POLS2044 WEEK 9
Probability and hypothesis testing

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

In Week 9 of POLS2044 we will be focusing on understanding the basics of probability and several core statistical tools in the modern political science toolbox.

This week I have two main goals. First, I want students to understand the basic elements of probability and how we can calculate probabilities. Second, I want students to learn several new ways of analysing the correlations between political and economic outcomes (Y variables) and potential causal factors (X variables), understand when each are useful, and how and why they are calculated the ways they are.

Reading notes and questions

There are two readings for this week.

Hibbs, Douglas A. 1978. "On the Political Economy of Long-run Trends in Strike Activity." ***British Journal of Political Science*** **8(2):153–175.**

This article is ancient in political science terms. However, it was substantially influential in the literature focused on labour strikes by outlining a clear theoretical argument for long-run industrial strike behavoir. It is also useful as an early effort at (1) creating an index for use as an outcome variable and (2) modelling the causal relationships between strike behaviour (Y) and

An early part of this paper describes three distinct dimensions of strike activity.

1. What are these three dimensions?
2. How are these three measures combined into a single indicator of strike volume?
3. What is your view on the usefulness of a combined indicator rather than using one dimensional measure?

Hibbs (1978) plots the volume of strike activity over time for each of the eleven countries he studies.

4. How much variation is there over time in these countries?
5. How easy or hard is it to compare countries' relative strike volume in the same year using these figures?
6. Is there another way that he could have displayed this information visually?

Hopefully, you will notice that in this paper he first defines his outcome, provides summary information about the outcome, and then he moves into a discussion of the causal mechanisms he argues shapes the frequency of this outcome.

Figures 5 and 6 report Hibbs (1978) main quantitative findings—correlations between several explanatory factors (Socialist/Labour party power and public sector spending) and the outcome (change in strike volume).

7. If you were to build on Hibbs (1978) for your final paper and run a multivariate regression analysis, what additional factors do you think are important to shaping strike activity that are not included in Hibbs' analysis?

Wheelan, Charles. 2013. "Chapter 5: Basic Probability: Don't Buy the Extended Warranty on your $99 Printer," in *Naked Statistics: Stripping the Dread from the Data*. London: W.W. Norton: 68-89.

This chapter's empirical examples are quite American-focused. Nevertheless, this is an approachable introduction to basic probability. By the way, if you are curious about the beer taste test example he begins the chapter with, you can view the commercial at https://youtu.be/P9a3K2vkvrU/ .

The chapter has a number of different examples of the use of probability from flipping a coint to DNA identification.

8. What is the probability of two independent events both happening?
9. Why is the assumption that these events are independent so important?
10. Why does Wheelan think calculating the expected value of an outcome is so important?
11. Would calculating the expected value of playing the Oz Lotto change your decision about whether or not to play next week?

Wheelan (2013) talks about the importance of outcomes' probabilities and expected costs to the insurance industry.

12. What is his broader lesson about when we should get insurance and when we should skip it?

LECTURE PART 1: Introduction

Why use statistics?

> "It's easy to lie with statistics, but it's hard to tell the truth without them" — Andrejs Dunjels

All research involves making choices.

What research question do I have?
What argument do I want to make?
What are the observable implications of my argument?
Which descriptive and inferential statistics do I want to use?

Mean vs. median

Neither the mean nor the median is hard to calculate, what is harder is deciding which one gives a more accurate measure of the middle in a particular situation.

What is harder is deciding which one gives a more accurate measure of the middle observation in a particular situation.

Medians are not sensitive to outliers.

Median example: household income

Graph of household income with mean substantially higher than median

Stephen Jay Gould (1941-2002)

Raw numbers vs. percentages

How many voters chose Labour in 2022?
vs.
What percentage of voters chose Labour?
Labour % = Labour vote/total vote
With percentages, you can break up values into groups.
This helps explain the gap between the mean and median in the previous slide.

Equalised disposable household income (EDHI) graph

Inflation (Consumer Price Index) graph

Raw numbers vs. indices

Should we use raw values or combine them to simplify complex phenomena into one index?

Which unit of analysis—people or countries?

Example: Is globalisation making income inequality better or worse?
Countries are getting more unequal (rich countries are growing faster than poorer countries).
People are getting more equal as the number of the world's poor is declining rapidly.

Four hurdles to establishing causality

1. Is there a credible mechanism connecting X and Y?
2. Can we rule out Y causing X (endogeneity)?
3. Is there covariation between X and Y?
4. Have we controlled for potential spuriousness (Z)?

