Week 12: Wrapping up

Dr. Richard Frank | Political Analysis, 2024 | The Australian National University

Image source: https://www.fxhash.xyz/gentk/146004

Today's discussion

1. Semester recap
2. Important terms
3. Final exam

1 Recapping the last eleven weeks

Image source: https://www.fxhash.xyz/gentk/146004

1 Course outline

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

Week 1: Scientific method

Week 2: Causal theorising

Week 3: Research design

Week 4: Concepts and measurement

Week 5: Surveys and sampling

Week 6: Descriptive inference & statistics

Week 7: Probability & statistical inference

Week 8: Bivariate hypothesis testing

Week 9: Bivariate regression

Week 10: Multivariate regression

Week 11: Regression pitfalls

1

Goal 1: Help you consume **research**

Table A.1: Observer Effects on Ballot Stuffing

	Ballot stuffing	Confidence Intervals
Observer Present (OP)	**-0.037**	(-.09, .01)
	(0.025)	
	[-1.51]	
Medium Saturation	**0.022**	(-.03, .07)
	(0.024)	
	[0.92]	
High Saturation	**0.010**	(-.02, .04)
	(0.016)	
	[0.63]	
Competition	**0.019**	(-.02, .06)
	(0.018)	
	[1.03]	
Urban	**-0.007**	(-.04, .03)
	(0.017)	
	[-0.41]	
Constant/Intercept	**0.052****	**(.01, .09)**
	(0.021)	
	[2.55]	
Observations	2,004	
R-squared	0.011	
F(5,59)	1.43,	*p*-value=.223

Note: Robust standard errors in parentheses. t-statistics in square brackets.
*** p<0.01, ** p<0.05, * p<0.1.
Note: Although we provide all statistics here, generally *either* t-statistics *or* standard errors are provided. Confidence intervals are often not provided.

Source: Long. 2016: 1.

1 Goal 2: help consume **information**

HOME PAGE | The Canberra Times | SUBSCRIBE | Lo

Women in high-status positions not trusted, research from ANU shows

By Steve Evans
Updated October 21 2022 - 7:35am, first published 7:30am | 0 Comments

Dr Eun Young Song. Picture by Jamie Kidston/ANU

ADVERTISEMENT

Ad

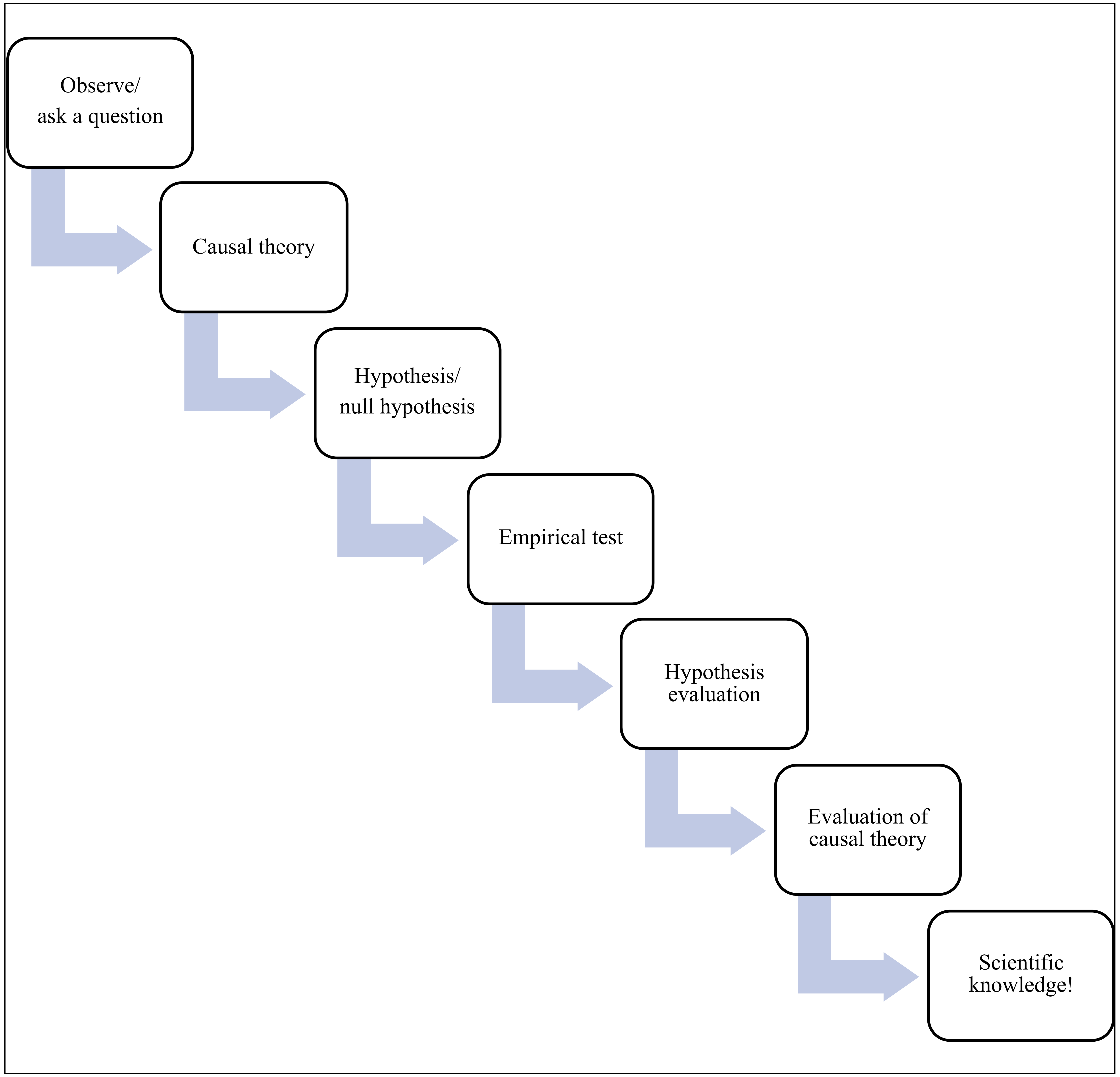
Women in high-status positions aren't trusted by the people they work with, according to new research done at the Australian National University

