University of Massachusetts Amherst Political Science 399

Using Statistical Techniques to Analyze Data

The word "data," plural of the Latin word *datum,* is used today to denote the numerical results of measuring some set of observable things using a common measuring device. Physical scientists use such devices as scales, radiation counters, thermometers, or spectroscopes. Social scientists can use them too, and also use such devices as currency units (dollars, euros, yen, etc.), votes, or ranking scales (such as the 5=strongly agree, 4=agree, 3=unsure, 2=disagree, 1=strongly disagree used in some opinion surveys).

Social scientists think of **data** as arranged in **data sets** that have four components:

1. A **unit of observation**, the thing whose properties, characteristics, or aspects have been measured. The unit can be an individual, a group of people, an object, or an event.

2. One or more **indicators**, the particular properties, characteristics, or aspects of the unit that have been measured. The indicators are chosen according to the question that the researcher is trying to answer. Thus a health policy researcher will be interested in what diseases a person has or has had while a welfare policy researcher is more likely to want to know about a person's income.

3. A set of **observations**, the results of the measurements taken. Each measurement of an indicator is a separate observation. Thus, a data set about individuals that included income, age, race, and religion would have four observations per individual. Alternatively, a researcher measuring individuals' income once a year for ten years would have ten observations per person.

4. A set of measurement **categories** into which the observations on each indicator are divided. These need to be mutually exclusive so that no single observation (one measurement of one unit on one indicator at one time) can be placed into two categories. Categories can be numerical, as when we report income as so many US dollars, Euros, Japanese yen, or other currency unit. They can be nominal (descriptive words), as when we indicate an individual's religion or gender. They can be ordinal (comparison terms) as when we rate something as high, medium, or low.

An indicator on which all observations fall into the same category (or, in an equivalent phrase, have the same value) is a **constant**. No matter how many researchers at sea level heat a container of water at the same time, or how many times any one of them does so, the thermometer in the water will always register 212 degrees F or 100 degrees C when it begins to boil. Though constants can be useful in constructing theories, social scientists are usually more interested in **variables,** the indicators on which different observations are in different categories (have different values).

Arraying Data Sets for Analysis

Social scientists have developed different techniques of statistical analysis for dealing with one variable (**univariate analysis**), two variables (**bivariate analysis**), and three or more variables (**multivariate analysis**). Yet these forms of statistical analyses share some basic characteristics. In each of them, the statistical calculations that can be used depend on the type of measurement employed. All can also rely on common methods of formatting data sets. Statistics is a special form of mathematics. Much of the time statistical calculations use arithmetical operations with which most people are familiar: addition, subtraction, multiplication, and division. Occasionally operations common to geometry (particularly calculation of areas – recall earlier discussion of the Gini coefficient) or calculus (determination of changes in rates of change) are used as well. In this unit focuses on statistical calculations that require arithmetic only.

The elaborateness of the mathematical operations that can be used depends on the mathematical quality of the data. Nominal data uses numbers as shortcuts for names of categories that are different without those differences having any mathematical relation to each other. (Despite long social traditions of regarding some races, ethnic groups, religions, or genders as "inferior" to others, social scientists treat these as categories that produce only nominal data.) The mathematical operations that can be done on data in nominal categories involve little more than counting, determining what percent of the total observations fall into any one of the categories, and comparing the distributions of two sets of observations. The same is true when the data has ordinal categories like high medium and low.

Only when researchers have numerical data expressed in interval and ratio categories, where the numbers have a clear mathematical relation to one another, can they use the whole array of statistical calculations. The examples used in this course involve interval or ratio data, but keep in mind that your choices of calculations are much more limited whenever your data reports observations on indicators (constants or variables) having only nominal or ordinal categories

Whatever the type of categories, a set of data can be displayed several ways. All researchers begin with a **raw distribution** listing each observation for each unit measured. With observations on lots of units, each unit is listed in the leftmost cell of one row of the table and the columns provide the observations on each indicator used in the study:

Individual	Income	Race/Ethnicity	Gender	Religion
1	45,375	white	female	Jewish
2	60,560	white	male	Catholic
3	38,700	white	male	Baptist
4	68,470	african-american	female	Baptist
5	42,460	hispanic	male	Catholic
6	63,325	asian	female	Buddhist
7	56,780	african-american	female	AME
8	49,340	hispanic	female	Catholic

Raw distributions focus on the individual units and help insure that all of the observations needed for the study have been made. However, they are not useful for finding patterns in the data. A **frequency distribution** listing the categories of the variable and how many observations fall into each category allows for some analysis of patterns. Focusing on the gender of the 8 individuals in the raw data above, would yield this frequency distribution:

male	3
female	5
(total)	8

Particularly with lots of observations having numerical categories it is often more useful to develop a **grouped frequency distribution** that bunches sets of categories into larger **groupings** (sometimes called "**intervals**" despite the risk of confusion with the term "interval data"). If our information about these 8 individuals were part of a larger data set, they would be part of a larger grouped frequency distribution on income that might look like this:

Income	Individuals (total= 500)
over 120,000	25
90,000-119,999	75
60,000-89,999	165
30,000-59,999	200
0-29,999	35

Researchers need to choose groupings (intervals) that are appropriately precise for answering the question being investigated. This grouped frequency distribution would be useful to voting behavior researchers, but they would not want to group income data into intervals of $250,000 (that is 0-249,999; 250,000-499,999; etc) because too many voters would be included in the 0-249,999 category for that grouped frequency distribution to be useful for answering the kinds of questions they typically investigate.

