POLS2044 WEEK 8
Bivariate Hypothesis Testing

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

In Week 8 of POLS2044 we will be focusing on ways of testing hypotheses related to the relationships between two variables. This week builds directly on the probability discussion from last week and takes us towards testing hypotheses about relationships between more than two variables in later weeks.

This week I have three main goals. First, I want to introduce you to hypothesis testing and how we gauge statistical significance. Second, I want to discuss two types of tests—a difference of means test and the correlation coefficient. Third, I want to give you an opportunity to practice using these tests and gauging statistical significance.

Reading notes

There is one assigned reading, Chapter 8 from the textbook. While the textbook does cover tabular tests of categorical X and Y variables, I will not be discussing it in lecture as it is rarely used in political science (in my limited experience) and it will enable us to talk more about the other two tests and p-values.

LECTURE PART 1: Introduction

My diagram of different types of statistics

Today's motivating questions

What are some of the most common forms of bivariate hypothesis testing?

How do they work?

Why are they useful?

LECTURE PART 2: 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.

It enables us to possibly pass the third causal hurdle (covariation).

Which test should 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 "we would see the observed relationship between the two variables in our sample data if there were truly no relationship between them in the unobserved population." (KW 2018: 164).
They include a null hypothesis.
They assume the selection of a random sample from the underlying population.
They represent a comparison between the actual X-Y sampled relationship to what we expect if there was no X-Y relationship in the underlying population.
The greater the difference between reality and null expectations the more confidence we can be in the X-Y relationship in the underlying population.

What do these tests also have in common? (Limitations)

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.

Crucial ceterus paribus assumption

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

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 (relatively uncommon) approach that is in the book but which we are going to skip over.

LECTURE PART 3: Difference of means test

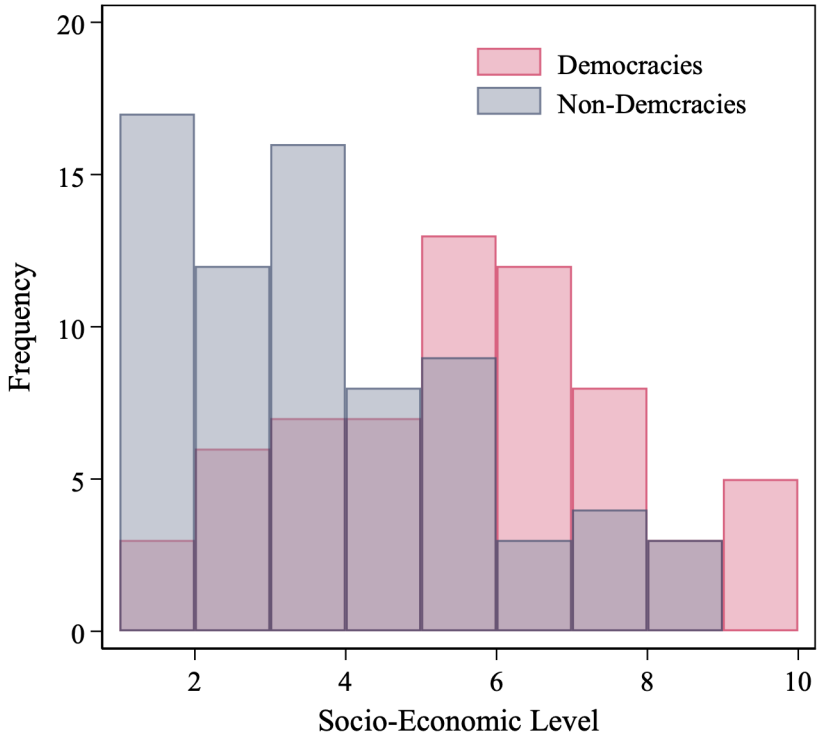
Difference of means test

Our first main bivariate hypothesis test
Useful when you have a continuous Y and a categorical X