Source: The Canberra Times (https://www.canberratimes.com.au/story/7950724/top-women-not-trusted-anu-research-shows/?cs=14329)

1

Goal 3: help you **produce** research

1 The scientific method



1 KKV's (1994) characteristics of scientific research

The goal is causal **inference**.

The **procedures** are public.

The **conclusions** are uncertain.

The content is the **method** not the subject matter.

Often the **scaffolding** of intellectual buildings are taken down after being built.

1 Developing new theoretical arguments

- Offer an answer to an interesting **research question.**
- Solve an interesting **puzzle**.
- Identify interesting **variation** (across **time** or **space**)
- Move from a **specific event** to more general theories
- Drop the **proper nouns**
- Use a new **Y**
- Use a new **X**
- Add a new **Z**
- Use the **literature**
- Make sure the theory can be **disproven.**

1 Developing good ideas

1. Intellectual **taste**
2. **Personality**
3. Our **interests**
4. **Logic**
5. Avoids **relabelling**
6. Stands the test of **time**
7. Can be **described to others** clearly and briefly.
8. **Simplifies** the world.
9. Learning from **bad ideas**

1 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)?

1

Useful to ask a consistent series of questions

What is the **research question** or **puzzle**?

What is the main theory(ies) or **argument(s)**?

What type of **research design** is used?

How well does the work surpass the **four hurdles**?

1

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)

1

Definitions

"A **population** is any group of people, organisations, objects, or events about which we want to draw conclusions; a *case* is any member of such a population." (Brians et al. 2011: 132)

"A **sample** is any subgroup of a population of cases that is identified for analysis." (Brians et al. 2011: 132)

"A **representative sample** is one in which every major attribute of the larger population from which the sample is drawn is present in roughly the proportion or frequency with which those attributes occur in that larger population." (Brians et al. 2011: 133)

1 The present state of the study of politics

The American Political Science Review

Vol. XV MAY, 1921 No. 2

THE PRESENT STATE OF THE STUDY OF POLITICS

CHARLES E. MERRIAM

University of Chicago

The original plan of this paper included a general survey and critique of the leading tendencies in the study of politics during the last thirty or forty years. It was intended to compare the methods and results of the various types of political thought—to pass in review the historical school, the juridical school, the students of comparative government, the philosophers as such, the attitude of the economist, the contributions made by the geographer and the ethnologist, the work of the statisticians, and finally to deal with the psychological, the sociological, the biological interpretations of the political process.

It would have been an interesting and perhaps a useful task to compare the scope and method of such thinkers as Jellinek, Gierke, Duguit, Dicey and Pound; the philosophies of Sorel and Dewey, of Ritchie and Russell, of Nietzsche and Tolstoi; to review the methods of Durkheim and Simmel, of Ward and Giddings and Small; of Cooley and Ross; and to discuss the developments seen in the writings of Wallas and Cole.

It would have been useful possibly to extend the analysis to the outstanding features of the environment in which these ideas have flourished, and to their numerous and intimate relations and interrelations. It might have been possible to discuss

