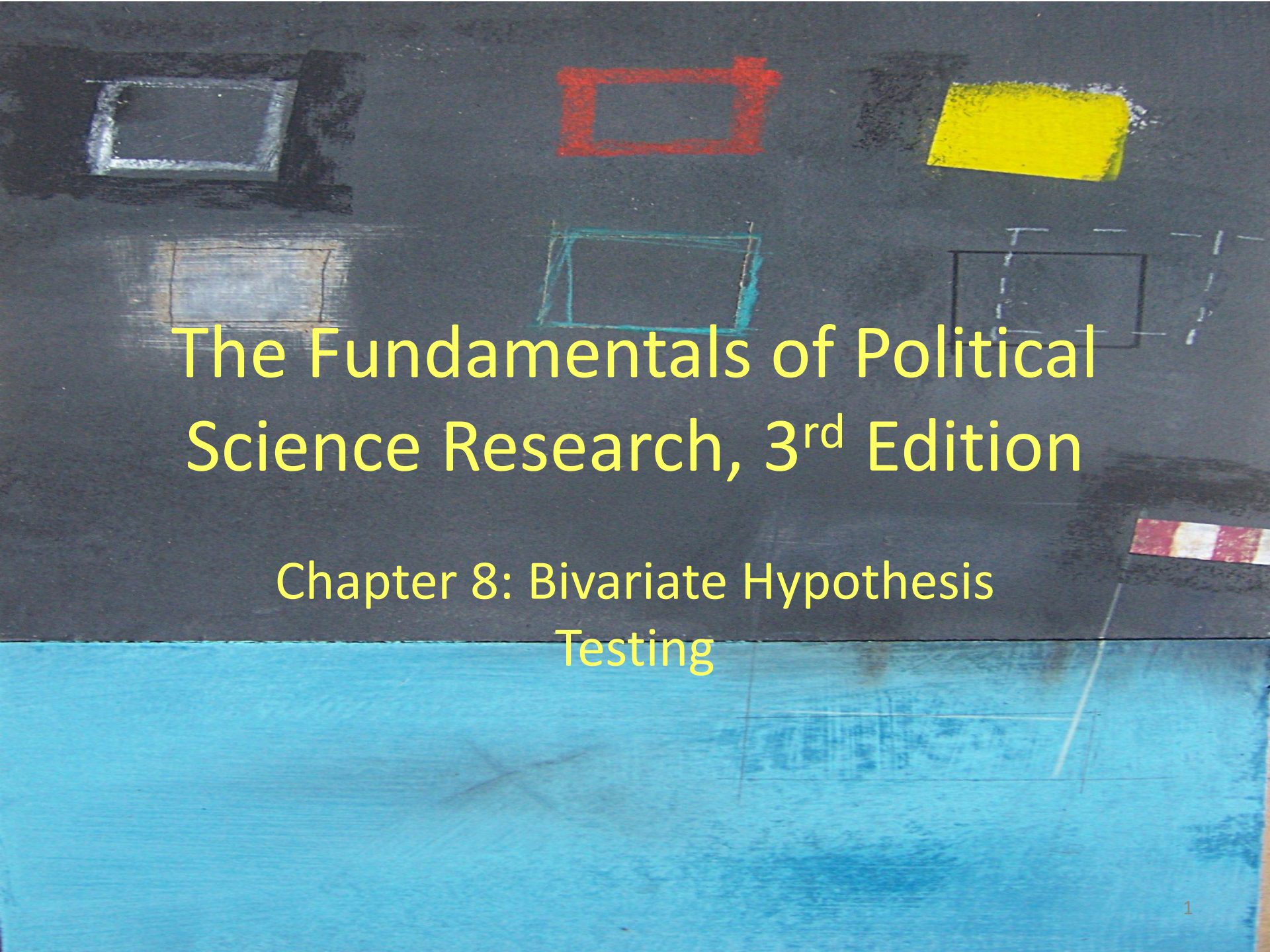
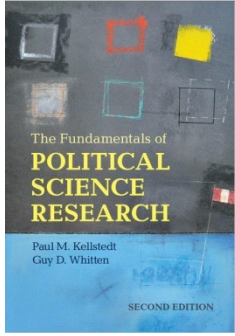


The Fundamentals of Political Science Research, 3rd Edition

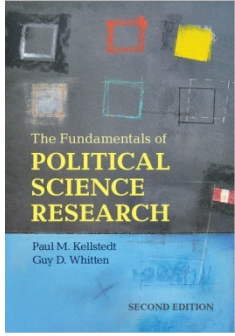
Chapter 8: Bivariate Hypothesis Testing

Chapter 8 Outline

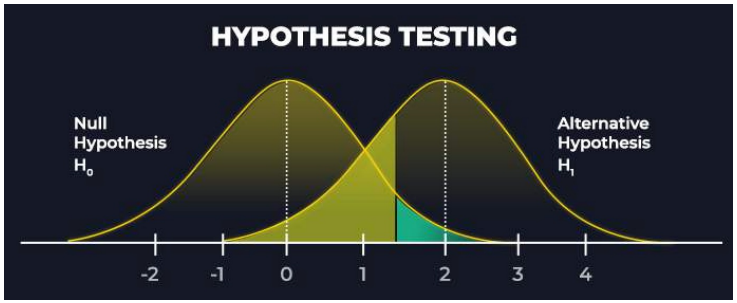


- Bivariate Hypothesis Tests and Establishing Causal Relationships
- All Roads Lead to p (the p-value)
- Three Types of Bivariate Hypothesis Tests

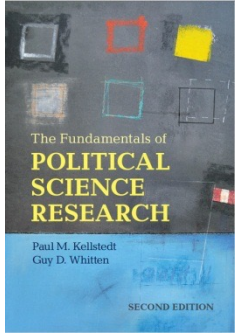
Bivariate Hypothesis Tests and Establishing Causal Relationships



- Bivariate hypothesis testing is less used now
- Omitted variables: It cannot help us with the important question, "Have we *controlled for* all confounding variables Z that might make the association between X and Y spurious?
  - But it can be pretty useful in experiments
  - Help us answer the question, "Are X and Y related?"
  - Lay the foundation for Multivariate hypothesis testing



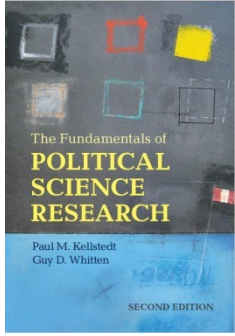
Choosing the Right Bivariate Hypothesis Test