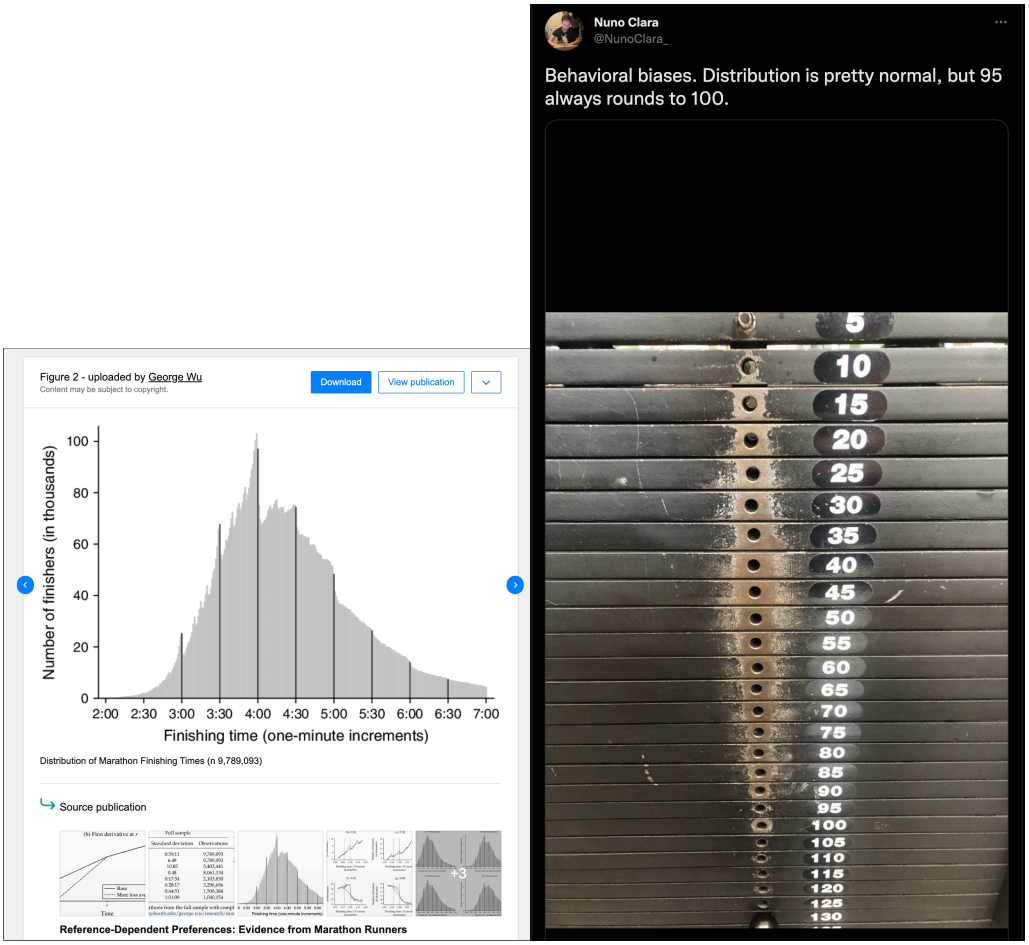
Difference of means example

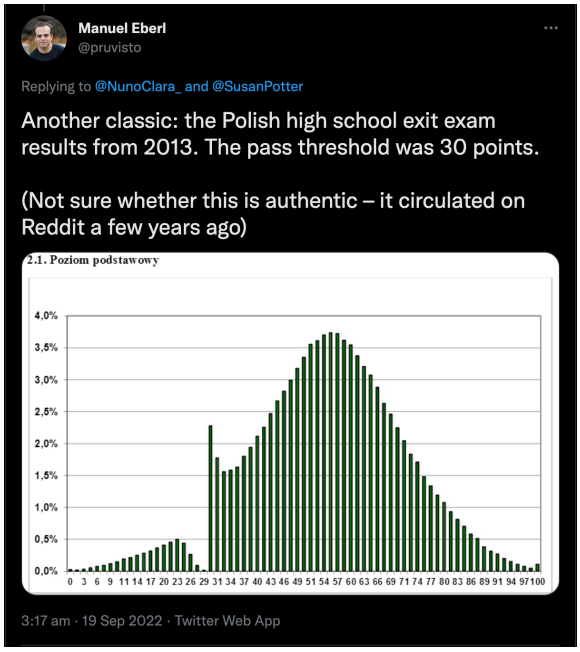
Say for example I have the following:
Research question: Does being a democracy increase a country's socioeconomic status?
Theory: _____ (insert theory here).
Research design: Collect the sample means of socioeconomic level in democracies and non-democracies
Hypothesis: These means come from different underlying distributions
Null hypothesis: These means come from the same underlying distribution

Socioeconomic level by political type



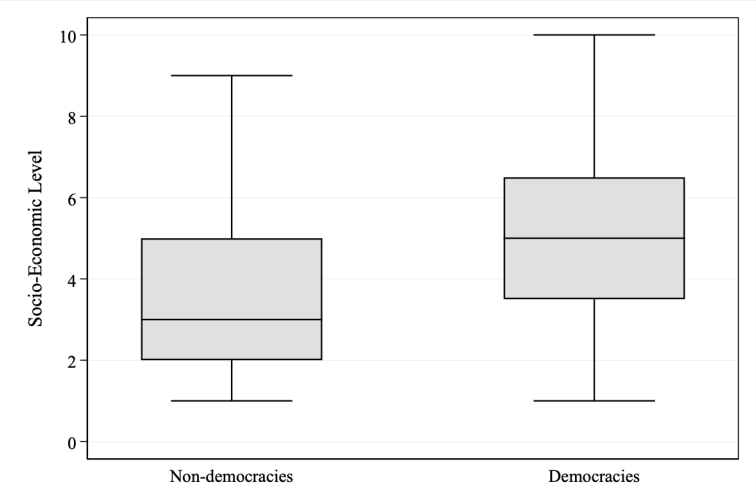
Know thy data (and what created them)





Example from Quality of Governance data –democracy and socio-economic level

Difference of means t-test



Summary statistics

	Observations	Mean	Standard error
Democracies	**64**	**5.14**	**0.28**
Non-democracies	**72**	**3.29**	**0.24**

Difference of means t-test intuition

The t-statistic basically is a measure of the difference of means over a measure of uncertainty around those means.

T-distribution & P-values

Difference of means example

T-statistic: -5.05
Degrees of freedom: 134
P-value: 0.000
Therefore, I conclude that there is less than less than 1 in 1,000 chance that we would see this relationship randomly in our sample if there was no relationship in the underlying population.

LECTURE PART 4: Correlation

Correlation

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.

Correlation does not imply causation.

Five main features of correlation (Tacq 2004)

1. The nature of the correlation

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

2. 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)

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

4. The strength of the correlation

5. 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

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

Example: Eurovision 2022 Finals

Why conduct a significance test?

So we have calculated a correlation coefficient.
However, 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?

Conducting a significance test
(rho) is the correlation coefficient.
Null hypothesis ( ): , there is not a significant linear correlation between x and y in the sample.
Alternative hypothesis ( ): , there is a significant linear correlation between x and y in the sample.
Now we conduct a Student's T-test. What is that?

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.

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.

Important Week 8 terms

Chi-squared test
Covariance
Correlation coefficient
Degrees of freedom
Difference of means test
P-value
T-statistic
Tabular analysis

WEEK 8 WORKSHOP

In this week's workshop, we will be discussing and applying the concepts of hypothesis testing to a difference of means test and a correlation coefficient as discussed in the readings and lecture. Like last week students are going to work on their own dataset while helping each other to work though the different formulae.