173

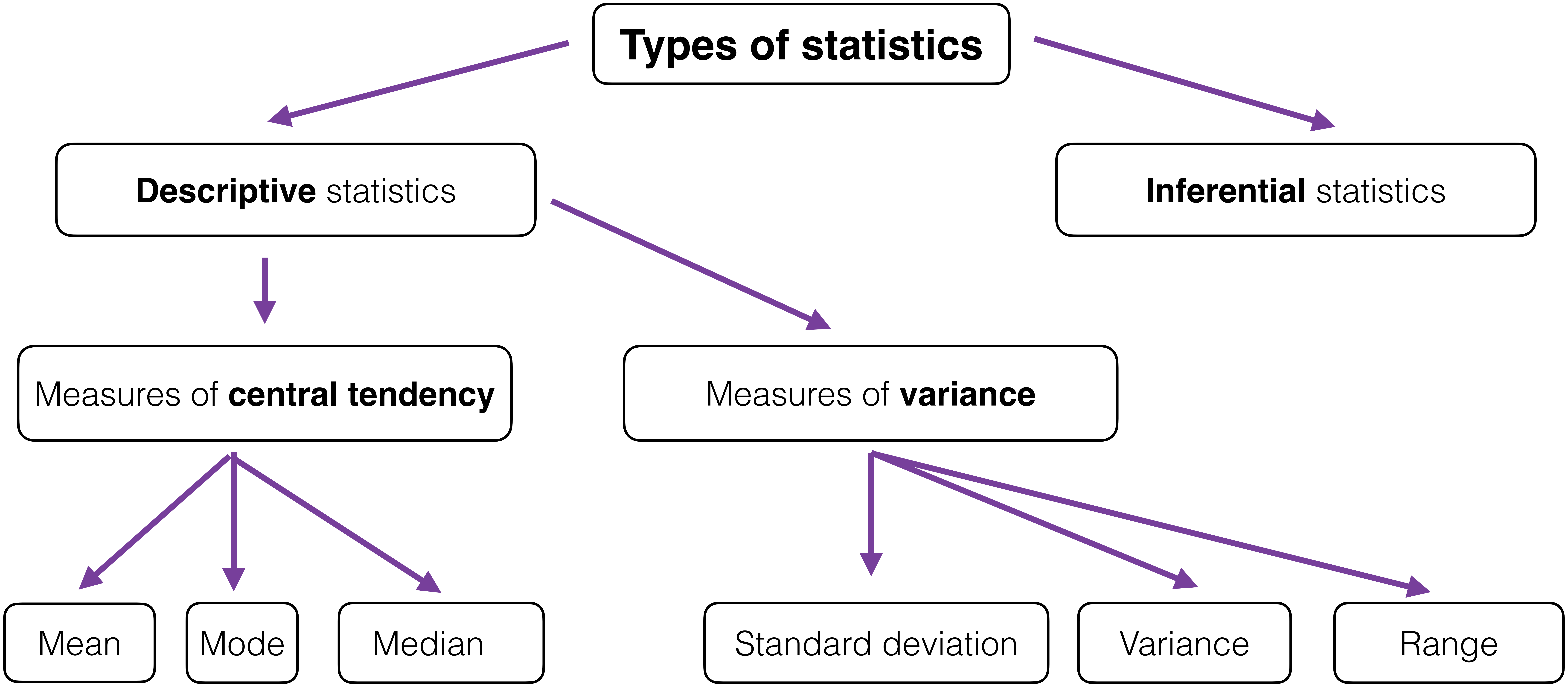
Statistics, to be sure, like logic can be made to prove anything. Yet the constant recourse to the statistical basis of argument has a restraining effect upon literary or logical exuberance; and tends distinctly toward scientific treatment and demonstrable conclusions. The practice of measurement, comparison, standard-

P. 179

clusions. We know that statistics do not contain all the elements necessary to sustain scientific life; but is it not reasonable to expect a much greater use of this elaborate instrument of social observation in the future than at present? Is it unreasonable to expect that statistics will throw much clearer light on the political and social structure and processes than we now have at our command?

P. 180

1 Statistics



1 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]

1 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?

1 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.

1

Categorical variables

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

Latest release

Data download

Language used at home (LANP)

Census of Population and Housing: Census dictionary

Reference period: 2021

Released 15/10/2021 **Next release** Unknown **Previous releases**

Definition

This variable identifies whether a person uses a language other than English at home and if so, records the main non-English language which is used. The purpose of this variable is to identify the main languages other than English which are used in households across Australia.

Scope

All persons

Categories

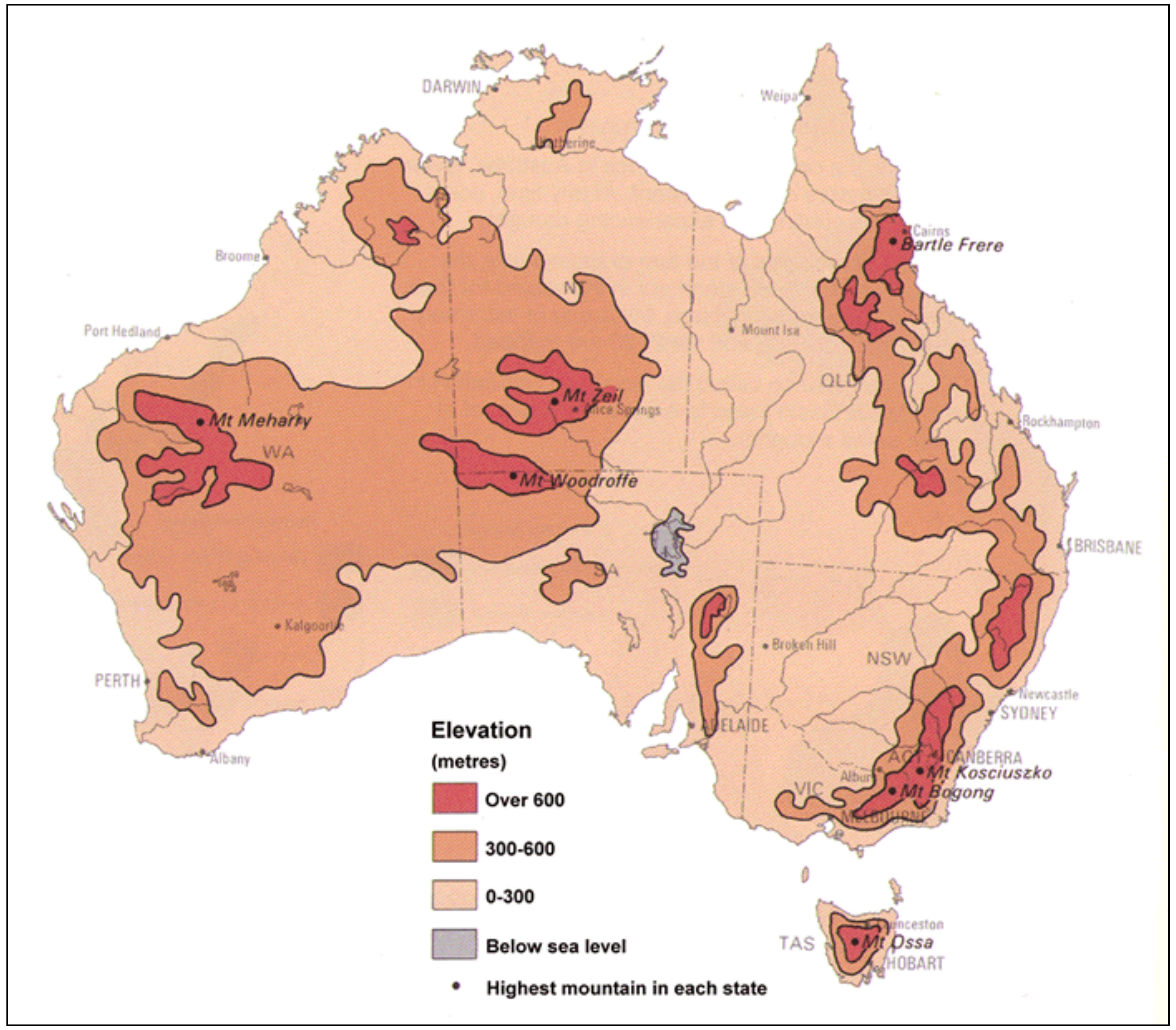
Language used at home (LANP) is classified using the Australian Standard Classification of Languages (ASCL), 2016. The categories are listed in groups below. The full list is available from the Data downloads on this page.