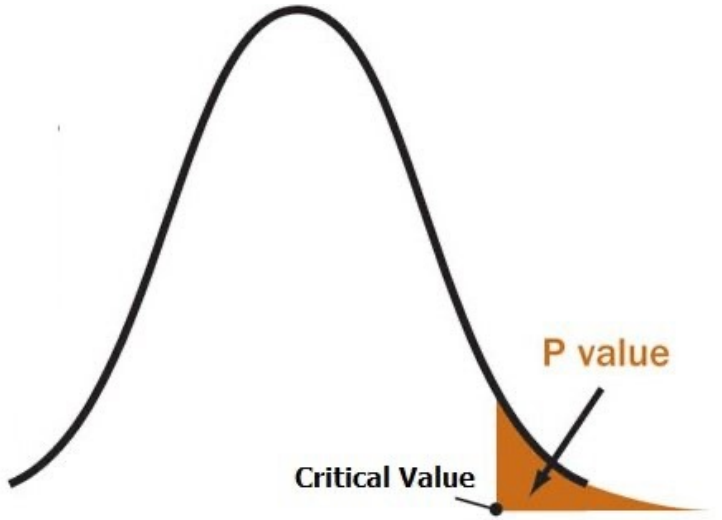
Outcome		Treatment: Independent variable type: Categorical	Treatment: Independent variable type: Continuous
Dependent variable type	Categorical	*Tabular analysis*	Probit/logit (Ch.11)
	Continuous	*Difference of means*	*Correlation coefficient*; bivariate regression model (Ch.8)

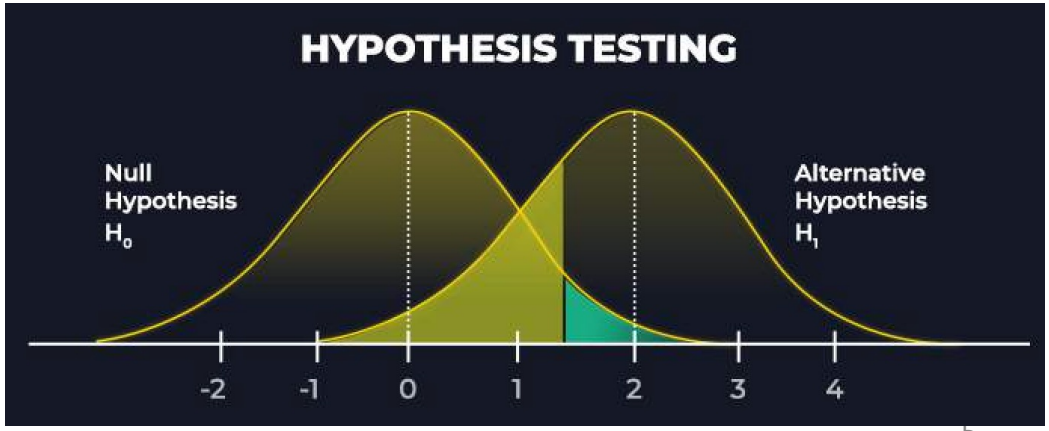
Note: Tests in italics are discussed in this chapter.

All Roads Lead to p

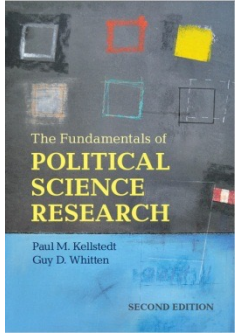


- One common element across a wide range of statistical hypothesis tests is the p-value.
- This value, ranging between 0 and 1, is the closest thing that we have to a bottom line in statistics.
- But it is often misunderstood and misused.



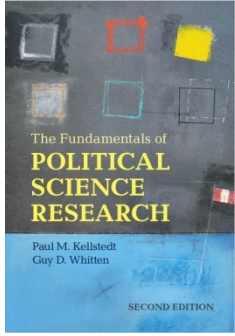


The logic of p-values

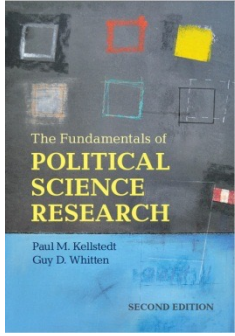


- The third hurdle: "Is there covariation between X and Y?"
  - We compare the actual relationship between X and Y *in sample data* with what we would *expect to find if X and Y were not related* in the underlying population.
- So what's P-value?
  - The null hypothesis: no relationship
  - A threshold we use to reject the null
- → The *more different* the empirically observed relationship is from what we would expect to find if there were *not* a relationship, the more confidence we have that X and Y are related in the population.
  - Basically, we want our H1 as different as possible to H0
  - Null hypothesis (H0): there is no relationship
  - Alternative hypothesis (H1): there is a relationship
- The statistic that is most commonly associated with this type of logical exercise is the p-value.