How are we going to do this? By finally using those dice I bought. Each table (of ideally four) students will take turns rolling a die. If a student gets the same number as a student who has already rolled, they should roll again until they get a number that has not already been selected. The number you have is the number of the dataset you will be analysing today.

Please go to Wattle and download the corresponding dataset in Wattle/Week 8/Workshop data.

Dataset	Description
1	Corruption and Polity
2	Quality of Governance indicators
3	Starbucks drinks menu nutritional information
4	Top Twitch streamers in 2020
5	City temperatures on 1 January 2013
6	Top 100 horror movies of all time

The first sheet of each dataset (except Dataset 1) includes data descriptions. The second sheet will be used to run a difference of means test. The third sheet will be used to calculate a correlation coefficient and significance test.

Please return the die when you are done to help make sure I have dice left to teach this class next year. Thanks!

When you are done with all parts of this week's workshop activities, please submit both your spreadsheet and your answers to the ***questions*** below to the Workshop submission link.

Part I. A difference of means test

For Part 1, please go to the second sheet ("difference of means") of your unique Week 8 workshop dataset. You will see in Column A the unit names (e.g., a country name or horror movie title). The other columns include a selected number of variables from the original (larger) dataset. The columns in green are the ones we are going to focus on.

Dataset	Description
1	Corruption in democracies and non-democracies
2	GDP in European Union (EU) and non-EU countries
3	Calories in hot and cold Starbucks drinks
4	Number of followers of English language and non-English language channels
5	Temperatures in Northern and Southern hemisphere cities
6	Revenue earned by movies that are either longer or shorter than 105 minutes

Hopefully, you will see that in each set of green columns there are two potential (continuous) dependent variables differentiated by a potential (categorical) independent variable. What we want to figure out is whether the average values of the columns are statistically different from each other. To do so, we will be running a difference of means test. Before we do so, we need to be clear about our theoretical expectations.

1. ***Please write (a) a suitable hypothesis and (b) a null hypothesis for your difference of means test.***

   Hint—Look at the H0 and H1 outlined in lecture and see if they can be reworked for your purposes here.

Next, run your difference of means test. While I did show you the underlying equations in lecture, I am not going to ask you to take it apart and put it back together today. Instead, we will ease back into Excel and use the Data Analysis ToolPak.

Side note—Please be sure that you have not installed the XLMiner Toolpak, which also shows up when you search for add-ons.

A screenshot on the right of the table shows the type of T-test you want to run as well as suggestions for the values to enter. We are using the test that assumes sample means with unequal variance because (1) we have different sample sizes in the columns, and (2) my preliminary analysis suggests that they do have different variances.

Once have your result, does the hypothesized mean difference make sense given what you might know about these data? More technically, what would it mean to change the alpha value of your difference of means test?

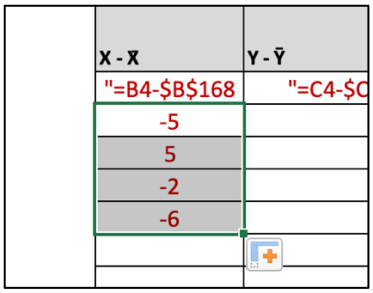
2. ***Do your results support rejecting the null hypothesis in favour of your alternate hypothesis or do you fail to reject the null hypothesis? How do you reach this conclusion?***

3. ***Given your research interests, can you think of a possible use for a difference of means test using data relevant to your interests? What might it be?***

Part II: Calculating a correlation coefficient

In this section, you will be calculating a correlation coefficient and run a Student's t-test of the correlation coefficient. This exercise is in the "Correlation" sheet of your spreadsheet. This time we will be breaking up the underlying equation into its constituent parts. 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 later). The formulas you will need in each cell are in **red** and your goal is to fill in all yellow cells.

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 the last observation.