Today's motivating questions

How can we try and not lie (to others or ourselves) with statistics?
How does basic probabilities underly most political science research?
Can we understand several more types of hypothesis tests?

Initial takeaways

There is no one right way to conduct quantitative analysis.
Rather we have a panoply of tools at our disposal.
To avoid consciously or unconsciously lying with data, we have to thoroughly understand:
(1) our data and methods,
(2) their underlying assumptions, and
(3) how to interpret our results.

LECTURE PART 2: Why should we care about probabilities?

What is probability?

"Probability is the study of events and outcomes involving an element of uncertainty." (Wheelan 2013: 71)

Why should we care about probability?

Because most of the time political scientists are dealing with samples instead of populations.

Probabilities help us determine which relationships are statistically significant.

In other words they are unlikely to occur by chance.

Probability has several key properties.

All outcomes have a probability ranging from 0 to 1.
The sum of all possible outcomes must be exactly 1.
If (and only if) two outcomes are independent, then the probability of those events both occurring is equal to the product of them individually.
The chance of either of two outcomes happening is the sum of their probabilities if the options are mutually exclusive.
If the events are not mutually exclusive, the probability of getting A or B consists of the sum of their individual probabilities minus the probability of both events happening.
Sources: Kellstedt & Whitten (2018: 147) and Wheelan (2013: 76)

Probability pitfalls

Assuming events are independent when they are not (e.g., rain today and tomorrow).
Assuming events are not independent when they are (e.g., hot streaks).
Clusters do happen (e.g. getting struck by lightening).
There is often reversion to the mean (e.g. doing well on an exam).
Moving from aggregate statistics to predicting individual behaviour (e.g., profiling).
Garbage in, garbage out (e.g., data quality).

Analytical tools are moving faster than our knowledge of what to do with results (e.g. predictive AI).

Central limit theorem

Sample size has to be large (say >30).
The sample mean will be distributed roughly as a normal distribution around the population mean.
The sample standard deviation will equal the population standard deviation over the square root of the number of sample observations.

The standard normal distribution graph

The standard normal distribution's properties

It is symmetrical about the mean
The median, mean, and mode are the same.
It has a predictable area under the curve within a specific distance of the mean.
Skewness and kurtosis are zero.

Standard deviation formula

The **standard deviation** is:

$$sd_{\bar{Y}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_1 - \bar{Y})^2}{n-1}}$$

Where:
$sd_{\bar{Y}}$ is the standard deviation of the sample mean.
Y_1 is a value of y for observation 1.
$\bar{Y}$ is the sample mean.
n is the sample size

How are the standard deviation and standard error related?

Here is the standard error equation.

$$\sigma_{\bar{Y}} = \frac{sd_Y}{\sqrt{n}}$$

Where:
$\sigma_{\bar{Y}}$ = standard error of the sample mean
sd =standard deviation
n =sample size

This allows us to calculate the confidence intervals around the mean value.

The standard normal distribution distribution of data

Important to note!

A distribution of sample values is the frequency distribution, which does not have to be normal.

However, because of the central limit theorem, the repeated sampling distribution mean will be naturally distributed, even if the underlying frequency is not.

Confidence intervals

The lower bound of the confidence interval is the mean minus two standard errors of the mean,

The upper bound is the mean plus two standard errors of the mean.

What would happen to the confidence intervals if the sample was cut in half?

$$\sigma_{\bar{Y}} = \frac{sd_Y}{\sqrt{n}}$$

Confidence intervals example from a recent paper I wrote

Black swan events

"Probability does not make mistakes,
people using probability make mistakes."
(Whitten 2013: 100)

Expected values matter.

Expected value of the Squid Game prize (~$49 million AUD) is E(P).
E(P) = (prize)*(probability of winning)
= ($49,185,319.58)*(.002 [1/456])
= $107,862

Probability takeaways

Probabilities involve uncertainty.
Political scientists need estimates of uncertainty as we have sample data instead of population data.
Probability theory comes with important assumptions, strengths, and weaknesses.
It is largely relevant to us when determining statistical significance.

LECTURE PART 3: Why conduct hypothesis testing?

Why conduct hypothesis testing?

It forces us to clearly link our theory to its real world implications.
It forces us to think about the null hypothesis.
It forces us to frame our implications in a falsifiable manner.
Variable types determine what type of hypothesis test you would use.

Which test do we choose?