The logic of p-values

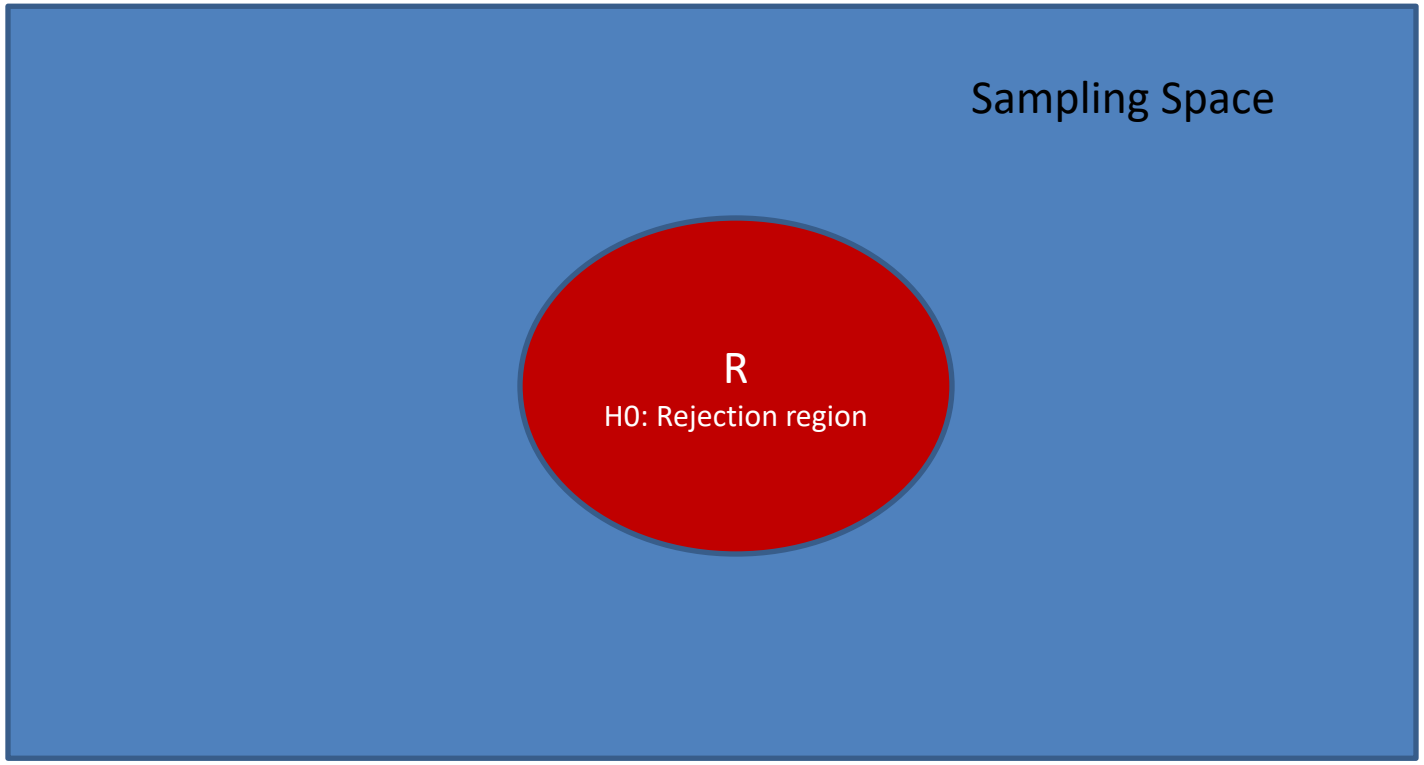


- P-value and the *confidence* level:
  - P-value, or probability value, tells you how likely it is that your data could have occurred under the null hypothesis. It does this by calculating the likelihood of your test statistic, which is the number calculated by a statistical test using your data
  - Thus, *the lower the p-value, the greater confidence we have that there is a "systematic relationship" between the two variables* for which we estimated the particular p-value.
  - E.g. when $p < 0.01$, it means that the chance that we don't see a relationship is very small
  - when $p < 0.001$, it means that the chance that we don't see a relationship is even smaller

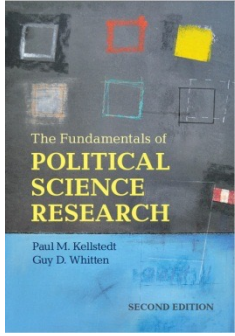


The logic of p-values

P-value defines the size of your H0 rejection region



Your point estimate $\boldsymbol{\theta}$

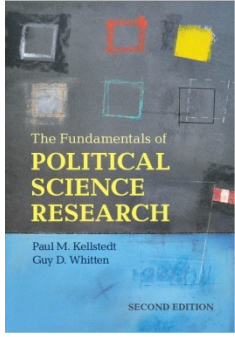


The logic of p-values

P-value defines the size of your H0 rejection region

Your point estimate $\boldsymbol{\theta}$

The limitations of p-values

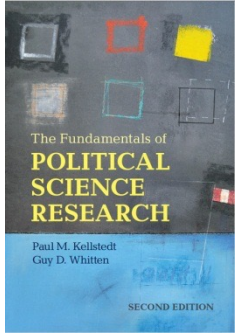


- The logic of a p-value is not reversible. In other words, p= .001 does not mean that there is a .999 chance that *something systematic is going on*.
  - The rejection region is calculated and defined by test statistics and its distribution

Caveats:

- 1. Although a p-value tells us something about our confidence that there is a relationship between two variables, it does not tell us whether that relationship is "*causal*".
  - Strong causal relationship requires good identification strategies
- 2. When a p-value is very close to zero, this does not indicate that the relationship between X and Y is very "*strong*".
  - Statistically significant at some level
  - Confidence level: Larger sample size → more confident (smaller p value)
- 3. A p-value does not directly reflect the quality of the measurement procedure for our variables

The limitations of p-values

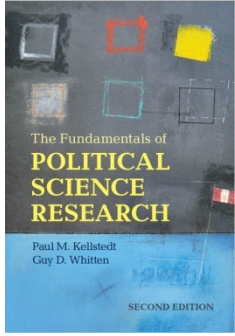


- 4. random sample assumption: p-values are always based on the assumption that you are drawing a "perfectly random sample from the underlying population."
  - Mathematically, this is expressed as

$$p_i = P \forall i$$

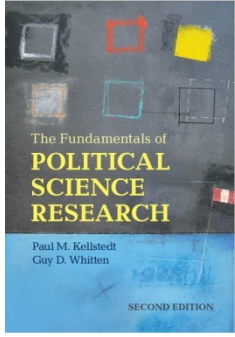
  - This translates into "the probability of an individual case from our population ending up in our sample, p_i, is assumed to equal P for all of the individual cases i."
  - If this assumption were valid, we would have a truly random sample.
  - The further we are from a truly random sample, the less confidence we should have in our p-value
  - However, more data indeed can increase p-value (and our confidence)

From p-values to statistical significance



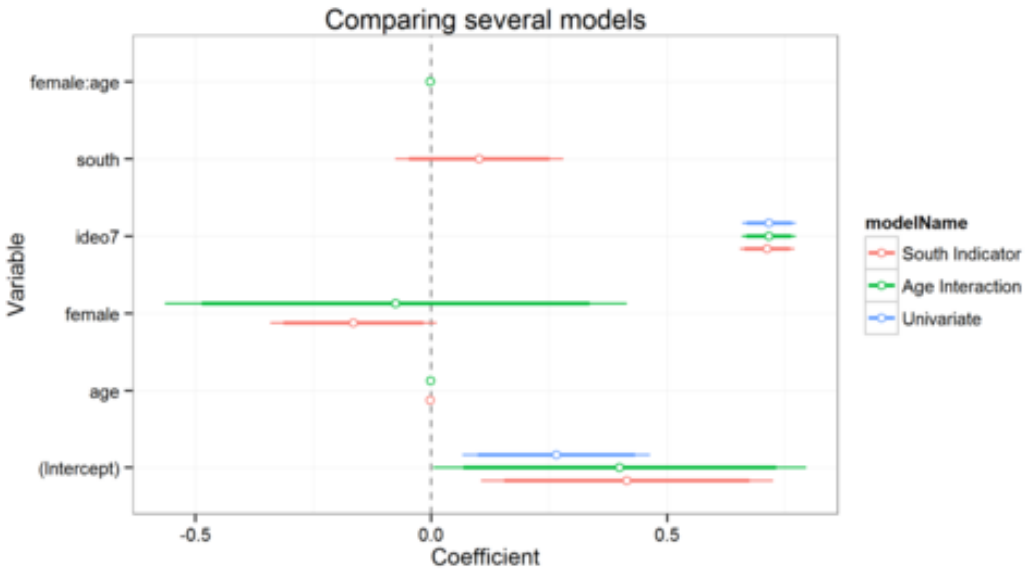
- The language to express found relationship:
  - A common way of referring to such a situation is to state that the relationship between the two variables is "statistically significant."
- An assertion of statistical significance depends on a number of other factors:
  - "Statistical significance" is achieved only to the extent that the **assumptions** underlying the calculation of the p-value hold.
  - There are a variety of different standards for what is a statistically significant p-value. **Most social scientists use the standard of a p-value of .05 (or 95% confidence interval).**
  - Avoid using a cut-point like $p<.05$; but use all thresholds ($p<.10$, $p<0.05$, $p<0.01$)
  - You will need to use these terminology when reporting your result

How do we show P-values

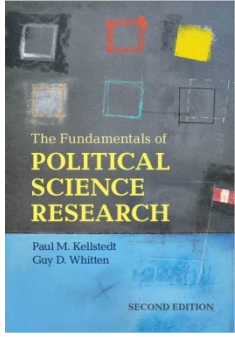


DV	More than 15yr sentence		
	(1)	(2)	(3)
New Law of Military Trial	-0.287***	-0.197**	-0.261***
Bias-corrected 95% CI	[-0.450, -0.123]	[-0.354, -0.040]	[-0.401, -0.121]
Robust 95% CI	[-0.480, -0.092]	[-0.380, -0.014]	[-0.419, -0.102]
Bandwidth (days)	721	1655	727
Effective N	435	2016	454
Polynomial order	Linear	Quadratic	Linear
Weight	Triangular	Triangular	Uniform

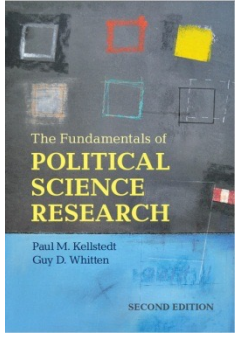
Note: Confidence interval (CI) clustered at the case level. Mean square error optimal bandwidth. Covariates of age, male, islander, (log) number of codefendants, weapon, committing subversion, leaking military intelligence, and president review again (rate) in $t - 1$ are included. **p<0.05; ***p<0.01.



