory

John Snow (1813-1858)

John Snow's map

Correlation does not imply causation

Example of ice cream and homicides.

Five main features of correlation (Tacq 2004)

1. Nature
2. Direction
3. Sign
4. Strength
5. Capacity for generalisation

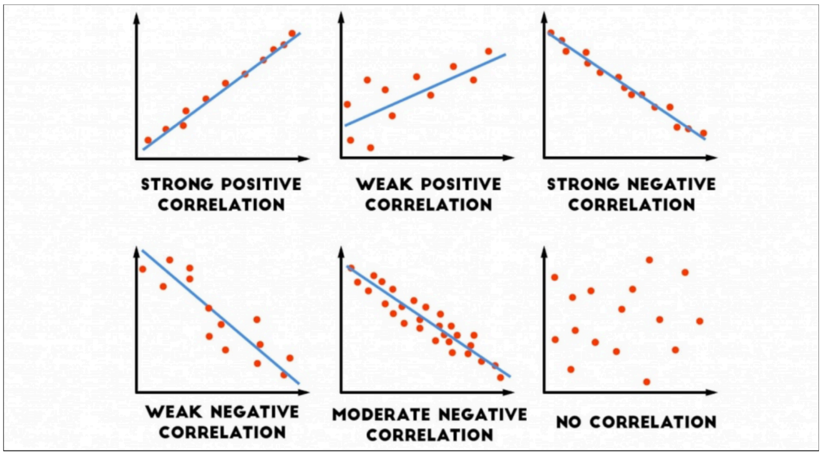
The nature of the correlation

The correlation between ***X*** and ***Y*** can be linear or non-linear (e.g., exponential, monotonic, logistic, quadratic, discontinuous).

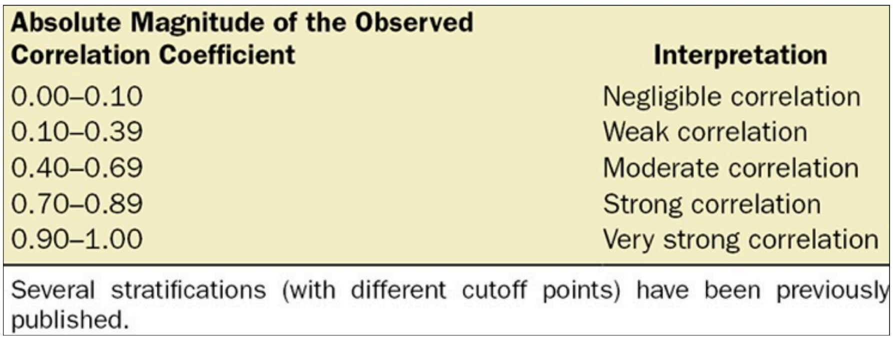
The direction of the correlation

1. ***y*** is dependent variable, ***x*** is the independent variable (x->y)
2. ***x*** is the dependent variable, ***y*** is the independent variable (y->x)
3. ***x*** and ***y*** are not plausibly causally related (e.g. Nick Cage & drowning)

The sign (-/0/+) of the correlation



The strength of the correlation



Absolute Magnitude of the Observed Correlation Coefficient	Interpretation
0.00–0.10	Negligible correlation
0.10–0.39	Weak correlation
0.40–0.69	Moderate correlation
0.70–0.89	Strong correlation
0.90–1.00	Very strong correlation

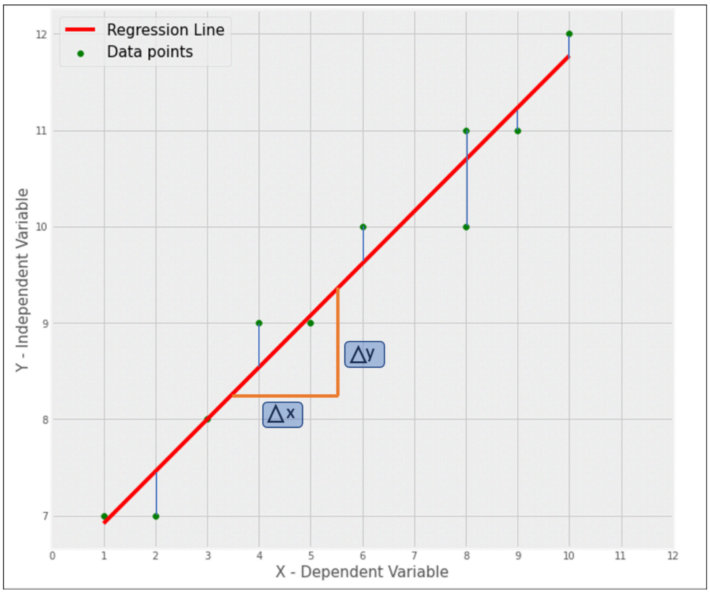
Several stratifications (with different cutoff points) have been previously published.