Comparing Data Sets

Researchers can begin analyzing data when they have something to compare. This might be two or more sets of observations about the same variable or variables taken in different places at the same time, in the same place at different times, or both. One basic form of analysis involves comparing the frequency distributions of the two or more sets of observations by calculating **relative frequencies** of various types. They could calculate a **ratio** for each data set that would allow comparing the frequency of a category or interval in one data set to the frequency of the same category or interval in another. Suppose that the four political parties in Slugonia each win this number of seats in the 150-member Slugonian parliament in two successive elections:

Party	2001	2006
Blue	37	33
Grey	19	16
Orange	71	78
Maroon	23	23

We can tell with an easy calculation (150 divided by 2 = 75, so 50% + 1 of the members is 76) that the Orange Party did not win a majority in the 2001 elections but gained one in the 2006 elections. Suppose, however, that we are interested in how well different ideological blocs are doing, and we know that the Orange Party is center-left, the Moroon Party is on the left, while the predominantly urban the Grey Party and the predominantly rural Blue Party are both on the center-right. We can calculate the ratio of seats held by the center-right as compared to the center-left as follows:

After the 2001 elections, the center-right held 37+19 = 56 seats. The center-left held 71. The ratio of center-right members to center-left members is 56/71 = 0.79. Since the 2006 election, the center-right has held 33+16 = 49 seats while the center-left holds 78. The ratio of center-right to center-left members has dropped to 49/78 or 0.63.

Political analysts tend to be more interested in the proportion of seats held by a party of bloc. A **proportion** is the ratio of the frequency of observations in one or more categories or intervals to the total number of observations. To determine the proportion of seats held by the center-right parties, we begin with the same addition: 37+19 = 56 for 2001 and 33+16 = 49 for 2006. This time, however, we divide that sum by the total number of seats, 150. Now we get 56/150 = 0.37 in 2001 and 49/150 = 0.33 in 2006.

When proportion calculations produce very small numbers, particularly numbers les than 1, researchers make reading them easier by calculating a **rate**. A country's birth rate is calculated by the dividing the number of live births in a year by the total population in that year, then multiplying the result by 1000. A country with 235,478 live births and a population of 31,367,289 in 2005 has a birth rate of 235,478/31,367,289 x 100 or 7.5 per 1000 population. Researchers can choose the number to use in the multiplication, and usually prefer numbers that will produce rates expressed in 2 or 3 digit numbers. If you think a moment, you can see that percentages are simply rates in which the multiplier is 100. Thus researchers wanting to know what percentage of a city's population is 70 or more years old would get the census data for that city, read from the data reporting ages how many people said they were 70 or older, divide that number by the number of people who live in the city, and multiply by 100. If 35,000 persons in a city of 870,000 are 760 or older, they are 4.02 percent of the population. If we take our proportions of center-right members of the Slugonian parliament and multiply them by 100, we get 37% in the 2001 elections and 33% in the 2006 elections. Whether expressed as a proportion (.33) or as a percentage (33%), the news is equally bad for the Slugonian center-right because their share declined. Yet the average Slugonian will understand the bad news better if it is expressed in percentages than in hundedths.

Researchers often calculate a **percentage change** to better understand shifts over time. These are calculated by subtracting the frequency of observations having a particular in the earlier data (n1) from the frequency of observations having the same value in the later one (n2), dividing the result of that subtraction by the earlier (n1), then multiplying by 100:

$$\frac{n2-n1}{n1} \times 100$$

Again focusing on the Slugonian parliament, we would calculate the percentage change in center-right members after the 2006 election as 49-56/56 = -0.125), x 100 = - 12.5%. Thus in going from 37% to 33% of the members, the center-right has lost an eighth of its parliamentary strength (100/8 = 12.5).

Researchers having two data sets of observations on the same variable expressed in ordinal, interval or ratio data can also calculate the median value of each set of observations and compare those medians. Recall from the discussion of averages that the median is the value splitting the observations into two equal-sized groups. This does not say much if the ordinal data has only three categories (low-medium-high) but does provide useful information when it has a larger number of categories. Interval and ratio data, being whole numbers and fractions, have enough categories for medians to be useful calculations. Means can be calculated only on interval or ratio data because the math involved assumes that each category is mathematically equidistant from the neighboring category.

Medians and means are sensitive to the distribution of the data. A table appearing in Dickinson McGaw and George Watson, *Political and Social Inquiry* (New York: John Wiley & Sons, 1976, p. 259) shows what can happen to medians and means when a data set with 7 observations has a close to normal distribution (column 1), an extreme value at one end (column 2) and a bimodal distribution (column 3):

income data set 1	income data set 2	income data set 3
25,000	100,000	32,000
22,000	15,000	32,000
19,000	14,000	32,000
18,000	11,000	16,000
17,000	11,000	11,000
14,000	10,000	11,000
11,000	10,000	11,000
mean = 18,000	mean = 24,714	mean = 20,714
median = 18,000	median = 13,300	median = 16,000
	mode = 10,000	modes = 11,000 & 32,000

As the table suggests, medians are more useful when data sets have one or a very few extreme values at one end; they are used in questions regarding incomes or income distribution so observations on Bill Gates and other extremely high income persons do not skew the analysis.

Neither means nor medians are particularly useful in bimodal or multimodal distributions. Statisticians do not worry about this very hard because in their collective experience most data distributions – particularly those with data covering 100 or more observations – are unimodal. However the wise researcher keeps the possibility of a bi- or multi-modal distribution in mind, and looks at the frequency distribution to get a sense of the shape of the data before beginning statistical analysis.

When researchers are interested in questions that cannot be answered by averages alone, they compare frequency distributions of two or more data sets. Careful statisticians also include variation measures in the descriptive statistics of their data sets, even when they intend only to analyze central tendencies, to show how representative the central tendency is of the whole data set. The lower the variation measure, the more representative the mode, median, or mean is because lower variation means that more observations cluster at values close to the central tendency.