Three Bivariate Hypothesis Tests



- So we've learned the null hypothesis, p-value (rejection region), confidence level
- We now turn to three different bivariate hypothesis tests:
  - 1. Tabular analysis
  - 2. Difference of means
  - 3. Correlation coefficient

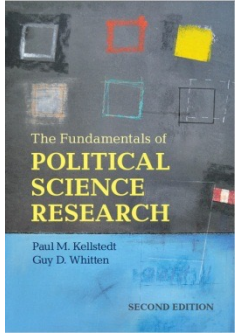


Example 1: Gender and vote in the 2008 U.S. presidential election (a hypothetical scenario)

Candidate	Male	Female	Row total
McCain	?	?	45.0
Obama	?	?	55.0
Column total	100.0	100.0	100.0

Note: Cell entries are column percentages.

- How to quantify this test?
  - H0:
  - H1:



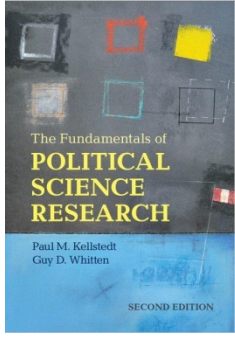
Example 1: Gender and vote in the 2008 U.S. presidential election (a hypothetical scenario)

Candidate	Male	Female	Row total
McCain	?	?	45.0
Obama	?	?	55.0
Column total	100.0	100.0	100.0

Note: Cell entries are column percentages.

- How to quantify this test?
  - H0: there is "no" relationship between gender and voting
  - H1: there is a relationship between gender and voting (this is our theoretical argument)

Gender and vote in the 2008 U.S. presidential election: Expectations for hypothetical scenario if there were no relationship



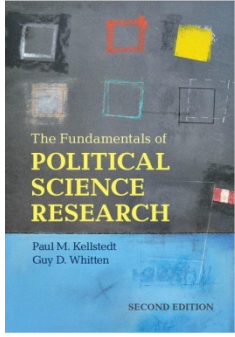
- Example: H0

Candidate	Male	Female	Row total
McCain	45.0	45.0	45.0
Obama	55.0	55.0	55.0
Column total	100.0	100.0	100.0

Note: Cell entries are column percentages.

- If there is no relationship, we should observe an identical voting pattern between men and women
- → This becomes our baseline comparison.

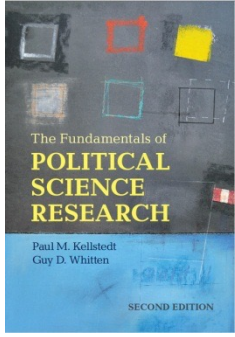
Gender and vote in the 2008 U.S. presidential election



Candidate	Male	Female	Row total
McCain	?	?	1,434
Obama	?	?	1,755
Column total	1,379	1,810	3,189

Note: Cell entries are number of respondents.

Gender and vote in the 2008 U.S. presidential election: Calculating the expected cell values if gender and presidential vote are unrelated

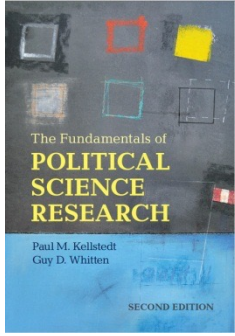


- What a H0 should look like (the comparison set)

Candidate	Male	Female
McCain	(45% of 1,379) $= 0.45 \times 1,379 = 620.55$	(45% of 1,810) $= 0.45 \times 1,810 = 814.5$
Obama	(55% of 1,379) $= 0.55 \times 1,379 = 758.45$	(55% of 1,810) $= 0.55 \times 1,810 = 995.5$

Note: Cell entries are expectation calculations if these two variables are unrelated.

Gender and vote in the 2008 U.S. presidential election

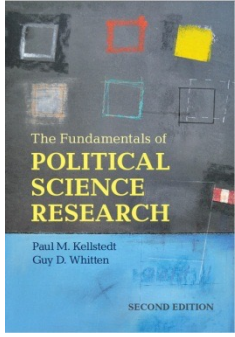


- The observed data:

Candidate	Male	Female	Row total
McCain	682	752	1,434
Obama	697	1,058	1,755
Column total	1,379	1,810	3,189

Note: Cell entries are number of respondents.

Gender and vote in the 2008 U.S. presidential election



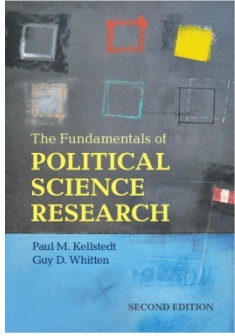
Candidate	Male	Female
McCain	$O = 682; E = 620.55$	$O = 752; E = 814.5$
Obama	$O = 697; E = 758.45$	$O = 1,058; E = 995.5$

Note: Cell entries are the number observed (O); the number expected if there were no relationship (E).

Democratic party: Male (-) ; Female (+)

- Among males, the proportion of observed voting for Obama is lower than what we would expect if there were no relationship between the two variables. Also, among males, the proportion voting for McCain is higher than what we would expect if there were no relationship.
- But for females, this pattern is reversed.
- The pattern of these differences is in line with the theory that ***women support Democratic Party candidates more than men do***.
- To assess whether or not these differences are "statistically significant", we turn to the chi-squared (χ^2) test for tabular association.

The χ^2 test for tabular association

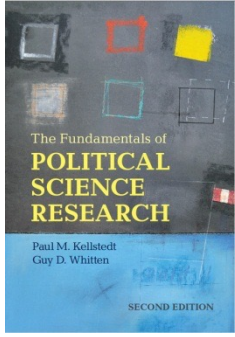


- The formula for the χ^2 statistic is (how different is it between O and E)

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

- The summation sign in this formula signifies that we sum over each cell in the table; so a 2x2 table would have "four cells to add up".
- If we think about an individual cell's contribution to this formula, we can see the underlying logic of the χ^2 test.
  - If all observed values were exactly equal to the values that we expect <u>if there were no relationship between the two variables, then $\chi^2 = 0$</u>.
  - **The more the O values differ from the E values, the greater the value will be for χ^2.**
  - Positive value: Because the numerator on the right-hand side of the χ^2 formula (O - E) is squared, any difference between O and E will contribute positively to the overall χ^2 value.