1 Northern European Languages
2 Southern European Languages
3 Eastern European Languages
4 Southwest and Central Asian Languages
5 Southern Asian Languages
6 Southeast Asian Languages
7 Eastern Asian Languages
8 Australian Indigenous Languages
9 Other Languages
Supplementary codes

1 Continuous variables

Sometimes called interval variables or ratio variables (if they have a meaningful 0).

They have **equal unit differences**.



Source: https://www.ga.gov.au/__data/assets/image/0013/12640/GA11759.gif

1 Describing **continuous** variables

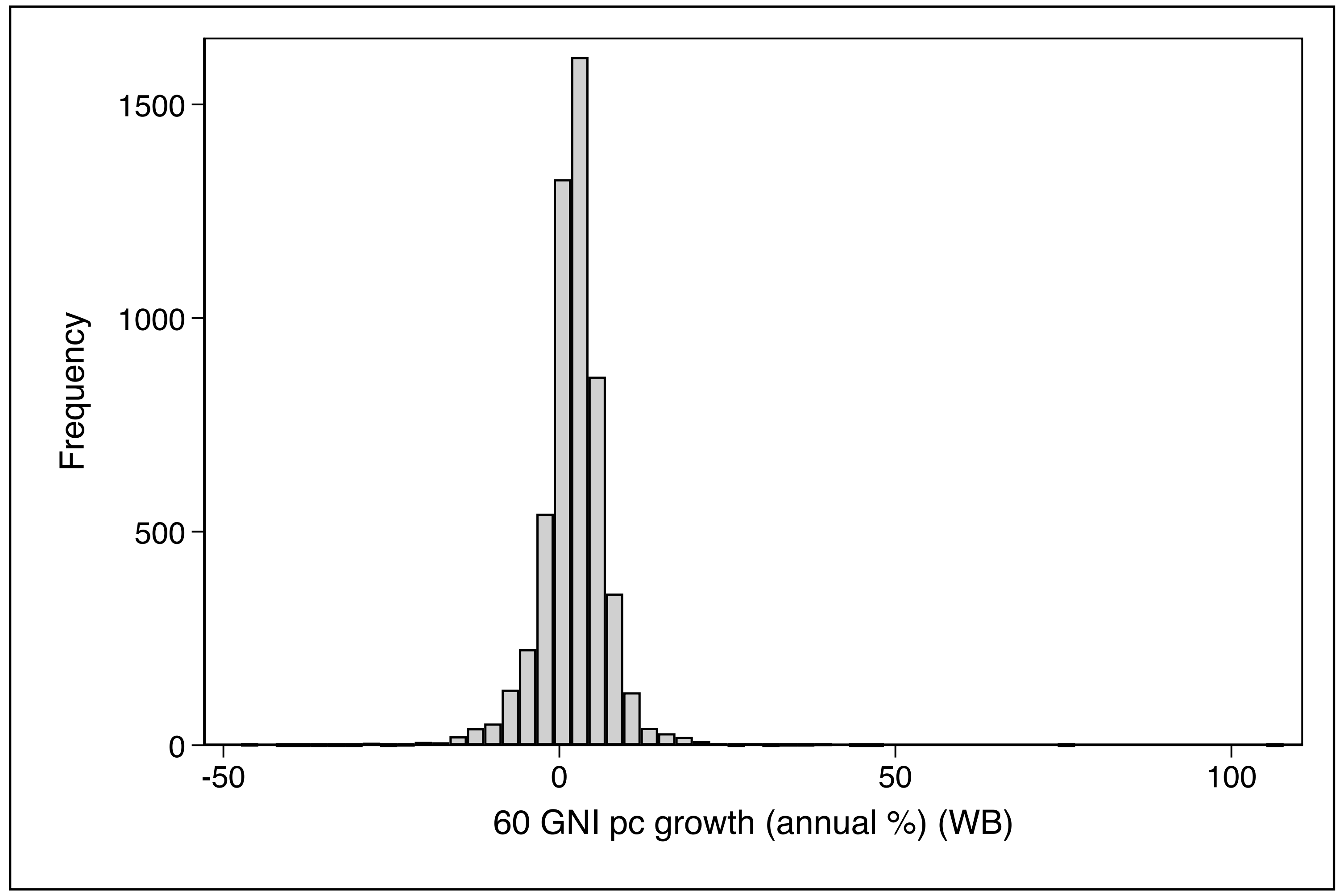
We are primarily interested in the **central tendency** and the **distribution** of values around this central tendency.