The statistical operations that can be used for indicating variation depend, again, on the type of data available. With nominal data, researchers are limited to variation ratios and similar techniques. A **variation ratio** is the proportion of observations falling in the non-modal categories of the variable. This is calculated by the formula

$$1 - \frac{\text{frequency of the mode}}{\text{number of observations}}$$

Recall the data about Slugonian parliamentary elections:

Party	2001	2006
Blue	37	33
Grey	19	16
Orange	71	78
Maroon	23	23

We have an obvious mode, the Orange Party-held seats (when there is more than one mode, calculating a variation ratio requires choosing one as "the" mode and treating the observations falling on the other modes as "off the mode"). Our 2001 and 2006 variation ratios are

$$1 - \frac{71}{150} = 0.527 \quad \text{and} \quad 1 - \frac{78}{150} = 0.48$$

A Slugonian analyst could then conclude that the parliament had become less ideologically diverse after the 2006 elections, something Slugonian citizens might have figured out intuitively by noticing that rather than form a coalition or govern as a "minority government," the Orange Party had attained a majority.

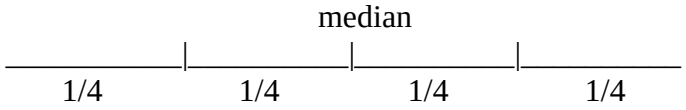
Yet this might be misleading. Suppose that an anarchist faction had developed in the Maroon Party, began competing for control of the party with the more conventional social democratic faction in 2004, and members of both factions had won seats in 2006. In this situation, the actual diversity of ideologies in parliament would have increased. The "take home message" of this little tale is that numerical data can hide significant conditions unless the indicators used are designed to focus on them.

Another way to summarize the amount of variation defines variation as pairs of observations falling in different categories. In the Slugonian parliament elected in 2001, we had 37 Blue Party members, 16 Grey Party members, 71 Orange Party members, and 23 Maroon Party members. To calculate the number of different pairs, start with the first category (Blue Party) and calculate how many pairs there are of observations in that category in relation to observations in each of the other categories (Grey, Orange, and Maroon Parties). Hence 37x19 (703) + 37x71 (2627) + 37x23 (851) or 4181. Then move to the second category (Grey Party). Since the number of different pairs with the first category has already been calculated, focus on the categories listed below it (Orange Party and Maroon Party). Now we have 19x71 (1349) + 19x23 (437) or 1786. Now move to the third category, and focus on the categories below it (in this case only one, Maroon Party) and do the same sort of calculation. Here we have 71x23 or 1633. Take all the sums and add them together to get the total number of different pairs: 4181 + 1786 + 1633 = 7600. The same calculation for the parliament elected in 2006 yields a smaller number, 7271. Again, the statistical calculation confirms what Slugonians probably figured out intuitively, but the calculations would be useful if Slugonians or social scientists wanted to track variety in parliament over more than two election cycles or were dealing with a nominal data set having a larger number of categories.

Since the total number of pairs with different categories can be very large, statisticians often simplify the reading of comparisons (but not their own work) by calculating an **index of qualitative variation**. This compares the actual data set with a hypothetical distribution of the same data containing the maximum possible variation. Maximum possible variation is calculated in two steps. First divide the total observations by the number of variable categories to yield an equal distribution in all categories. Then calculate the number of different pairs in this hypothetical data set. With this figure in hand, calculate the ratio of the variation in the actual data to the hypothetical maximum. For the 2006 Slugonian parliament, the hypothetical maximum variation set would have 4 categories (parties) with 37.5 elected members each (it's a good thing this is hypothetical; parliaments do not include half members). Doing the pairs calculation gives a maximum variation of 8437.5. The ratio, then, is 7271/8437.5 or 0.86.

With ordinal data, analysis can use measures that indicate dispersion of data around the median. The **range**, as noted in the discussion of averages, is the difference between the highest and the lowest value in the data set. This may not say very much, so analysts like measures that also take the frequency of values into account. They may use the **inter-quartile range**, which is the difference between the highest and lowest values within which the middle 50% of the observations fall. 25% of the observations will be below the lower value, 25% between the lower value and the median, 25% between the

median and the higher value, and the last 25% above the higher value. In line diagram form:



Inter-quartile ranges are particularly useful when either or both data sets being compared have observations with extreme values at the high end, the low end, or both. "Data set 2" on page 6, with one person having a $100,000 income, has an extreme value at the high end. One or a very few observations with extreme values (called **outliers** because they have values distant from those of most observations in the data set) will artificially expand the range. Clipping the range by using the inter-quartile range instead focuses attention on the main cluster of observations.

Analysts having interval or ratio data can use more fine-grained measures of variation that work by calculating the distance of each observation in the data set from the mean. The **variance** is calculated in four steps; 1) subtracting the mean from the value of each observation (this produces positive numbers with values above the mean and negative numbers with values below), 2) squaring (multiplying by itself) the result of each subtraction (this produces all positive numbers because a negative multiplied by a negative = a positive), 3) adding up all the squared numbers, and 4) dividing by the total number of observations in the data set.

This is the calculation of variance for "income data set 1" on page 6, the one with a mean of $18,000:

25,000 – 18,000 = 7,000, squared = 49,000,000		
22,000 – 18,000 = 4,000, squared = 16,000,000		
19,000 – 18,000 = 1,000 squared = 1,000,000	$\frac{132,000,000}{8}$	= 16,500,000
18,000 – 18,000 = 0, squared = 0		
17,000 – 18,000 = -1,000, squared = 1,000,000		
14,000 – 18,000 = - 4,000, squared = 16,000,000		
11,000 – 18,000 = - 7,000, squared = 49,000,000		
132,000,000		

16,500,000 is a rather large number, and it does not seem to have any relation to the data. It would seem less bizarre if we had two income distributions we were comparing because even though the variance of the other distribution would also be a very large number, we would know from whether that number was greater or less than 16,500,000 if the other income distribution included greater differences of income than this one.