$\chi 2$ calculation



- Here are the calculations for χ^2 for our gender and voting example:

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

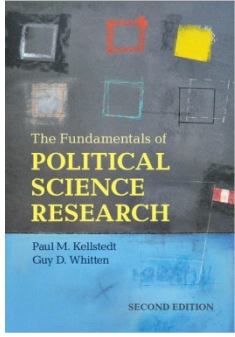
$$= \frac{(682-620.55)^2}{620.55} + \frac{(752-814.5)^2}{814.5} + \frac{(697-758.45)^2}{758.45} + \frac{(1,058-995.5)^2}{995.5}$$

$$= \frac{3,776.1}{620.55} + \frac{3,906.25}{814.5} + \frac{3,776.1}{758.45} + \frac{3906.25}{995.5}$$

$$= 6.09 + 4.8 + 4.98 + 3.92 = 19.79.$$

- Critical value of χ^2 (in R or Table): We need to compare that 19.79 with some predetermined standard, called a "critical value", of χ^2.
- If our calculated value is greater than the critical value, then we conclude that there is a relationship between the two variables; and if the calculated value is less than the critical value, we cannot make such a conclusion.
- D.F.: To make this evaluation, we need a piece of information known as the "degrees of freedom" (df's) for our test. df = (r-1)(c-1), where r is the number of rows in the table, and c is the number of columns in the table. In our table there are two rows and two columns, so (2-1)(2-1) = 1.
- Look at the table in Appendix: Calculate critical values and the level of significance (rejection region) based on the χ^2 distribution

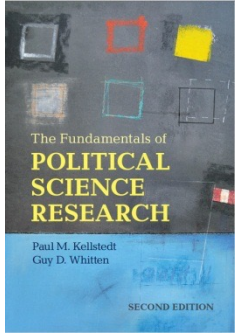
χ2 calculation



	Level of significance				
df	0.10	0.05	0.025	0.01	0.001
1	2.706	3.841	5.024	6.635	10.828
2	4.605	5.991	7.378	9.210	13.816
3	6.251	7.815	9.348	11.345	16.266

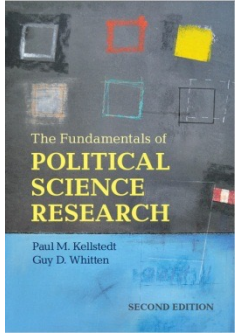
- You can find a table with critical values of χ^2 in Appendix A.
- If we adopt the standard p-value of .05, we see that the critical value of χ^2 for df = 1 is 3.841.
- A calculated χ^2 value of 19.79 is well over the minimum value needed to achieve a p-value of .05.
- At this point, we have established that the relationship between our two variables meets a conventionally accepted standard of statistical significance (i.e., $p < .05$).

From χ^2 to statistical significance



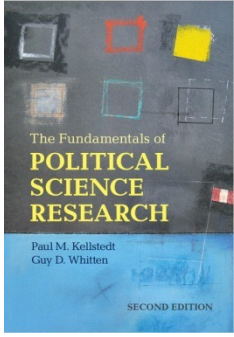
- Remember: Although this result is supportive of our hypothesis, we have not yet established a "causal relationship" between gender and presidential voting.
- With a bivariate analysis, we cannot know whether some other variable Z is relevant because, by definition, there are only two variables in such an analysis. So, until we see evidence that Z variables have been controlled for, our scorecard for this causal claim is [y y y n]. → *mechanisms (v), reversed causality (v), covariation (v)* etc.

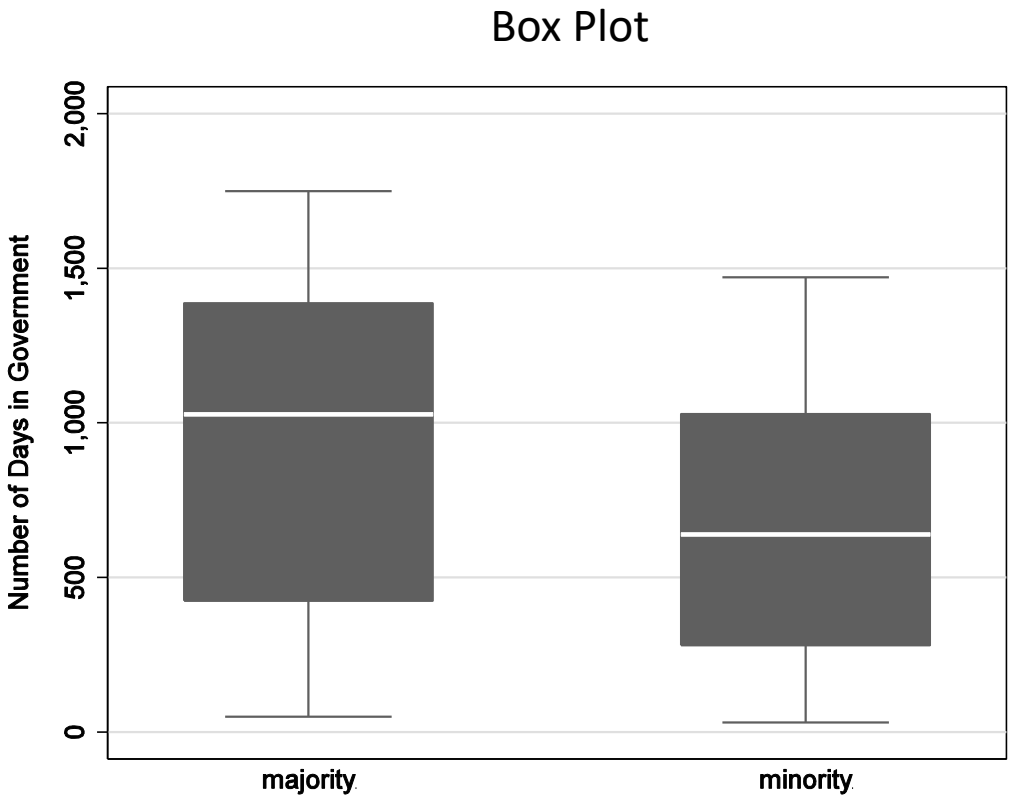
Example 2: Difference of means



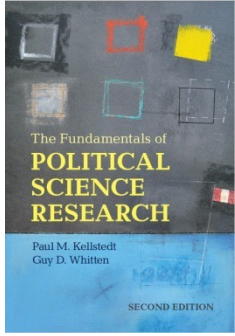
- In our second example, we examine a situation in which we have a "continuous dependent variable" and a categorical independent variable.
- Difference of mean test: In this type of bivariate hypothesis test, we are looking to see if the means are different across the values of the independent variable.
  - Treatment: binary
  - Outcome: continuous
- We use the sample means and standard deviations to make inferences about the unobserved population.
- With continuous variables, it is useful to *graph our data* to get an initial look at what is going on. Two helpful graphs for this type of situation are a box-whisker plot and a kernel *density plot.*