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**.



1 Moving from theory to test

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

Operationalisation | Operationalisation

Measured proxy------------------------------ **Hypothesis** -------->**Measured dependent variable**

Null hypothesis

Figure adapted from Kelstedt & Whitten (2018: 10).

1 Probability's key properties

1. All outcomes have a **probability** ranging from **0 to 1**.

2. The **sum** of all possible outcomes must be exactly **1**.

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.

Sources: Kellstedt & Whitten (2018: 147) and Wheelan (2013: 76)

1 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).

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

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).

1 Central limit theorem

Sample size has to be large (say greater than 30 observations).

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.

1 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.

1 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).

1 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

1

Research design tradeoffs

		Dependent variable variation	
		Yes	***No***
Explanatory variable variation	***Yes***	**A** Quant. design (ideally)	**B** Selecting on DV
	No	**C** Shotgun approach	**D** A case study

1 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.

1 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.

1

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 will be largely relevant to us when determining **statistical significance**.

1

Finding the **standard deviation**

A sample's **standard deviation** (*sd*) is given by $sd = \sqrt{variance(y)}$

Or more concretely:

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

Where:

$\bar{x}$ is your variable's mean.

x_i is an individual value.

n is the sample size.

1 How are the standard deviation and standard error related?

Here is the standard error equation.

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

Where:

$\sigma_{\overline{Y}}$ = standard error of the sample mean

sd =standard deviation

n =sample size

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

1 How to calculate the **margin of error**?

Sampling error at 5% significance level: $1.96\sqrt{Var/\mathrm{n}}$

With **variance** (var)=*p*(1–*p*) where is the number of respondents, and *p* is the proportion favouring an outcome.

1 From margin of error to confidence intervals

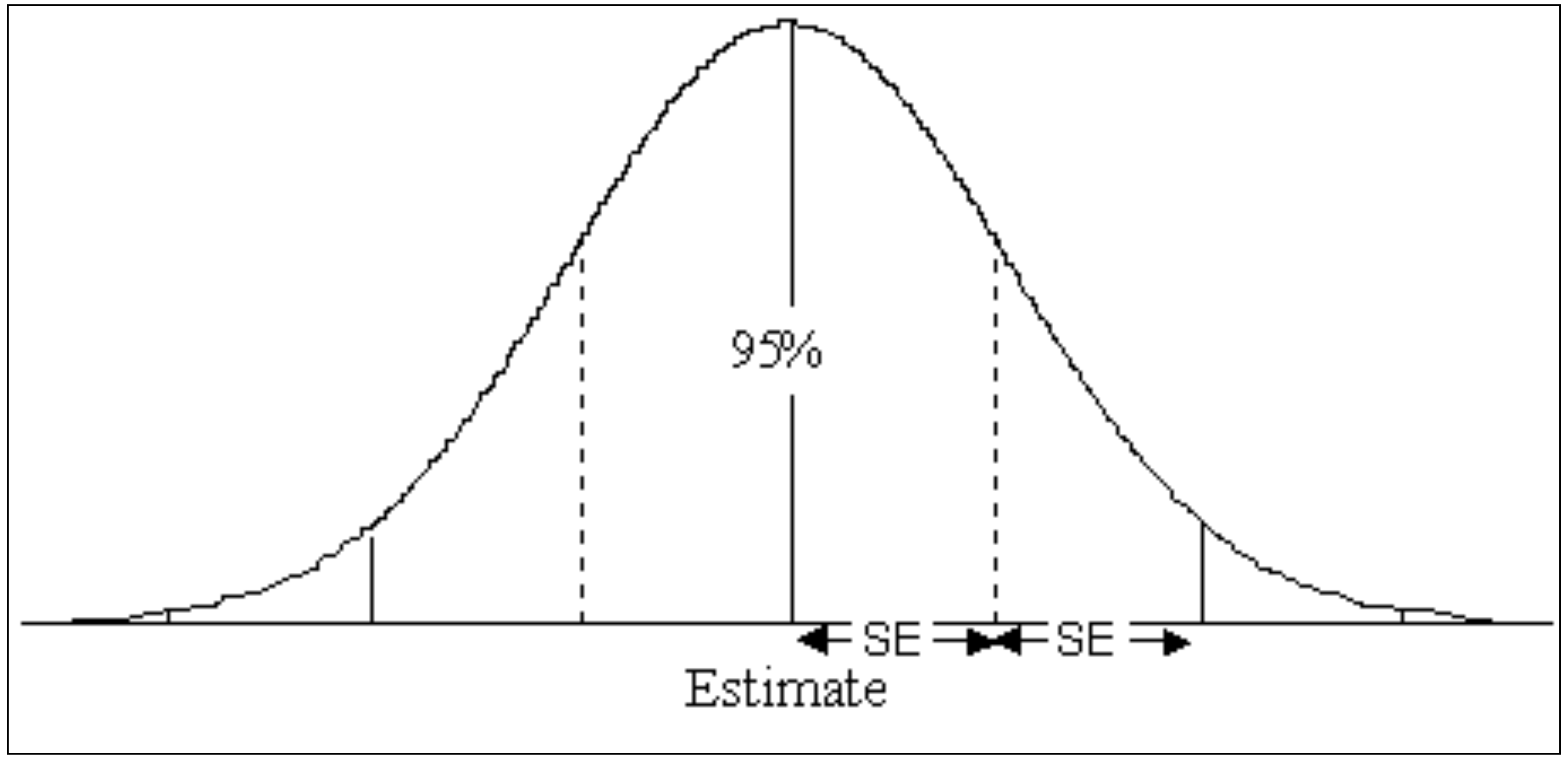
How can we connect our knowledge of probability to better understand polling results?