As you can guess from this short example, calculating variance for a large data set was a daunting process – much harder than the different pairs calculations done on pages 7-8 for the Slugonian elections– when it had to be done with pencil and paper or adding machine. It is now easily accomplished by using the "descriptive statistics" tool in

statistical software, Excel, or any other software that has a statistical analysis tool. You enter the data into a spreadsheet, and the computer does the math for you.

Once the variance has been calculated, it is easy to calculate a much more comprehensible summary statistic, the **standard deviation**. This is the square root of the variance (the number that multiplied by itself equals the variance). For this income distribution, the standard deviation is 9859.52. Again, we could calculate the standard deviation of another income distribution and compare it to the 9859.52. that we got for this one. These smaller numbers are a lot easier to process than big ones of the variance.

The standard deviation does not tell us much about the shape of the data distribution; we still need a frequency distribution table for that. When (and only when) a data set has a normal distribution, we can use the standard deviation to calculate the values that observations are likely to have. In a normal distribution, 34.13% of the observed values will be between the mean and standard deviation, and 68.26% (or a little more than 2/3 of all observations) between the values calculated by subtracting the standard deviation from the mean to get the lower side value and adding the standard deviation to the mean to get the higher side. In a normal distribution with a mean of 18,000 and a standard deviation of 9,859.52, we would expect 68.26% of the observations to have values between 8,140.48 (18,000-9,859.52) and 27,895.52 (18,000 + 9,859.52). In a normal distribution, 95% of the observations will lie within 2 standard deviations of the mean, and 99% within 3 standard deviations.

Even though the standard deviation of a set of interval or ratio data tells us nothing about the range of values in that data set or about the frequency distribution of the values in it, statisticians use the standard deviation as a measurement of variation because it has mathematical properties helpful in calculating tests of **statistical significance.** The basic notion of statistical significance is that it measures the likelihood that the pattern of observations in a data set results from a correlation or a causal relation rather than being the product of random variation among observations. This is important in inferential analysis, as you will see.

Univariate Analysis

Because a univariate data set shows the distribution of observations on a single variable, analysis of a single data set cannot go much further than determining the central tendency (mode, median, or mean depending on whether the data is nominal, ordinal, interval, or ratio) the relative frequency of each value among the observations, and from the relative frequencies the distribution of the data.

Researchers can use two or more univariate data sets in comparisons with each other. Such comparisons might involve comparing observations taken at one time (“time 1”) with observations taken at another time (time 2”), as we did earlier when comparing the results of the 2001 and 2006 Slugonian parliamentary elections. They might involve comparing observations taken in two places at the same time. If the three income distributions shown earlier were the income distributions for three different cities in

2005, we could compare incomes in three cities and learn things about the conditions of life or the politics of each to the extent that income affects political attitudes or outcomes.

If we have a large univariate data set, we might be able to divide it into segments and make comparisons between the segments. This is common in studies of voting and elections, where the comparisons are used to see if there are any notable non-random inter-regional differences. Thus studies of voting in the USA often divide the 50-state data sets into four (northeast, south, midwest, west) or more regions. Political analysts and citizens in the USA share a sense that regional differences exist and matter; the statistical analyses help them see if the differences are as large as people think and how, if at all, they affect voting or attitudes on the main national issues of the day.

Bivariate Analysis

Bivariate analysis operates on data sets reporting simultaneous observations of each observed unit on two variables. We can think of the regionally-divided voting data as "bivariate" if we treat region itself as a variable. While the fact region is a nominal variable limits the mathematical operations we can use on the data, it does open up the wider resources of bivariate data analysis. Assessing whether there is a "gender gap" on various issues, and if so how large it is has been inspiring a lot of bivariate analysis in recent years. In these studies, gender is one variable and responses to public opinion polls on the issue is the other.

Bivariate data is displayed in a **cross-classification table**. One of the variables observed is used for the column headings, and the other for the rows. Each cell includes the number of observations having the row value and the column value simultaneously. Thus, the top left cell of the table below informs us that 55 of the urban dwellers polled said Issue A is the most important issue facing the country. Typically the observations on the "independent variable" (the one being treated as the antecedent) are displayed in the columns and the observations on the "dependent variable" (the one being treated as the consequent) are displayed in the rows. A bivariate table of the frequency of 450 individuals' responses to the question "what is the single most important issue facing the country today" might look like this:

	urban	rural	total
Issue A	55	78	133
Issue B	105	28	133
Issue C	35	34	69
Issue D	55	60	115
total	250	200	450

The absolute numbers are hard to compare directly, so analysts often **percentage the table** to make comparison. To percentage a table, an analyst first determines which is the independent and which the dependent variable. Here, the decision is suggested by the fact that type of place of residence occupies the columns and issues occupy the rows. The implicit theory is that city dwellers and rural dwellers have different ideas about

what is the "most important" issue. Percentaging begins by calculating the percentage of total responses in a row. This tells us what percentage of all observations on the dependent variable appear in each row (here all respondents in the survey mentioning each issue as the most important one). Thus we calculate:

133/450 x 100 = 29.6 %; 133/450 x 100 = 29.6%; 69/450 x 100 = 15.5%; 115/450 x 100 = 25.6%.
The right bottom cell gives the total number of units observed. You may have noticed that adding the row percentages (29.6+29.6+15.5+25.6) equals 100.1%. This happened because the row calculations were rounded off to the nearest tenth. s long as the sum of the percentages is very close to 100, things are ok.