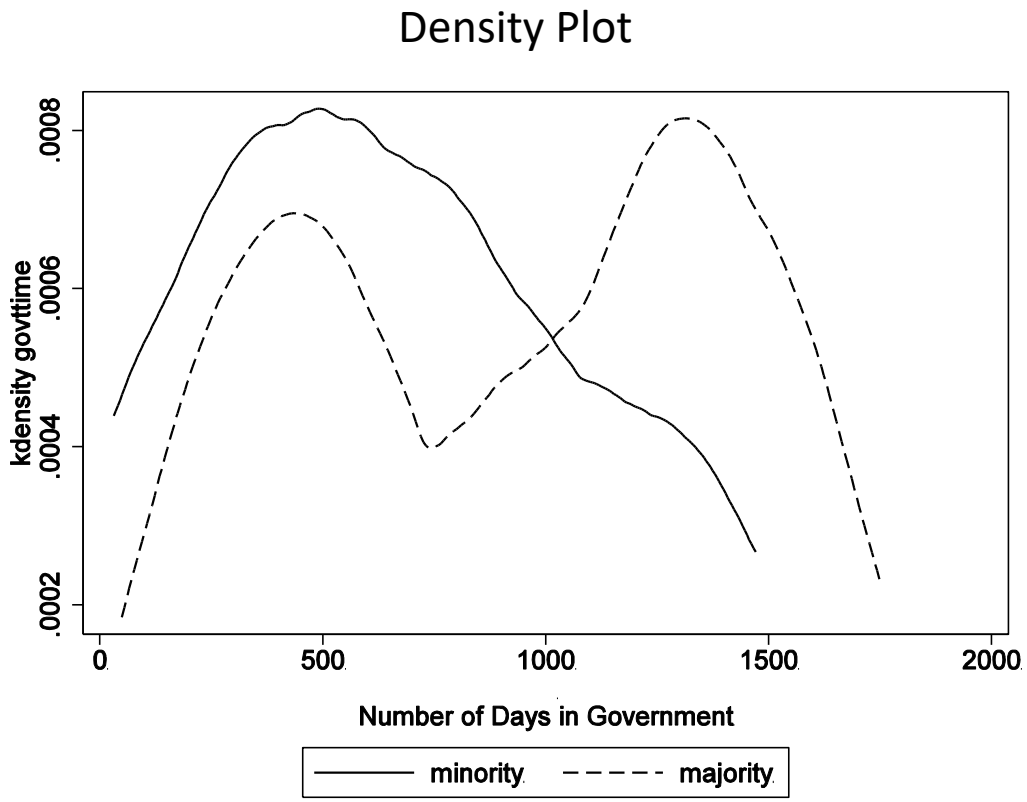
Majority/minority governments → Gov duration



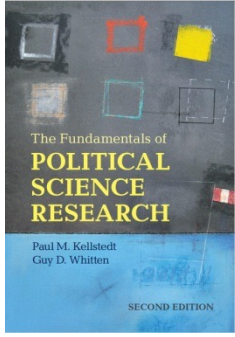


Kernel density plot of Government Duration for majority and minority governments



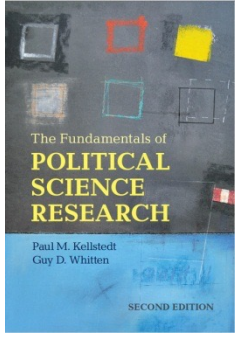


Difference of means test



- From both of these plots, it appears that majority governments last longer than minority governments.
- Statistically significant: we turn to a difference of means test.
- H0: In this test we compare what we have seen in the two figures with what we would expect if there were "no" relationship between Government Type and Government Duration.
- If there were no relationship between these two variables, then the world would be such that *the duration of governments of both types were drawn from the same underlying distribution*. If this were the case
  - 1. the mean or average value of Government Duration would be the same for minority and majority governments.
  - 2. The standard deviation would be the similar.

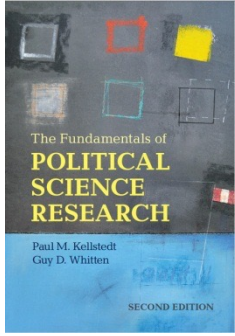
Difference of means test



- One possible test of this null hypothesis is a t-test (because it follows the t-distribution).
- The formula for this particular t-test is

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

- where $\bar{Y}_1$ is the mean of the dependent variable for the first value of the independent variable and $\bar{Y}_2$ is the mean of the dependent variable for the second value of the independent variable.
- 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.
- The further apart the two means are and the less dispersed the distributions (as measured by the standard deviations s_1 and s_2), the greater confidence we have that $\bar{Y}_1$ and $\bar{Y}_2$ are different from each other.



Government type and government duration

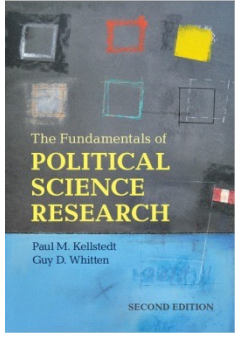
Government type	Number of observations	Mean duration	Standard deviation
Majority	124	930.5	466.1
Minority	53	674.4	421.4
Combined	177	853.8	467.1

- From the values displayed in this table we can calculate the t-test statistic for our hypothesis test. The standard error of the difference between two means ($\bar{Y}_1$ and $\bar{Y}_2$), se ($\bar{Y}_1$ - $\bar{Y}_2$), is calculated from the following formula:

$$\text{se}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\left(\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}\right)} \times \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

- where n_1 and n_2 are the sample sizes, and ${s_1}^2$ and ${s_2}^2$ are the sample variances.

Difference of means test



- If we label the number of days in government for majority governments $\bar{Y}_1$ and the number of days in government for minority governments $\bar{Y}_2$, then we can calculate the standard error as

$$\text{se}(\bar{Y}_1 - \bar{Y}_2) = \sqrt{\left(\frac{(124-1)(466.1)^2 + (53-1)(421.4)^2}{124+77-2}\right)} \times \sqrt{\left(\frac{1}{124} + \frac{1}{53}\right)}$$

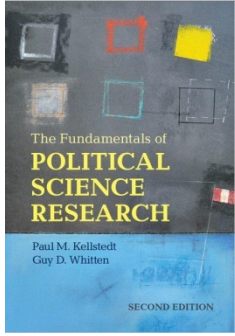
$$\text{se}(\bar{Y}_1 - \bar{Y}_2) = 74.39.$$

- Now that we have the standard error, we can calculate the t-statistic:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2}{\text{se}(\bar{Y}_1 - \bar{Y}_2)} = \frac{930.5 - 674.4}{74.39} = \frac{256.1}{74.39} = 3.44.$$

- Now that we have calculated this t-statistic, we need one more piece of information before we can get to our p-value. This is called the "degrees of freedom" (df's).