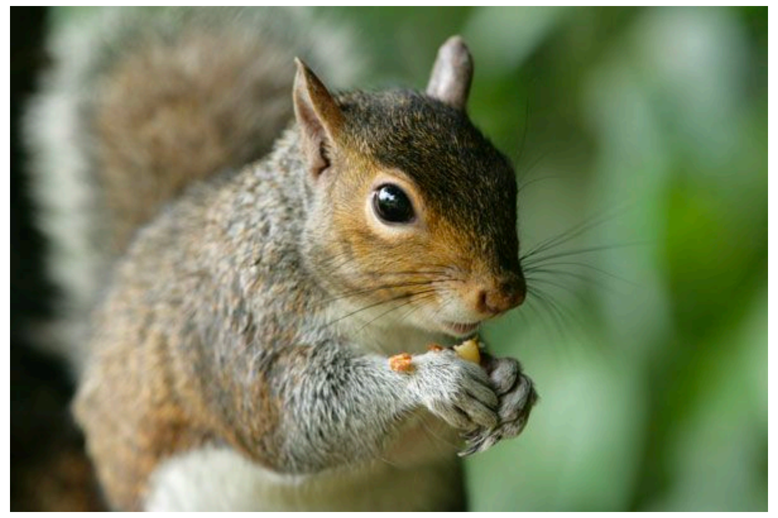
By converting the polling results into measures of confidence that the population mean is within a certain range around the sample mean.

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

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

1 **Confidence intervals** explained





There is a 95% chance that the confidence interval which extends to two standard errors on either side of the estimate contains the "true value".

This interval is called the 95% confidence interval and is the most commonly used confidence interval. The 95% confidence interval is written as follows:

95% confidence interval for outcome y = [y - [2 * se(y)] , y + [2 * se(y)]]

Source: Australian Bureau of Statistics (https://www.abs.gov.au/websitedbs/d3310114.nsf/home/Basic%20Survey%20Design%20-%20Errors%20in%20Statistical%20Data)

1 Confidence intervals

To calculate the confidence interval you need four things:

The number of observations (*n*)
The mean ($\bar{X}$)
The standard deviation (s)
The desired confidence level (let's say 95%) you go to the *Z* table and find the *Z*(0.95) score, which is 1.96.

Then you plug these values into the following equation

$$\bar{x} \pm Z\frac{s}{\sqrt{n}}$$

1 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.

for these purposes. The test statistic for this is known as a *t*-test because it follows the *t*-distribution. The formula for this particular *t*-test is

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\mathrm{se}\left(\bar{Y}_1 - \bar{Y}_2\right)},$$

$$\frac{\text{Difference of means}}{\text{Uncertainty about that difference}}$$

where $\bar{Y}_1$ is the mean of the dependent variable for the first value of the independent variable, $\bar{Y}_2$ is the mean of the dependent variable for the second value of the independent variable, and se(…) is the standard error of the difference between the two means (see below). We can see from this formula that the greater the difference between the mean value of the dependent variable across the two values of the independent variable, the further the value of *t* will be from zero.

In Chapter 7 we introduced the notion of a standard error, which is a measure of uncertainty about a statistical estimate. The basic logic of a standard error is that the larger it is, the more uncertainty (or less confidence) we have in our ability to make precise statements. Similarly, the smaller the standard error, the greater our confidence about our ability to make precise statements about the population.

Source: Kellstedt & Whitten (2018: 176)

1 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.

1 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**.

1

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 <u>independent</u> variable

x hat is the average value of x

y is each individual value (*i*) of the <u>dependent</u> variable

y hat is the average value of y

n is the number of observations

1 Conducting a significance test

ρ (rho) is the correlation coefficient.

Null hypothesis (H_0): $\rho = 0$, there is **not** a significant linear correlation between x and y in the sample.

Alternative hypothesis (H_1): $\rho \neq 0$, there **is** a significant linear correlation between x and y in the sample.

Now we conduct a **Student's T-test**. What is that?

1

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

1

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 pairs of samples (**difference of means test**) or whether two variables are **correlated**?

Rather we want a more complex understanding of the **directionality** and **significance** in the relationship between an X and Y?

Or perhaps we want to **predict** our outcome as we vary values of our independent variable?

1 The bivariate regression model

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

Where:

Y is the <u>outcome</u> you are trying to explain.
X is the main <u>explanatory</u> variable.
α (alpha) is the value of Y when X=0.
β (beta) is the estimated relationship between X and Y.
ϵ is the systematic error.
ε is the random error.

1 Two-tailed hypothesis testing of slope coefficient

Here are my regression results for happiness regressed on GDP: β **= 0.845; se= 0.060**.

My theory's main empirical hypotheses are:

H0 (null hypothesis): β **= 0**

H1 (alternative hypothesis): $\beta \neq$ **0**

To test these hypotheses we do a t-test, in this case we set β_{null} = 0.

$$t = \frac{\beta - \beta null}{se(\hat{\beta})}$$

t **= (0.845-0)/0.06 = 14.083.**

With ~118 degrees of freedom, with a two-tailed test at the 0.05 level the threshold t statistic is 1.984. The estimated ***p-value*** *is 0.000*. I therefore **reject the null hypothesis** in favour of the alternate hypothesis.

