POLS2044 WEEK 7
Probability and statistical inference

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

In Week 7 of POLS2044 we will be focusing on understanding the basics of probability and apply them to real world examples.

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 have the opportunity to calculate several statistics themselves.

Reading notes and questions

There is one reading for this week, Chapter 7 (pp. 143-160) from the Kellstedt and Whitten (2018) textbook.

Some students have already asked me in person whether it is necessary to know the equations and their meaning in this and other chapters. The short answer is yes. We will work through them in more detail during the workshop, but spending time working at your own speed should help you get initially comfortable with them. They will be used repeatedly over the rest of the semester.

There are four main elements of the sampling distribution I want you to get to know and be able to apply this week:

Sample mean (p. 136)
Standard deviation (pp. 137 & 151)
Standard error of the mean (p. 152)
Confidence interval lower and upper bounds (p. 153)

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

Today's motivating questions

How can we try and not lie (to others or ourselves) with statistics?
Can we better understand probabilities through a few examples?

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.

In this section

1. Key properties of probability
2. Central limit theorem
3. Standard normal distribution
4. Confidence intervals
5. Expected values

Probability has several key properties.

1. All outcomes have a probability ranging from 0 to 1.
   a. Example of probabilities of a six-sided die
   b. Probability=number of relevant outcomes/total number of possible outcomes

2. The sum of all possible outcomes must be exactly 1.
   a. Example of distribution of sum of two dice probabilities

3. If (and only if) two outcomes are independent, then the probability of those events both occurring is equal to the product of them individually.

4. The chance of either of two outcomes happening is the sum of their probabilities if the options are mutually exclusive.

5. 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.

Probability pitfalls

1. Assuming events are independent when they are not (e.g., rain today and tomorrow).

2. Assuming events are not independent when they are (e.g., hot streaks).

3. Clusters do happen (e.g., getting struck by lightning).

4. There is often reversion to the mean (e.g., doing well on an exam or class).

5. Moving from aggregate statistics to predicting individual behaviour (e.g., profiling).

6. Garbage in, garbage out (e.g., data quality).

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

Black swan events

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

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.

Key point: The sampling distribution is normally shaped even though the underlying frequency distribution is not normally shaped.

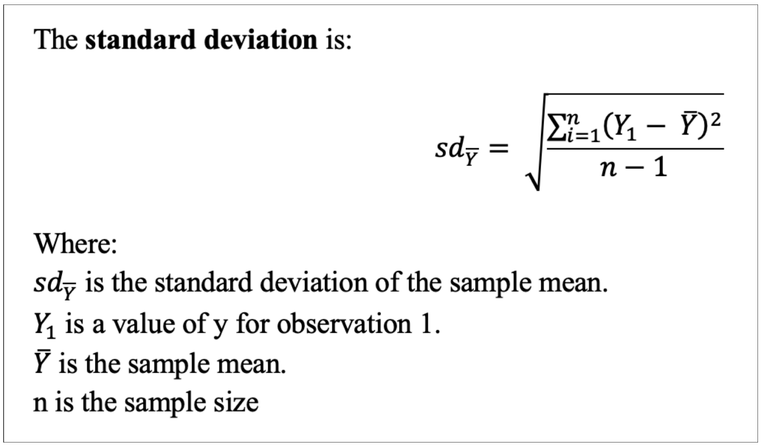
The standard normal distribution graph

The 68-95-99 rule

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

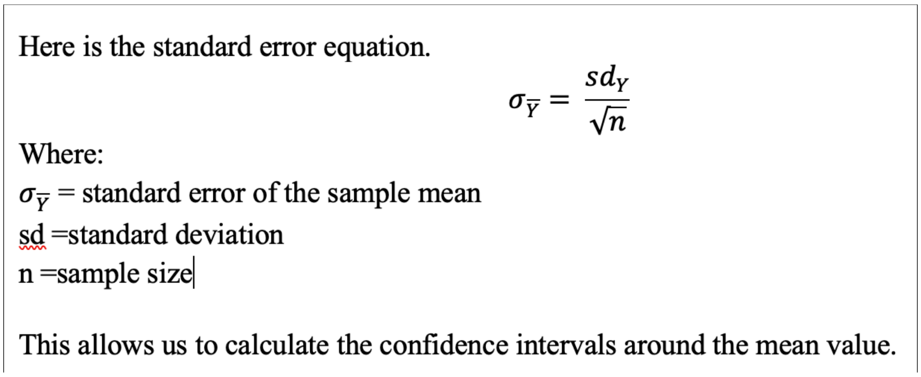


The **standard deviation** is:

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

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.