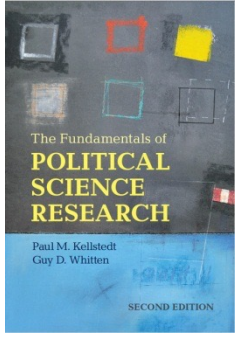
Difference of mean test—degrees of freedom



- Degrees of freedom reflect the basic idea that we will gain confidence in an observed pattern as the amount of data on which that pattern is based increases. → More data, greater df
- If we turn to Appendix B, which is a table of critical values for t, we can see that it reflects this logic.
- This table also follows the same basic logic as the χ^2 table:
  - The way to read such a table is that the columns are defined by targeted p-values, and, to achieve a particular target p-value, you need to obtain a particular value of t.
  - The rows in the t-table indicate the number of degrees of freedom.
  - As the number of degrees of freedom goes up, the t-statistic we need to obtain a particular p-value goes down. → easier to achieve the conf. level
- We calculate the degrees of freedom for a difference of means t-statistic based on the sum of total sample size minus two. Thus our degrees of freedom is

$$n_1 + n_2 - 2 = 124 + 53 - 2 = 175.$$

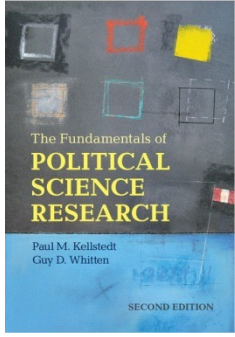
Difference of means test—p-value



- From the p-value, we can look across the row for which df = 100 and see the minimum t-value needed to achieve each targeted value of p.
- T = 3.44, df = 175 (sig. at 95% conf. int.)

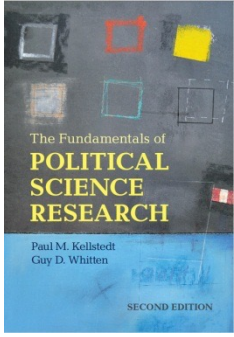
	Level of significance					
df	**0.10**	**0.05**	**0.025**	**0.01**	**.005**	**0.001**
100	1.290	1.660	1.984	2.364	2.626	3.174

Example 3: Correlation Coefficient

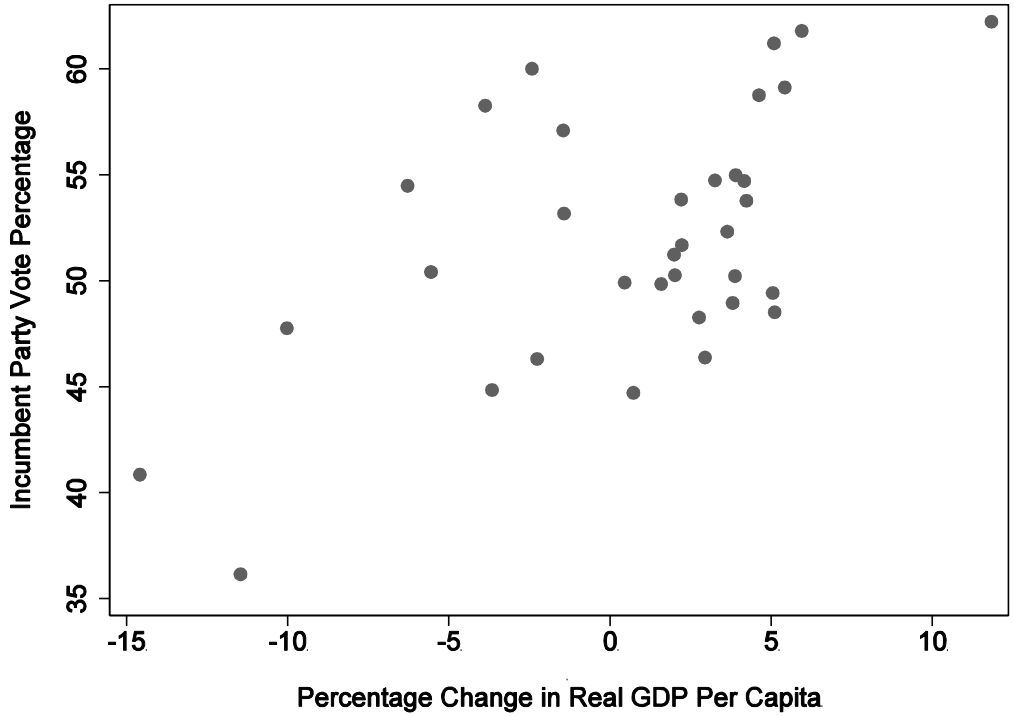


- In our final example of bivariate hypothesis testing we look at a situation in which both the independent variable and the dependent variable are continuous.
  - γ: Pearson's correlation coefficient
  - Treatment: continuous
  - Outcome: continuous
- Scatter plots are useful for getting an initial look at the relationship between two continuous variables:

Scatter plot of change in GDP and incumbent-party vote share

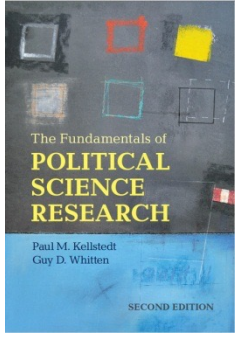


Outcome → Y



X

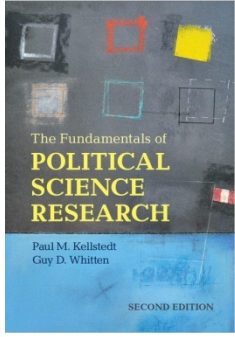
Covariance



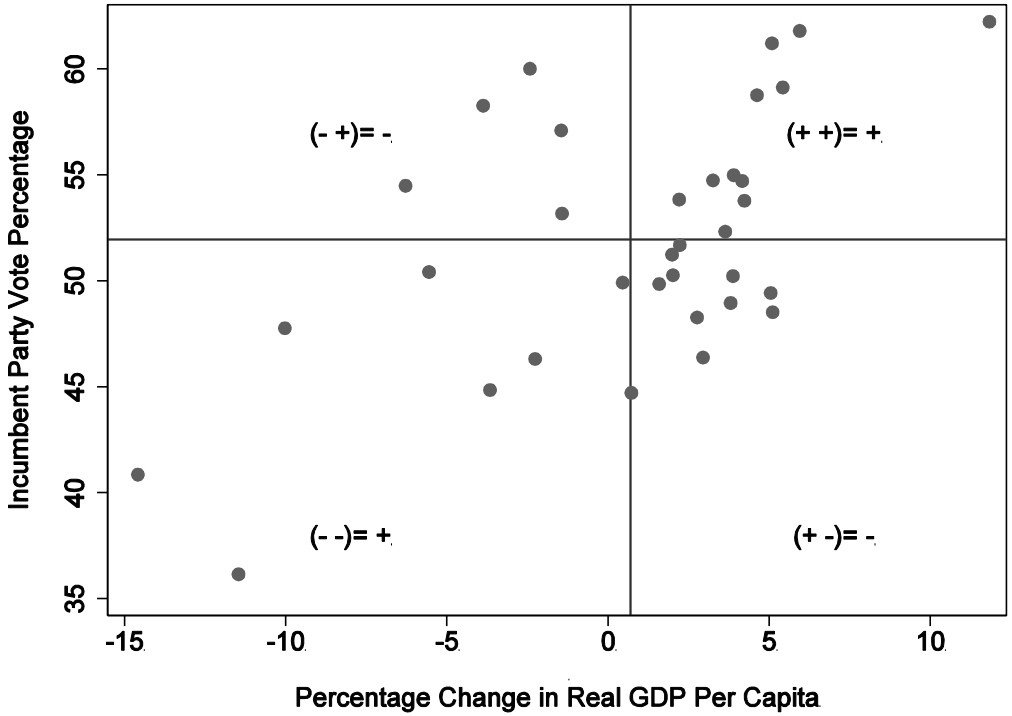
- Covariance is a statistical way of summarizing the general pattern of association (or the lack thereof) between two continuous variables. The formula for covariance between two variables X and Y is