The generalisability of the correlation

Focus on **statistically significant correlations**.
Statistical significance can be affected by the size of the sample including its relation to population size.
We want to have some confidence that the relationship is due to some relationship in the population rather than random chance.

How is a correlation measured?

Similar to how we use linear regression we are interested in the speed of change (e.g., dy/dx)



Pearson’s correlation coefficient

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Where:
r is the coefficient of correlation between x and y
x is each individual value (i) of the independent variable
x hat is the average value of x
y is each individual value (i) of the dependent variable
y hat is the average value of y
n is the number of observations

In-lecture example of Eurovision 2022 finals

An enduring interest to IR/Comparativists

Why conduct a significance test?

We want to be sure that the correlation is not an artefact of random chance or what sample we have.

How do we actually conduct a significance test?

Rho is the population correlation coefficient.
Null hypothesis (H0): rho=0, there is not a significant linear correlation between x and y in the population.
Alternative hypothesis (H1): rho is not equal to 0, there is a significant linear correlation between x and y in the population.
Now we conduct a Student’s T-test. What is that?

Student’s t-test creator

William Sealy Gosset
(aka “Student”)
b.1876-d.1937

Student’s t-test

$$t_{score} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Where:

r is the Pearson’s correlation coefficient

n is the sample size

T-distribution

The area under the curve equals 100% probability of all values of an outcome being observed.

Know thy data (and what created them)

Examples of weight rack, marathon times, Polish student exams.

T-distribution and p-values

In-class example of conducting significance tests

Correlation: important takeaways

A correlation is the statistical association between two variables.
It has five important characteristics (nature, direction, sign, strength, statistical significance).
Calculating a correlation coefficient and its statistical significance is straightforward.
Interpreting what it means is a different thing and requires thinking causally.

LECTURE PART 3: Data visualisation

What is wrong with these three graphs

POLS2044’s three commandments

1. Know thy “what” and “why”.
2. Know thy data.
3. Write it down.

Know thy data

We can learn a lot about the world through descriptive statistics.
However, humans are often visually focused creatures.
Therefore, descriptive statistics work hand-in-hand with visual examination and visual description when trying to understand our world.

"Getting information from a table is like extracting sunbeams from a cucumber."
Farquhar, Arthur, and Farquhar, Henry. 1891. *Economic and Industrial Delusions*. New York: G. P. Putnam's Sons: 55 (emphasis added).

Wainer's (1984) rules for data graphics

Graphics done well	Graphics done poorly
Show data	Show as few data points as possible, and hide what you do include.
Show data accurately	Only order matters, graph data out of context.
Show data clearly	Change the scales mid-axis; emphasise the trivial and ignore the important; ignore natural baselines; ignore natural ordering; obscure labels.

Wainer's (1992) rules to creating tables

1. Order the rows and columns in a way that make sense.
-Often it helps to put largest first or order chronologically.
2. Round numbers.
-Most readers cannot understand more than two digits
-Measurement error also makes more than two digits seem like we know more than we do.
3. Summary rows or columns help with comparison.

Bertin's (1973) three levels of visual analysis

Elementary level questions
What votes did Eurovision participants receive?
Intermediate level questions
How did Eurovision voting patterns for Australian singers vary from 2015 to 2022?
Overall level questions
How did Eurovision voting patterns change over time (e.g.,since the end of the Cold War)?

Tufte's principles of graphical excellence

"Graphical excellence is the well-designed presentation of interesting data—a matter of substance, of statistics, and of design."
Edward Tufte. 2001.*The Visual Display of Quantitative Information*. 2nd ed. Cheshire CT: Graphics Press: 51 (emphasis added).

William Playfair (1759-1823)

A Scottish engineer, political economist, and secret agent against the French.

He developed the line, bar, area, and pie charts.

Examples of graphs from Playfair and modern examples

Abraham Wald (1759-1823) data on WWII plane holes and survival

Data visualisation important takeaways

1. Visualising data is a crucial way to understand yourself and describe to others the "what" (descriptive inference) and the "why" (causal inference).
2. It is often easier to interpret than data tables.
3. Several popular visualisation types include maps, line charts, bar charts, scatterplots, and histograms.
4. Visualisations for others are useful only if they are clear, accurate, and are conveying a direct message.