Now comes the more complicated part, percentaging the distributions of observations within each independent variable category. This involves dividing the number of observations reported in each cell of each row (hence the name **cell frequency**) by the total number of observations in that column. Here the percentage, or cell frequency, for urban residents naming Issue A as the most important is 55/250 x 100 or 22.0%. Doing the same for the rest of the cells in the urban column and all of the cells in the rural column yields this percentage table when we also include the total percentages calculated previously:

	urban	rural	total
Issue A	22.0%	39.0%	29.6%
Issue B	42.0%	14.0%	29.6%
Issue C	14.0%	17.0%	29.6%
Issue D	22.0%	30.0%	25.6%
Total	100.0%	100.0%	100.1%
(N)	(250)	(200)	(450)

The table tells us that 22% of the urban dwellers responding and 39% of the rural dwellers responding think Issue A is the most important. Analysis can then calculate the difference between percentages in each independent variable category by subtracting the smaller from the larger. This on issue A is 39%-22% = 17% and we are encouraged by this poll result to conclude that rural dwellers are more likely than urban dwellers to regard A as the most important issue. The "gap" is even bigger on issue B – 42%-14% = 28%. This is a bigger difference, one we might put into words by comparing the 42 to the 14 and saying that "three times as many city people as rural people think A is the most important issue facing the country today."

Percentaging helps readers comprehend observed results, but professional data analysts prefer working with **measures of association**. These are a way of determining the extent to which the values of the dependent variable in a particular set of observations are influenced by the value of the independent variable. Association can range from none to complete, as exemplified by these two tables of votes in the US Senate at a time when

it has only Democratic and Republican members:

Vote on Bill 1

	Dem	Rep	Total
yes	20	30	50
no	20	30	50

Vote on Bill 2

	Dem	Rep	Total
yes	40	0	60
no	0	60	60

In the vote on Bill 1, there is variation on the variable "vote" (which has the values "yes" and "no") in the rows. With a two-category variable we can calculate this with the formula

V = 1 – total # of observations in first row / # observations taken

In these votes, V = 1 – 50/100 or .50 in the vote on Bill 1, and 1 – 60/100 or .40 in the vote on Bill 2.

What we want to know is whether any of this variation resulted from ("can be attributed to" in social science jargon) Senators' party affiliation (the value of the other variable observed in this data set). This is calculated by determining the amount of variation within each category of that variable. Calculating within-category variation for this table involves using the number of observations in the top cell of the each column and the number of observations in the second cell. In the vote on Bill 1, the by-party variation for Democrts is 1 – 20/40 = .5 and for Republicans s is 1 – 30/60 = .5. In the vote on Bill 2, the by-party variations are V (Dems) = 1 – 40/40 = 0 and V (Reps) 1-60/60 = 0.

It does not take any math to figure out that party affiliation absorbed all the variation in the vote on Bill 2. The situation on Bill 1 is less obvious and requires using some additional calculation to arrive at a measure of association. This involves using the values of the **original variation** – the variation on the dependent variable alone, and of the **unexplained variation** – the amount of variation left after the independent variable is brought into the analysis. The difference between the original and the unexplained variation is the **explained variation** – the variation in the dependent variable that disappears when the independent variable is included in the analysis. In these tables, vote is the dependent variable and party the independent variable (while the independent and dependent variables could be reversed in some analyses, there is no theory or observation of legislator behavior that says votes cause party affiliation and a lot that says party affiliation at least partly causes votes).

To calculate the measure of association, use the formula:

$$\frac{\text{original variation – unexplained variation}}{\text{original variation}} = \frac{\text{explained variation}}{\text{original variation}} = \text{proportion of o.v accounted for}$$

Recall that we have two figures for variation by party. The unexplained variation is the mathematical average of the values of the variation ratio for each category of the independent variable. In the vote on Bill 1, then, the unexplained variation is .5+.5/2 or .5.

Now we can calculate with the formula that $\frac{.5 - .5}{.5} = \frac{0}{.5} = 0$

This result that none of the variation is accounted for by the independent variable is unusual, and stems from the fact that equal numbers of Democrats and Republicans voted yes and no. If we had had a more usual result, say one in which 10 Democrats voted yes and 30 voted no while 60 Republicans voted yes, our measure of association would be

original variation = 1-70/100 = .7
within-Dem variation = 1-10/40 = .4
within Rep variation = 1-60/60 = .0
unexplained variation = .4+0/2 = .2 and therefore

the proportion of original variation accounted for by party is $\frac{.7-.2}{.7} = \frac{.5}{.7} = 0.71$
An association coefficient of .71 means that party affiliation explains 71% of the variation in votes.

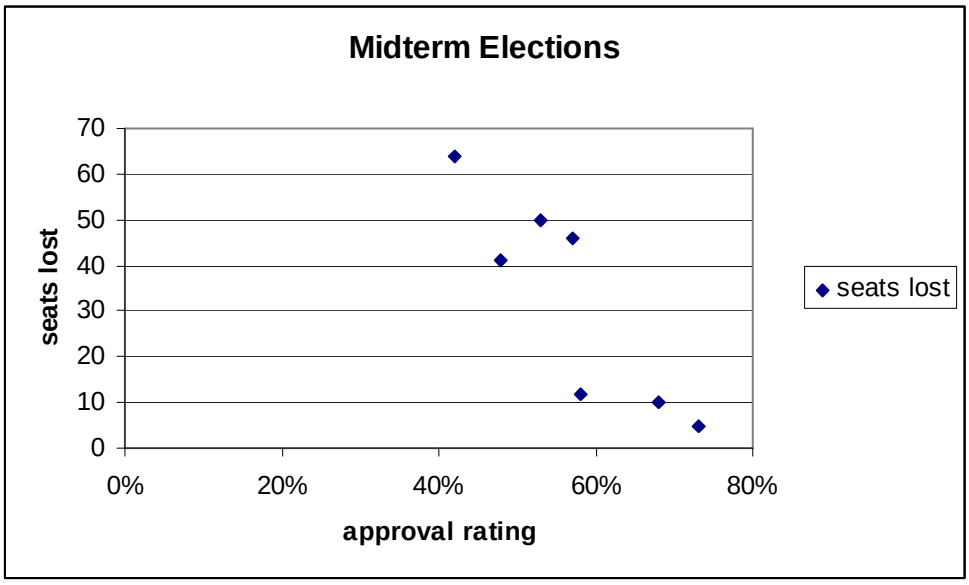
When a bivariate data set contains only interval or ratio data, we can also calculate the mathematical relationship between the two variables to estimate the value of an observation on one variable when the value of an observation on the other variable is known. Calculating this uses the technique of **regression analysis.** The first step involves calculating a **regression coefficient** for the data, which states the change in the value of the variable being guessed per unit of change in the value of the other variable. That other variable is called the **predictor variable,** and is the one for which an observation is either available or assumed. When we have a causal or correlational hypothesis, the predictor variable is the one we are treating as the independent variable.