$$\text{cov}_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n}.$$

- Whether two variables vary in the same direction
- To better understand the intuition behind the covariance formula, it is helpful to think of individual cases in terms of their values relative to the mean of X ($\bar{X}$) and the mean of Y ($\bar{Y}$) :
  - In addition to the level of difference, we start to have directions!!
  - Same direction: If $X_i - \bar{X} > 0$ and $Y_i - \bar{Y} > 0$, that case's contribution to the numerator in the covariance equation will be positive.
  - If $X_i - \bar{X} < 0$ and $Y_i - \bar{Y} < 0$, that case's contribution to the numerator in the covariance equation will also be positive, because multiplying two negative numbers yields a positive product.
  - If a case has a combination of one value greater than the mean and one value less than the mean, its contribution to the numerator in the covariance equation will be negative because multiplying a positive number by a negative number yields a negative product.

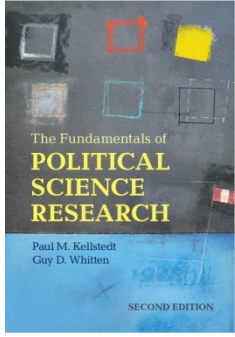


Scatter plot of change in GDP and incumbent-party vote share with mean-delimited quadrants



What would a null hypothesis/relationship look like?

Example 3: Covariance table for economic growth and incumbent-party presidential vote, 1880--2004

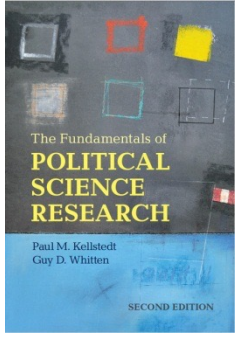


	Vote	Growth
Vote	35.4804	
Growth	18.6846	29.8997

Covariance (two variables)

- In a covariance table, the cells across the main diagonal (from upper-left to lower-right) are cells for which the column and the row reference the same variable.
- In this case the cell entry is the variance for the referenced variable.
- Each of the cells of of the main diagonal displays the covariance for a pair of variables.

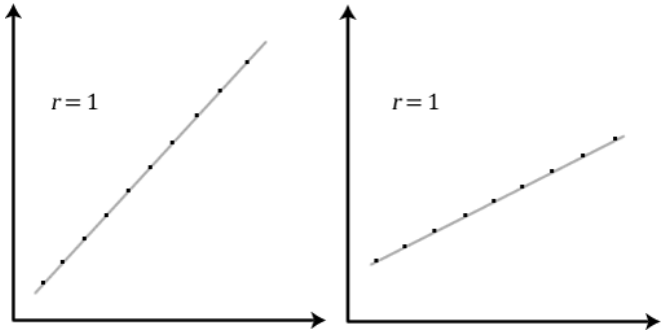
From covariance to correlation

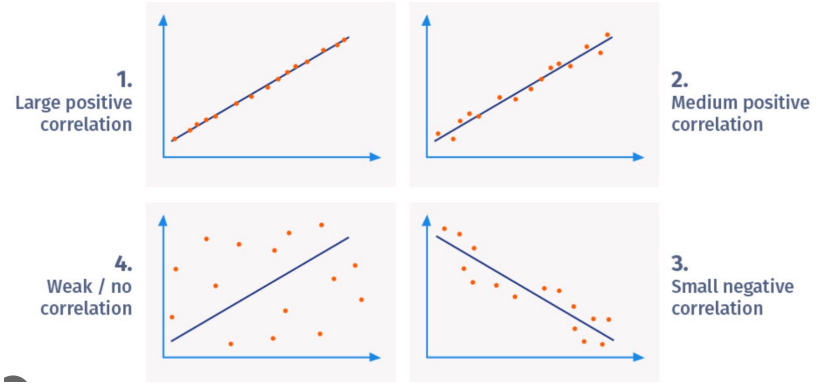


- H0 and H1: The covariance calculation tells us that we have a **positive or negative relationship**, but it does not tell us how **confident we can be that this relationship is different** from what we would see if our independent and dependent variables were not related in our underlying population of interest.
- To make this assessment, we turn to a third test developed Karl Pearson, *Pearson's correlation coefficient*. This is also known as "Pearson's r," the formula for which is

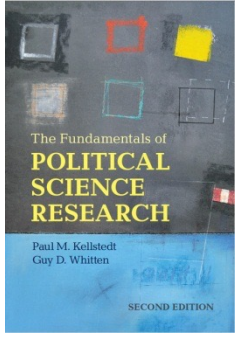
$$-1 \leq r = \frac{\text{cov}_{XY}}{\sqrt{\text{var}_X \text{var}_Y}} \leq 1$$

- There are a couple of points worth noting about the correlation coefficient:
  - If all of the points in the plot line up perfectly on a straight, positively sloping line, the correlation coefficient will equal 1.
  - If all of the points in the plot line up perfectly on a straight, negatively sloping line, the correlation coefficient will equal -1.





t-test for r

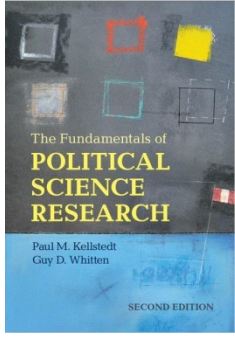


- We can calculate a t-statistic for a correlation coefficient as

$$t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}},$$

- with n - 2 (two variables) degrees of freedom, where n is the number of cases.
- In this case, our degrees of freedom equal 34 - 2=32.
- With the degrees of freedom equal to 34 (n = 34) minus two, or 32, we can now turn to the t-table in Appendix B. Looking across the row for df = 30, we can see that our calculated t of 3.96 is greater even than the critical t at the p-value of .001 (which is 3.385).
  - Steps: (1) calculate test statistics (2) calculate d.f. → go to the table
- This tells us that the probability of seeing this relationship due to random chance is less than .001.

Wrapping up



- We have introduced three methods to conduct bivariate hypothesis tests—tabular analysis, difference of means tests, and correlation coefficients.
- Which test is most appropriate in any given situation depends on the measurement metric of your independent and dependent variables.
- In the next chapter we introduce the final method for conducting bivariate hypothesis tests covered in this book, namely bivariate regression analysis.