1 Confidence intervals

We can estimate confidence intervals using the following equations:

$$\widehat{\beta} +/- [t * se(\widehat{\beta})]$$

$$\widehat{\alpha} +/- [t * se(\widehat{\alpha})]$$

So my **slope's confidence interval** is [0.726, 0.963].

My **intercept's confidence interval** is [-3.627, -1.324].

1 Interpret the **estimated coefficient and standard error**

Regression Statistics	
Multiple R	0.79188047
R Square	0.62707468
Adjusted R Square	0.6239143
Standard Error	0.70291684
Observations	120

ANOVA

	df	*SS*	*MS*	
Regression	1	98.0363889	98.0363889	19
Residual	118	58.3028664	0.49409209	
Total	119	156.339255		

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-2.4701756	0.5790136	-4.2663511	4.0311E-05	-3.616880299	-1.3236709	-3.6168803	-1.3236709
gdp	0.84466562	0.05996462	14.0860655	4.8976E-27	0.725919338	0.9634119	0.72591934	0.9634119

The slope of the regression line (β) is also called the estimated coefficient.

The β and its standard error (α) lets you a hypothesis test using the same t-statistic approach as last week to see if we can conclude that it is statistically significant.

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2},$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}.$$

1 8. Interpret the regression statistics: R-square

Regression Statistics	
Multiple R	0.79188047
R Square	0.62707468
Adjusted R Square	0.6239143
Standard Error	0.70291684
Observations	120

ANOVA

	df	*SS*
Regression	1	98.0363[illegible]
Residual	118	58.302[illegible]
Total	119	156.339[illegible]

	Coefficients	*Standard E[illegible]*						*[illegible]5.0%*
Intercept	-2.4702756	0.57901[illegible]						[illegible]6709
gdp	0.84466562	0.05996[illegible]	14.0[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.9[illegible]4119

The **R-square** or R^2 is the **coefficient of determination**. In other words the proportion of the DV variation accounted for by the model.

$$R^2 = \frac{MSS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})}{\sum_{i=1}^{n}(Y_1 - \bar{Y})^2}$$

1

10. Interpret the **regression statistics:** Standard error

Regression Statistics	
Multiple R	0.79188047
R Square	0.62707468
Adjusted R Square	0.6239143
Standard Error	0.70291684
Observations	120

The regression's **standard error** is average distance that the observed values fall from the regression line.

The better the regression fit the smaller this value will be.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	98.0363889	98.036388	[illegible]	[illegible]
Residual	118	58.3028664	0.494092		
Total	119	156.339255			

$$Regression\ standard\ error = \frac{\sum_{i=1}^{n}(Y_1 - \hat{Y}_i)}{n}$$

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-2.4702756	0.57901366	-4.2663511	4.0311E-05	-3.616880299	-1.3236709	-3.6168803	-1.3236709
gdp	0.84466562	0.05996462	14.0860655	4.8976E-27	0.725919338	0.9634119	0.72591934	0.9634119

1

11. Interpret the regression statistics: F statistic

Regression Statistics	
Multiple R	0.79188047
R Square	0.62707468
Adjusted R Square	0.6239143
Standard Error	0.70291684
Observations	120

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	98.0363889	98.0363889	198.417241	4.89759E-27
Residual	118	58.3028664	0.49409209		
Total	119	156.339255			

	Coefficients	*tandard Erroi*	*t Stat*	*I*	*5.0%*	*Upper 95.0%*
Intercept	-2.4702756	0.57901366	-4.2663511	4	8803	-1.3236709
gdp	0.84466562	0.05996462	14.0860655	4	1934	0.9634119

The regression's **F statistic** is a measure of the regression's overall significance measured using analysis of variance (ANOVA).

With the F statistic, you can do a statistical significance test using the F-distribution for 1 and n-2 degrees of freedom

For a two-variable regression:

$$F = \frac{(\widehat{\beta_1} \sum y_i x_{1i})}{\sum \widehat{u_i^2}/(n-2)}$$

Equation source: Gujarati (2003: 140)

1 **Multivariate** regression model

$$Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + u_1$$

Where:

Y is the outcome you are trying to explain.
X is the main explanatory variable.
Z is an additional explanatory/control variable
α (alpha) is the value of Y when X=0 & Z=0.
β_1 (beta) is the estimated effect of X on Y holding constant the effects of Z.
β_2 (beta) is the estimated effect of Z on Y holding constant the effects of X.
u = population error term/residual

1 Estimating the relationship between X and Y, controlling for Z

Y=Happiness; **X**=GDP; **Z**=Freedom

Bivariate: $Y_i = \alpha + \beta X_i$

= -2.47 + 0.85**X**

$\widehat{Y_{Australia}}$= -2.47 + 0.85(10.82) = <u>7.27</u> (actual value is 7.11)

Multivariate: $Y_i = \alpha + \beta_1 X_i + \beta_2 Z_i + u_1$

$\widehat{Y_i}$ = -4.19 + 0.72X + 3.74Z

$\widehat{Y_{Australia}}$= -4.19 + 0.72(10.82) + 3.74(0.91) = <u>7.38</u> (actual value is 7.11)

All intercepts and slope coefficients are statistically significant at the 0.001 level.

1 Interpreting a regression table

Tell your readers in words **what you want them to take away** from your table.

Often focus is on both **statistical** and **substantive** significance.