		Independent variable type	
		Categorical	*Continuous*
Dependent variable type	*Categorical*	Tabular (goodness of fit) analysis	Logit/probit
	Continuous	Difference of means test or regression	Pearson's correlation coefficient or regression

What do these tests have in common?

They use p-values in their hypothesis tests.
These p-values range from 0 to 1.
They represent the probability that the relationship we are finding is due to random chance.
They include a null hypothesis.
They do not tell us that the relationship is causal.
They do not tell us how strong the relationship is.
They do not tell us anything about the quality of our measures.

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

Ordinary least squares regression

Y= alpha + Beta * X + E + e

Where:

Y is the outcome you are trying to explain.
X is the main explanatory variable.
(alpha) is the value of Y when X=0.

(beta) is the estimated relationship between X and Y.
E is the systematic error.
e is the random error.

Ceterus paribus assumption

Latin for "other things equal."
Also short for "all other things being equal."

Goodness-of-fit example

Table 1. Canberra first preferences in 2022 Federal election

Candidate	In person		Postal		Row average
	number	%	number	%	(%)
Tim Hollo (Greens)	19,240	25.15%	2,065	20.71%	22.93%
Alicia Payne (Labor)	34,574	45.20%	4,402	44.16%	44.68%
Total voters by type	53,814		6,467		

Source: AEC (https://results.aec.gov.au/27966/Website/HouseDivisionPage-27966-101.htm)

We have the following hypotheses:

H0=Candidate vote totals are independent of how people voted.

H1=Vote totals are not independent of how people voted.

Table 2. Expectation if there was no relationship between how people voted.

Candidate	In person	Postal	Row average
Tim Hollo (Greens)	22.93%	22.93%	22.93%
Alicia Payne (Labor)	44.68%	44.68%	44.68%
Total voters by type	53,814	6,467	

Table 3. Calculating expected cell values if voting type and vote choice were unrelated.

Candidate	In person	Postal
Tim Hollo (Greens)	22.93% of 53,814= .2293*53,814=12,339.55	22.93% of 6,467= .2293*6,467=1,482.88
Alicia Payne (Labor)	44.68% of 53,814= .4468*53,814=24,044.10	44.68% of 6,467= .4468*6,467=2,889.46

Table 4. Observed and expected values

Candidate	In person	Postal
Tim Hollo (Greens)	Observed (O) = 19,240	Observed (O) = 2,065

	Expected (E) = 12,339.55	Expected (E) = 1,482.88
Alicia Payne (Labor)	Observed (O) = 34,574 Expected (E) = 24,044.10	Observed (O) = 4,402 Expected (E) = 2,889.46

Now we can use Karl Pearson's chi-squared (χ^2) test for tabular association.

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(19{,}240-12{,}339.55)^2}{12{,}339.55} + \frac{(2{,}065-1{,}482.88)^2}{1{,}482.88} + \frac{(34{,}574-24{,}044.10)^2}{24{,}044.10} + \frac{(4402-2289.46\,)^2}{2{,}889.46} =$$

$$= \frac{(6900.45)^2}{12{,}339.55} + \frac{(2582.12)^2}{1{,}482.88} + \frac{(10529.9)^2}{12{,}4044.10} + \frac{(2112.54\,)^2}{12{,}889.46} =$$

$$= 3{,}858.83 + 4{,}496.21 + 4{,}611.48 + 346.24 = 13{,}312.76$$

Degrees of freedom is df=(r-1)(c-1) where r is the number of rows and c is the number of columns. In our case the answer is (2-1)(2-1)=1

We then go to a table of critical values of Chi-squared, and with 1 degree of freedom the critical value at .05 level is 3.841.

You can also calculate the chi-2 in Excel using the following command.

=CHISQ.DIST.RT(x, deg_freedom)

Hypothesis testing important takeaways

There are different types of hypothesis tests for different types of data and hypotheses.
They all involve some form of significance test.
These significance tests rely on probabilities related to the distribution of the sample means.
A tabular approach is one (currently uncommon) approach.

LECTURE PART 3: Why should we run a regression?

Why run a regression?

What if we are interested not just if there is a statistically significant difference in a sample (goodness of fit)?
Or a more complex understanding of the directionality and significance in the relationship between an X and Y than a simple correlation can tell us?

Estimating the relationship between X and Y

It can be shown that the least-squares estimators of α and β, which we call $\hat{\alpha}$ and $\hat{\beta}$, are given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}$$

where

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{and} \quad \bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

Industrial strikes example

Correlation coefficient and R-squared

The Pearson correlation coefficient (r) is used to identify patterns between variables.
The coefficient of determination (R^2) is used to identify the strength of a model.