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 the margin of error (or two standard errors) of the mean,

The upper bound is the mean plus the mean's margin of error.

Confidence intervals example from a paper I wrote

Probability and expected values

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

Expected value of this week's Powerball

Expected value of this Thursday's Powerball ($20 million AUD) is E(P).
E(P) = (prize)*(probability of winning)
E(P) = ($20,000,000.00)*(.00000000744 [1/134,490,400])
E(P) = $0.15

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: Probability in action

Example #1: The Voice public opinion surveys

Professor Simon Jackman's graph of public opinion over time

https://simonjackman.github.io/poll_averaging_voice_2023/poll_averaging.html

Essential research poll example

Chosen because they have relatively clear information about their methodology and are a relative outlier in their latest poll.

Referendum sources of information overview graph

Essential research methodology disclosure statement

Essential research's margin of error

From margin of error to confidence intervals

How can we connect our knowledge of probability to better understand polling results?

By converting the polling results into measures of confidence that the population mean is within a certain range around the sample mean.

The margin of error is half the width of the confidence interval.

The confidence interval is thus twice the margin of error centred on the sample mean.

Essential research confidence intervals

So if the Sept 2023 poll suggests that 48% no, 42% yes, 10% undecided with a 3.1% margin of error…

Response	Percentage	Lower 95% bound	Upper 95%
No	48	44.9	51.1
Yes	42	38.9	45.1
Undecided	10	6.9	13.1

If you are interested in polling research on the Voice

Groot, Murray. 2023. "Support in the Polls for an Indigenous Constitutional Voice: How Broad, How Strong, How Vulnerable?" *Journal of Australian Studies* 47(2): 373-397.

Example #2: Detecting election fraud

Based on Beber and Scacco (2012) analysis of Senegal 2000 & 2007 and Nigeria 2003 elections

Important Week 7 terms

Central limit theorem
65-95-99 rule
Normal distribution
Sampling distribution
Population distribution
Standard error of the mean
Standard deviation
Statistical inference
Expected value
Margin of error
Confidence intervals

WEEK 7 WORKSHOP

In today's workshop, we are going to get our hands dirty with several exercises geared towards getting familiar with the following four fundamental terms.

PART 1: CALCULATING MEANS, SD, SE, AND CONFIDENCE INTERVALS

This first part is submitted individually. Please download the spreadsheet entitled "POLS2044 2024 Week 7 Workshop". Feel free to work through the steps as a group, but you need to calculate and submit your own spreadsheets with your own U-number.

Step 1: Enter your university ID numbers into column E in the grey cells.
Step 2: Calculate the mean and standard deviations for these values in cells E12 and E13.
Step 3: Calculate your squared standard deviations in F2-F8.
Step 4: Sum these deviations in Cell F17, divide the sum by n-1 in Cell F18, and take the square root in Cell F19. Your value in Cell F19 should equal that in E13. If not, go back and fix any mistakes.
Step 5: Calculate the standard error by dividing the standard deviation by the square root of the number of observations.
Step 6: Now check whether this is the same amount as given in the Analytics TookPak. Follow the options I included in the screenshot in the spreadsheet.
Step 7: Now calculate the 95% confidence bounds both within Excel and manually.

Hopefully, during this process you have developed a better understanding of how these basic statistics are calculated. Of course, there is not some underlying population mean values for your u-number, but if you use these techniques with real data, you would be able to estimate statistics about the sample and make connections to the underlying population. If you ever get stuck, do raise your hand and Sajjad and I will come by and try and help.

Rename your completed spreadsheet with your name and upload this document to Wattle when you are done (POLS2044/Week 7/Workshop/Item 7.1).

PROBABILITY GAMES

Next, you will be working together as groups to play two games that centre on reinforcing our understanding of probabilities. I want us to focus on wrapping our heads around probabilities and how our intuition can, occasionally, lead us astray.

Part 2: 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 (0.5) 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?***

Paste (a) your completed Table 1, (b) your answers to questions 1 and 2, and (c) the names and U-numbers of your group when you are done to POLS2044/Week 7/Workshop/Item 7.2.

Part 3: 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. Then run 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?***

Paste (a) your completed Table 2, (b) your answers to questions 3-5, and (c) the names and U-numbers of your group when you are done to POLS2044/Week 7/Workshop/Item 7.3.

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