This sort of bivariate regression analysis begins with a **scatter diagram** of the data plotting the values on both variables. The values of the independent variable are put on the horizontal dimension (the X axis) and the values of the dependent variable on the vertical (the Y axis). Suppose political analysts in Prova, a country with a presidential system, want to see whether Prova midterm elections are anything like those in the USA where the president's party usually loses more seats in midterm Congressional elections if the president's approval rating is low at the time. They have the following data on approval ratings of the president's performance and losses of seats by the president's

party in the Prova National Assembly after each midterm election in 1980 - 2004:

Election	approval	seats lost
1980	42%	64
1984	53%	50
1988	48%	41
1992	58%	12
1996	68%	10
2000	57%	46
2004	73%	5

It is hard to see a relation from this report of raw data. A pattern does seem to emerge when the data is arrayed in a scatterplot:



This data appears to express a linear relation, which is fortunate for us because **linear regression** is the least mathematically complex form of regression analysis. It involves drawing a line through the plot in a way that minimizes the distance between the line and each point (joint observation on the two variables being analyzed) in the scatterplot. Obviously, linear regression assumes that the relation between the two variables is linear – that as variable X changes one unit, the value of variable Y changes in equal increments.

The basic regression prediction formula is

$$Y' = \alpha + \beta X$$

Y' is the predicted value of Y, α (alpha) is the intercept point where the regression line drawn along the scatter diagram hits the Y (vertical) axis, β (beta) is the slope of the regression line, and X is a selected value of the predictor variable.

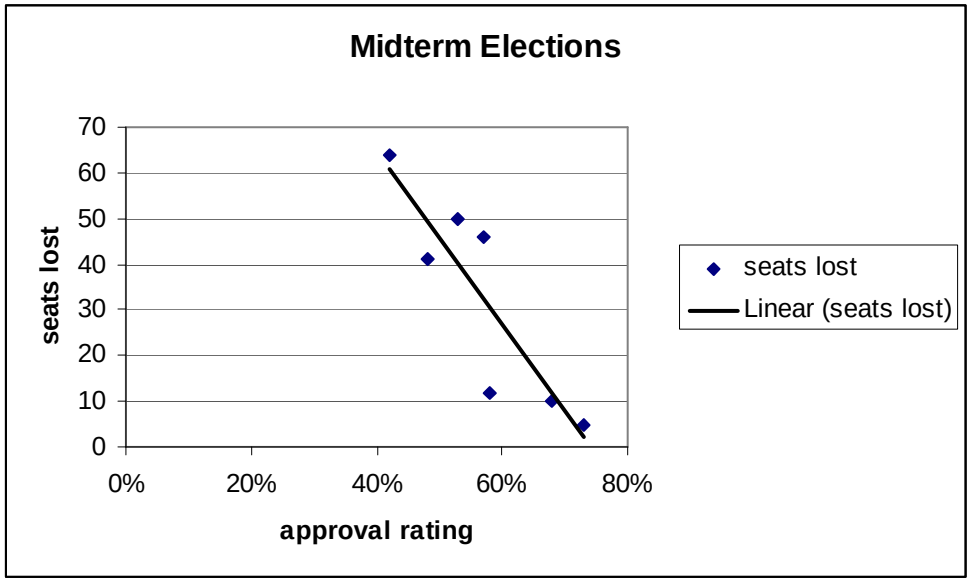
The calculation of β, the slope of the regression line, cannot be done until the line has been fitted to the data. This is done, as noted above, by minimizing the difference between the points on the line and the squared deviations of observed Y values from the line. Though statistics textbooks give a different formula when defining β, the preferred way to calculate it is:

$$\beta = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}$$

N = number of cases observed, ΣXY = sum of the crossproducts of the scores (that is, the score on X times the score on Y for each individual unit measured)
ΣX = sum of all scores on X, ΣY = sum of all scores on Y, and
ΣX^2 = sum of the squared scores on X

These calculations are rather lengthy once the number of units observed goes above 20. This is where computers are very handy. Excel and most similar spreadsheet software can fit lines and calculate α and β for you.

Here is the line for our data set on midterm elections:



In this line, $\beta = -1.92$, indicating a loss of 1.92 seats for every 1% drop in the president's approval rating. With β known, we can calculate α (where the regression line would cross the vertical (x) axis using the fairly easy formula α = the mean of $(Y - \beta)$ multiplied by the mean of X. = 140.43.

Suppose it is close to the 2008 election, and the president's approval rating is 42%. The party leaders will anticipate from past elections that their party will lose 140.43 + (-1.92x42) or 59.79 seats. They will round that up to 60 seats, and be looking hard for a way to get the president's approval rating up unless they perceive there has been some really big political change that makes predictions based on past elections irrelevant for 2008.

Meanwhile, political scientists in Prova will be interested in whether the two variables (approval rating and loss of seats) are causally related. They can see from the fact that the slope of the regression line is not zero that there is an association. That $\beta = -1.92$ tells them that every 1% decrease in approval rating means the loss of an additional 1.92 seats. What they want to know is the strength of the relation between the two variables, and whether it is reasonable to regard the data as showing a real connection between the variables rather than the result of random variation.