Regression takeaways

Ordinary least squares regression is about fitting a line that minimises the (squared) distance between sample values and the line.
A basic regression provides us with two important estimates:
(1) the slope of the line summarising the relationship between x and y
(2) the expected value of y when x=0
The intuition helps us understand what it can and cannot tell us.

Today's motivating questions

How can we try and not lie (to others or ourselves) with statistics?
How does basic probabilities underly most political science research?
Can we understand several more types of hypothesis tests?

WEEK 9 TUTORIALS

In today's tutorial, we are going to get our hands dirty with two games that centre on reinforcing our understanding of probabilities. Do not worry! You will have a chance to play around with regressions in Excel soon. Today, however, I want us to focus on wrapping our heads around probabilities and how our intuition can, occasionally, lead us astray.

As normal, we are going to work in groups of four students while regularly checking in with the group and tutor to see whether your findings are matching what others are finding.

Part 1: Flip a coin

Our first game is a classic. Unfortunately, I do not have a stack of coins laying around to share with tutors and students, so we are going to have to rely on an online version of this game. The basic idea is simple. Toss a coin in the air and write down whether it comes down heads or tails. A fair coin has an equal population probability (.5) of landing on heads or tails. What we are going to do now is see how sample heads and tails percentages can vary from this underlying population probability.

Go to the following website: https://flip-a-coin-tosser.com/. Try clicking the blue "Start Flipping Coin" button. Does it come up heads or tails? I got tails. See on right side of the screen the heads and tails percentages. For me, the tails percentage is 100%.

Well, this is not as much fun as flipping a coin in person, but then we do not have to worry about debating whether flipping the coin on the back of our hands or on the floor is fairest (the literature suggests the latter).

Now let us see how the heads and tails percentages change as our sample size changes. On the top of the page, there are several options, from flipping two times to 10,000 times. Given the lecture and readings, we should expect that the standard deviation of our summary percentages of heads and tails should decrease as the sample increases as the $\sqrt{n}$ is in the denominator of the standard deviation equation. Let us run the coin toss for a few sample sizes. We are stopping at 100 due to time constraints.

Table 1. Coin flips

	Column 1	Column 2	Column 3	Column 4
How many coin flips?	Heads % (A)	50% - A (e.g., 50-46=4%)	Tails % (B)	50% - B (e.g., 50-49=1%)
1				
10				
50				
100				

1. *Did the proportions of heads get closer to 50% as the sample size increased? Put differently, did the numbers in columns 2 and 4 get smaller?*

Pretend that you could download the coin flip data and run some descriptive statistics on the. Also assume that we assign values to heads and tails turning up. For instance, let a heads = 1 and a tails = 0. So, if I flipped 6 heads and 4 tails, the sample mean would be 0.6.

2. *How do you think the standard deviation values of the calculated means would change as the samples increased?*

Part 2: Who is Monty Hall and what is with all these goats?

Next, we are going to play a bit more complicated game, called the Monty Hall game or the Monty Hall problem. It is probably the most famous example of how our intuition can lead us astray (probability-wise). An online version of the game can be found at https://www.mathwarehouse.com/monty-hall-simulation-online/.

On the website, you will see three doors. Click on a door. After you do so, a second door will open and reveal a goat. Do you keep your original choice, or do you switch to the second unopened door? It is up to you. Play a few times to get a hang of the game.

Given time is short, let us speed the process of running the game multiple times. Below the doors, you will see a tab labelled "simulate." Click it. A dotted box appears, and you can now simulate multiple runs of this game. You are going to divide your group into one half that keeps their door choice, and the second half switches choices.

Run the simulation three times. The first running the game 10 times, then 100 times, finally 1,000 times with either "change choice" or "keep the choice" in the middle drop-down option. I would recommend you changing the setting in the final dropdown box to "instant." Keep track of your results in Table 2.

Table 2. Monty Hall game

	Change choice		**Keep choice**	
	number	*percentage*	*number*	*percentage*
10 times				
Car				
goat				
100 times				
Car				
goat				
1,000 times				
Car				
goat				

3. *What percentages for cars and goats seem to be the mean value that the samples are converging to as the sample increases?*

4. *Does this process suggest that changing choice or keeping choice maximises the probability of getting the car?*

5. *Is this what you expected? Why or why not?*

For a recent take on the Monty Hall Problem see https://youtu.be/ggDQXlinbME.