Step 1: Calculate the mean values for your variables in Columns B and C and enter them into the two yellow cells below your data.
Step 2: Complete the yellow cells in Table 2 down to the same row as the last observation of your data.
Step 3: Calculate the sums for the three cells below these cells
Step 4: Calculate your correlation coefficient
Step 5: Run the built in correlation equation in the yellow cell to the right of "Pearson's Correlation Coefficient"
Step 6: Complete Table 3 to collect the elements necessary for the Student's t-test
Step 7: Now calculate the T-score.
Step 8: Figure out what the threshold t-score is for statistical significance using the screenshot below Table 3.
Step 9: Conclude whether the correlation is statistically significant.

4. ***What is the value of the Pearson's correlation coefficient you calculated? Is it as high (or low) as you were expecting?***

5. ***What are at least two of the five correlation features (discussed in lecture) of your correlation? Why did you choose those features to discuss?***

6. ***What is the value of correlation coefficient you calculated using the built in Excel command? Is it the same as the one you calculated? If you had problems matching these values, why do you think this happened?***

7. ***What is your t-score?***

8. ***In Figure 1 what is the threshold T-score given your degrees of freedom?***

9. ***Is your t-score greater than the threshold value?***

10. ***What does this tell you about the relationship between these two variables?***

Part III: Data visualisation

In the previous two parts of the workshop, you conducted two bivariate hypothesis tests. I tried to include a variety of different datasets and relationships that are both statistically significant and not significant at traditional levels. You have now spent a good amount of time staring at tables—probably more than you would chose to do voluntarily. This process reminds me of a

quote from Arthur and Henry Farquhar, two very early advocates of data visualization: "Getting information from a table is like extracting sunbeams from a cucumber." [1]

Therefore, in this section you are going to try and graph at least one variable included in the difference of means sheet. This connects us back to Week 6's discussion about descriptive statistics and the importance of learning about your data including how they is distributed.

The most straightforward types of graphs to start with are *histograms* and *scatterplots*. If you need a brief overview of how to create a histogram or scatterplot in Excel, I found these brief videos helpful:

How to Make a Histogram in Excel (https://youtu.be/yh5ihdHwmTk?feature=shared)
How to Make a Scatter Plot in Excel (https://youtu.be/MfEAEmdFOBo?feature=shared)

Step 1: Go back to the "difference of means" sheet and decide what variable(s) you want to graph. I would suggest aiming for continuous variables if possible.
Step 2: Decide whether you want to create a scatterplot or histogram.
Step 3: Create your graph of choice on the same sheet to the right or below your results from Part 1.
Step 4: Feel free to tweak the graph to try and make the graph as easy to read as possible.

11. What variable(s) did you choose and why did you choose them? Did the graph conform to your expectations? Why/why not? If you created a histogram, does the distribution approximate a bell curve or is it skewed or distributed in any other notable way? If you created a scatter plot, can you discern any clear pattern in the relationship between your variables?

Finally, I want to give your group an opportunity to take a step back and discuss these efforts at conducting bivariate hypothesis tests and how they connect back to the other steps of the research design.

12. What were the hardest parts of these calculations for your group? The simplest? What connections do your statistical significance tests enable us to make between our data sample and the underlying population we assume the samples were randomly selected from?

Part IV: For those who made quick work of the sections above (optional)

So, you still have time left before the end of workshop and are up for another challenge? This section is for you.

13. Go back to the difference of means table and run a correlation and Student's t-test on two variables you have not previously correlated and report (and substantively describe) your results here.

---

[1] Farquhar, Arthur, and Farquhar, Henry. 1891. *Economic and Industrial Delusions*. New York: G. P. Putnam's Sons: 55.

14. ***Go to the Quality of Governance's "Basic Dataset" site (https://www.qog.pol.gu.se/data/datadownloads/basicdataset) and download the cross-section XLSX file and codebook. If you feel more comfortable using another statistics package, feel free to use that software. Poke around the data and codebook and run some hypothesis tests and visualisations that randomly occur to you. Summarise your results here. Did anything stand out to you as being particularly interesting/weird/unexpected?***