LECTURE PART 4: A FEW ASSESSMENT TIPS

Revised problem statement tips

Do the reading; attend/watch the lectures; attend tutorials; ask questions.
You are outlining your plan for a <u>3,000</u>-word essay that takes the best theory and qualitative evidence bits from your last essay and then using a different form of evidence.

Revised problem statement

1. The relevant <u>literature</u> (can use text from previous assessments)
2. Your research <u>question</u> (ditto)
3. <u>Argument</u> and observable <u>hypothesis</u> (ditto again)
4. <u>Evidence</u> (aha!, this is the major change)

    (a) What quantitative data are you using; (b) why do you think they measure your theoretical concepts; (c) what is your unit of analysis; (d) what is your spatial and temporal coverage; (e) what kinds of descriptive and inferential methods will you use; (f) how will you tie your results back to your argument and hypothesis?

Final essay suggested structure

1. The relevant <u>literature</u> (can use text from previous assessments)
2. Your research <u>question</u> (ditto)
3. <u>Argument</u> and observable <u>hypothesis</u> (ditto again)
4. <u>Research design</u>

5. Results description and discussion (important points; alternative explanations; strengths and weaknesses of your approach)
6. Conclusions

Final essay tips

1. Use the lectures, readings, and exercises to your advantage.
2. Use the descriptive and inferential statistics we cover.
3. Show that you understand what they are, what they are telling us, and why it matters for your argument.
4. Use a clear structure (subheadings, page numbers) and proofread your work.
5. Be creative.

WEEK 7 TUTORIALS

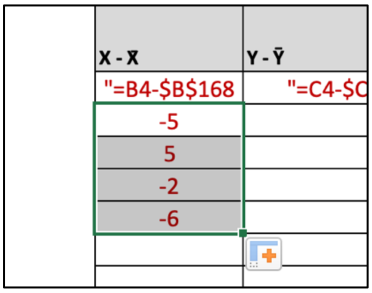
In this week's tutorial, we will be discussing and applying the concepts of correlation and visualisation discussed in the readings and lecture.

Part I: Calculating correlations (groups of 3 or 4 students)

In this section, you will be following the same process I used in lecture to calculate a correlation and in Part II run a Student's t-test of the correlation coefficient.

I have uploaded an updated version of the Polity and CPI datasets we analysed in Week 6 into Wattle/Week 7/Tutorial. I have merged these two datasets keeping one substantive variable from each, the 2018 value of polity2 and the overall 2021 CPI measure. I also included the 165 names of the countries we have data on both values for.

Complete Table 2. I have found it easier to break the elements of more complicated formulae into separate columns instead of trying to put it all into one long formula (which one can do if desired). The formulas you will need in each cell are in red in row 3. The goal here is to do the process without wasting too much time figuring out the right command. All you must do is copy the values in red without the (") mark that is in the front of each command. You want to copy the formula in row 4 for each column then drag the bottom right of the cell down to row 166 (Zimbabwe's value).



	X - X̄	Y - Ȳ
	"=B4-B168	"=C4-$C
	-5	
	5	
	-2	
	-6	

Please submit all values using the one link on Wattle/Week 7/Tutorial.

1. What is the value of the Pearson's correlation coefficient you calculated in cell I171?

2. What is the value of correlation coefficient you calculated using the one command in cell I173? Is it the same as the value in I171?

If you get ambitious you can also use the "Correlation" command in the Analytical TookPak you downloaded and used in Week 6. Is the answer the same?

Part II: Seeing if the correlation is statistically significant (same groups)

Now let's see if that value in cell I171 is statistically significant using a Student's t-test. To do so, **complete Table 3.** You need to fill in four values before you calculate the T-score.

3. What is your t-score?
4. In Figure 1 what is the threshold T-score given 161 degrees of freedom?
5. Is your t-score greater than the threshold value?
6. What does this tell you about the relationship between Polity and CPI?

Part III. Data visualisation (same groups)

Now that you are experts on how these two variables are correlated, let us see if you can visualise these variables in a few ways. Given the chart types we discussed in lecture, try:

7. Generate a histogram of each variable. Screenshot it (and those below) and submit them to Wattle. To generate these (and other) charts and figures. Go to the dropdown menu/Insert/Chart/then the type of chart you want to generate. You can also play around with the formatting to try use the best practices we read about this week.
8. Generate a scatterplot of both variables.
9. Extra credit if you can (1) move the Y-axis to the left side and (2) add a linear trendline and describe what it tells you and whether you can see any connections between this trend and the correlation coefficient you calculated.

Part IV. Connections to your research (same groups & as a tutorial)

If your group has any extra time before the end of the tutorial, try and make connections between this week's material and your own research paper ideas.

10. Can you see using any of these methods in your final paper? If so, how? If not, why?