Political scientists in Prova have good training, and know that when dealing with any method – not just statistical analysis – it is possible to make mistakes. They also know they have to worry about two sorts of error – accepting a hypothesis as true when it is actually false, and rejecting a hypothesis as false when it is actually true. Robert Miller ("Hypothesis Testing" in Miller and Brewer, eds. *The A-Z of Social Research*, 2003, p. 145) summarizes the possible results of statistical analysis this way:

	hypothesis is actually INCORRECT	hypothesis is actually CORRECT
researcher accepts the hypothesis	**Type I error – the worst**	the right call
researcher rejects the hypothesis	the right call	**Type II Error – less bad**

Notice that accepting an incorrect hypothesis as true is regarded as a bigger mistake than rejecting a correct hypothesis. Social scientists prefer to err on the side of caution and not accept a hypothesis as true unless the statistical (or other) confirmation is very strong.

In their quest to determine whether they have an actual relationship, they start with determining the strength of the relationship by calculating a **standard error of estimate** which indicates the amount of error that occurs in estimating the value of Y by using the regression line rather than an actual observation. A small standard error of estimate indicates that the regression line lies close to the actual observed values of Y, a large one indicates that it does not. (If the regression line hits all the observed values on Y exactly, it has a standard estimate of error = zero.)

The standard error of estimate is calculated with the equation

$$\sqrt{\frac{\Sigma(Y-Y')^2}{N}}$$

(Read this as the square root of the entire expression; I can't get the math symbols to work quite right at the moment)

Again letting the computer do the mathematical calculations, the standard error of estimate on the Slugonian midterm election data is 12.00. The maximum value the standard error of estimate can attain is the standard deviation of Y. This is helpful for one line, but does not help when comparing lines generated from two sets of data. Over the years, statisticians have determined that comparing lines can be done best with the correlation coefficient. This is handy because it can also be used to assess whether the relationship is real or the result of random variation.

Because they have ratio data, the Provan political scientists can use measures of association appropriate for interval and ratio data rather than depending on the less precise measures that must be used with nominal or ordinal data. The measure for interval and ratio data is called **Pearson's r.** Like the measures used with nominal and ordinal data, Pearson's r is constructed to vary from zero (no association) to +1 (perfect positive association) or -1 (perfect negative association). Like beta for regression lines, Pearson's r has a definition formula that is somewhat awkward, so statisticians prefer the calculational formula:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2]\,[N\Sigma Y^2 - (\Sigma Y)^2]}}$$

(Again, √ means "the square root of" and applies in this formula to all the mathematical operations below the line.)

The math here is sufficiently complex that computer calculation is a real time-saver. Some simpler spreadsheet software does include Pearson's r and related measures in one of the standard routines, but if you have a really large data set you need to use full-feature statistical software like SPSS, STATA, or R.

For the results of Prova's midterm elections, we get Pearson's r = 0.881732441, which suggests a fairly strong relation.

Squaring r yields **r^2,** the **coefficient of determination.** This indicates the proportion of the variation in the dependent variable (in this example seats lost) explained by changes in the independent variable (in this example the president's approval rating at election time). The r^2 for this analysis of Provan midterm election is 0.777452097, suggesting that almost 78% of the variation in the number of seats lost can be attributed to the level of voter approval of the president's performance.

Tests of statistical significance indicate the likelihood that the **"null hypothesis"** – the relation seen in the data is the result of randomness rather than any actual relation between the independent (X) and the dependent (Y) variables – is true.

While a **t test** can be used when an independent variable has two categories, the **F-test** associated with **analysis of variance** (ANOVA) is better because it can be used with independent variables having any number of categories. Though they are pretty sure there is a relation between presidential approval rating and midterm election results in Prova, the Provan political scientists want to calculate an F score for their data. I will not carry out the calculation here because it is complex; computer software can do that for you. I will summarize the elements that go into the calculation so you understand what the machine is going. The process involves dividing the total variation into two components, the SSW – sum of squares within categories – and the SSB – the sum of squares between categories. Each sum of squares is associated with a number indicating the degrees of freedom in the data, a hypothetical construct of how many observations can vary once the value of one is known. The degrees of freedom within (dfw) is calculated as number of cases (N) minus number of categories (k) while the degrees of freedom between (dfb) is calculated by number of categories (k) minus 1. Each sum of squares is divided by its degrees of freedom to yield the terms of the F ratio:

$$\frac{SSW}{dfw} \text{ for the mean square within and } \frac{SSB}{dfb} \text{ for the mean square between}$$

Dividing the mean square within by the mean square between yields the F ratio. At an 0.05 confidence level (that is, giving ourselves a 95% probability that the null hypothesis is true), we then consult a table to determine whether the F ratio we calculate is within or outside the range of results which lend credence to the null hypothesis. We look across the columns to find the degrees of freedom between and the rows to find the degrees of freedom within. For the Prova data, analysts would need to look at the entry corresponding to the intersection of the 7-2=5 column and the 2-1 +1 row. To reject the null hypothesis, they need an F test greater than 6.61, and since it comes out to 17.46, the Provan political scientists can be quite sure that there is a real relationship here, not just a random play of miscellaneous factors. If you are using Excel or another spreadsheet, it will calculate F and also give you the statistical significance because the computer has the F-ratio table in its memory when running Excel or statistical software. If the "significance" entry in the column just to the right of F shows 0.005 or a lower number, there is a less than one-half percent chance that the relationship is random and you can be very confident that the result is not random. Most scial scientists conclude there is a relation if the significance number is no higher than 0.01; a few use 0.05. The first means there is a 1% chance the relation is random, the second that there is a 5% chance. I am cautious myself and feel good only with significance numbers at or less than 0.01.