Connect results back to your **theory** and **hypotheses**.

Research pitfalls

#1: Correlation does not equal causation.
#2: Spurious/omitted variable problem
#3: Endogeneity
#4: Multicollinearity
#5: Transforming (or leaving) variables
#6: Stepwise regression
#7: Data mining/garbage-can regressions/overfitting
#8: Dichotomous or categorical dependent variables
#9: Time series vs. cross-sectional sample?
#10: Simpson's Paradox (trend can be eliminated/ reversed by splitting data into groups.
#11: Overlooking cross-validation
#12: Extrapolating beyond the data you have
#13: Using a regression on a non-linear relationship
#14: Publication bias
#15: Theoretical biases (Confirmation bias, interpretation bias, fundamental attribution error)
#16: Empirical biases

2 Important terms

Image source: https://newsable.asianetnews.com/world/world-celebrates-international-pillow-fight-day

2

Learning terms this semester

A modest proposal…

1. Write down the definition.
2. Describe why scholars think this term is useful.
3. Think of an example that resonates with you.

2

Weeks 1 & 2

Causality
Correlation
Data
Dependent variable
Generality
Hypothesis/null hypothesis
Hypothesis testing
Independent variable
Measure
Parsimony
Theory
Variable

Complete & incomplete information
Expected utility
Formal theory
Transitive & intransitive preferences
Preference ordering
Rational choice
Rational utility maximisers
Spatial & temporal dimensions
Utility
Bivariate vs multivariate
Confounding variable
Deterministic vs. probabilistic relationship
Endogeneity
Spuriousness

2

Weeks 3 & 4

Experimental research design
Observational research design
Treatment and control groups
Placebo
Survey experiment
Field experiment
Natural experiment
Internal and external validity
Convenience sample
Replication
Datum and data
Cross-sectional and time-series data

Conceptual clarity
Reliability
Measurement bias
Face validity
Content validity
Construct validity

2

Weeks 5 & 6

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

Population
Sample
Representative sample
Random sample
Systematic random sample
Cluster or multistage sampling
Stratified random sample
Selection bias
Nonprobability sample
Convenience sample
Volunteer sample
Purposive sample
Snowball sample
Sampling error
Confidence interval
Variance
List experiment

2

Week's 7 & 8

Sample mean
Standard deviation
Standard error of the mean
Confidence interval
Lower & upper bounds

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

2 Weeks 9 & 10

Directional and non-directional hypotheses
Parameters
Parameter estimate
Population error term
Residual
Model standard error
R-squared
Stochastic
T-ratio
T-test
Slope
Intercept
Ceterus paribus
Autocorrelation

Bias
Omitted variable bias
Perfect multicollinearity
Substantive significance
Vector
Matrix

2

Week 11

Confirmation bias
Data mining
Dummy variable
Extrapolation/interpolation
Fundamental attribution error
Index/indices
Interactive effect
Interactive model

Interpretation bias
Leave-one-out cross-validation
Limited dependent variable
Multicollinearity
Publication bias
Stepwise regression
Transformed variable

3 Final exam

Image source: https://www.fxhash.xyz/gentk/146004

3 Final exam

Course Codes

POLS2044

You can search for multiple courses by putting a comma between your course codes.

SEARCH : FINAL TIMETABLE

Displaying records 1 to 3 of 3

Exam Code	Exam Title	Assessment Type	Date	Time	Writing Time (minutes)	Reading Time (minutes)	Venue	Building	Room
POLS2044_Semester 2	Contemporary Political Analysis	Normal	Wednesday 13/11/2024	2:00pm	120	15	Copland G29	24	G29
POLS2044_Semester 2	Contemporary Political Analysis	Normal	Wednesday 13/11/2024	2:00pm	120	15	Copland G30	24	G30
POLS2044_Semester 2	Contemporary Political Analysis	Normal	Wednesday 13/11/2024	2:00pm	120	15	Copland G31	24	G31

Displaying records 1 to 3 of 3

3 Final exam

POLS2044
Contemporary Political Analysis

2023 FINAL EXAM

Reading time: 15 minutes
Writing time: 120 minutes

This exam is an opportunity for you to demonstrate your ability to identify major elements of contemporary political analysis, explain why they matter, and apply them. The exam is closed book and taken in person. Please be sure to check your answers and make sure your full name is on your script book.

This exam has three sections that are broken down as follows:

10 multiple choice questions (20%),
10 shorter answer questions (50%), and
2 longer answer questions (30%).

Please be sure to answer all 22 questions, write all your answers in the script book (not on this exam), and leave time to complete all the questions. Please write as legibly as possible (we cannot mark what we cannot read/decipher). Please write the multiple-choice answers on the first page of your script book in the following format:

1. a
2. b
3. c
4. d
5. a
6. b
7. c
8. d
9. a
10. b

I have really enjoyed teaching this class and engaging in our workshop discussions. I hope you found this term interesting and useful. Thank you for all your hard work!   -Richard

3 Final exam

MULTIPLE CHOICE
(20% total, each question is worth 2% of your final mark)

Please answer the following ten multiple-choice questions. When reading the questions please be sure to read them carefully and answer the question asked.

SHORT(ER) ANSWER
(50% total, each question is worth 5% of your final mark)

In this section, please answer the following ten questions, making sure to answer all parts of the question. These questions are designed to take two to five sentences to answer adequately.

LONG(ER) ANSWER
(30%, each question is worth 15% of your final mark)

Please answer the following two questions using full sentences, complete paragraphs, and clearly structured answers.


Thank you for a great semester!