Political scientists who have nominal or ordinal data must use different measures to determine whether their statistics reveal a relation between the two variables. For nominal data, the sort of data where numbers are simply codes for verbal descriptions, researchers rely on the χ^2 (chi-square, usually pronounced as "kai-square") test. For this

test (as with F) the "null hypothesis" – the alternate possibility that we want to reject – is that the variable named in the row captions and the variable named in the column captions of the table are independent of each other; that is, there is no correlation between the two. Suppose government statisticians in Ochto, a developing country, are studying whether subsistence farmers are less likely than factory workers to send their children to school long enough for them to complete a primary education. They have the following data after some survey research in rural areas and in cities:

	subsistence farming	factory worker	total
do not complete	75	25	100
complete	50	150	200
total	125	175	300

For this data set, the null hypothesis is that parent's occupation has no relation to how far a child progresses in education.

The standard formula for χ^2 is $= \Sigma \frac{(\int o - \int e)^2}{\int e}$

$\int o$, the observed frequencies in the cells is read from the table.
$\int e$, the expected frequencies if the two variables were independent of each other, is calculated on the formula $\frac{\text{row marginal x column marginal}}{N}$

For the Ochto data, then, the expected frequencies are:

$\frac{100 \times 125}{300}$ or 41.6, $\frac{100 \times 175}{300}$ or 58.3, $\frac{200 \times 125}{300}$ or 83.3, and $\frac{200 \times 175}{300}$ or 116.6

Thus we calculate χ^2 with its standard formula as follows:

$$\chi^2 = \frac{(75-41.6)^2}{41.6} + \frac{(50-58.3)^2}{58.3} + \frac{(50-83.3)^2}{83.3} + \frac{(150-116.6)^2}{116.6}$$

$$= \frac{33.4^2}{41.6} + \frac{-8.3^2}{58.3} + \frac{-33.3^2}{83.3} + \frac{33.4^2}{116.6}$$

$$= \frac{1115.56}{41.6} + \frac{68.89}{58.3} + \frac{1108.89}{83.3} + \frac{1115.56}{116.6}$$

$$= 26.81 + 1.181 + 13.31 + 9.57$$

$$= 50.87$$

So far, so good, but this does not yet say anything about whether the relationship is random or not. As with using F, we need to work out the statistical significance of this

value. As with F, there is a table that provides the probability that a relationship is random at different values of chi-square. To use the table, we need one more calculation, the **degrees of freedom** of the original table. This is easy, df = (number of rows – 1) x (number of columns – 1). Going back to our table, we see that the table has two rows and two columns of data. Thus df = (2-1) x (2-1) or 1 x 1 = 1. We then use the value 1 to look things up in the table. In general, the higher the value of χ^2, the higher the likelihood that the relationship is not random. However, degrees of freedom matter. With 1 degree of freedom, the value of χ^2 with a probability of 0.001 (that is, a one-half of 1% chance of being random is 10.827; with 30 degrees of freedom the value of χ^2 with the same probability is 59.703. Our table has 1 degree of freedom, and out χ^2 is 50.87. We can be vert confident that there is a relation between father's occupation and a child's completion of primary education. When using a statistical software program, the computer will do all of this for you and report significance as well as χ^2.

Remember that many bivariate relations are not linear, so cannot be analyzed using linear regression. Some can be analyzed with more complex **curvilinear regression** methods. There are different polynomial (useful for S-curves), logarithmic, exponential, and moving average techniques. The mathematics are more complex, so I will not discuss them here. If you ever get to the point of doing that kind of analysis, there are textbooks that can guide you through the process and very sophisticated computer software will do the "grunt work" of calculations for you.

With all the different levels of measurement available, and the need to use different statistical calculations in analysis of data at the different levels of measurement, it can be hard to remember which measure of association and test of statistical significance goes with what type of data.

Here is a handy table covering the various forms of bivariate analysis:

Level of Measure-ment of the Data	Measures of Association	Range of Measures of Association	Tests of Statistical Significance
two nominal variables	Lambda Phi Cramer's V Tau в	0 to +1.0 0 to +1.0 0 to +1.0 0 to +1.0	Chi²
Two ordinal variables	Gamma Kendall's Tau в Kendall's Tau c	-1.0 to +1.0 -1.0 to +1.0 -1.0 to +1.0	Chi²
Two interval variables	Pearson's r	-1.0 to +1.0	F test
One nominal and one interval variable	Eta	0 to +1.0	F test t-test Difference of Means

[From Alan D. Monroe, *Essentials of Political Research* (Westview Press, 2000), p. 101.]

Multivariate Analysis

Bivariate analysis allows more comparison than univariate analysis, but often understates the complexity of correlational or causal relations in the real social world. Most of the time more than one factor is influencing what we observe. Take, for example, an effort to explain popular attitudes on immigration. Suppose survey researchers in Imnar, where immigration has been a hot issue in recent years, report the following results of a survey on whether people support or oppose the government's proposal to reduce the number of immigrants admitted to the country each year by 50% and to crack down on illegal immigrants by finding them and deporting them to their home country:

	Urban	Rural
Support the proposals	37%	45%
Oppose the proposals	63%	55%
(N) (total responses)	(120)	(120)

With this data alone, an interested citizen would conclude that city dwellers are more receptive to immigration than rural dwellers. Their perspectives might change if they saw this table instead:

	Self-employed	Self-employed	Employee	Employee
	Urban	Rural	Urban	Rural
Support	70%	50%	30%	20%
Oppose	30%	50%	70%	80%
(N)	(20)	(100)	(100)	(20)

This starts everyone thinking about other possibilities. If people think that the self-employed don't have to worry as much about being laid off (losing their whole income at once) then they might draw a connection between greater confidence in continued income and receptiveness to immigration.

Multivariate analysis can focus on any of three types of effects of change in one variable on change in another:

1. **control effects** – the additional variable or variables are used to further test the validity of a bivariate relationship or to specify the conditions under which the bivariate relationship does or does not remain true, or conditions under which it is weakened or strengthened by the presence of other factors.

2. **relative effects** — analysis focuses on determining the relative separate impact of each variable assumed to be an independent variable in the model or hypothesis being tested using statistical techniques to "control for" the other variables. This relies on more complex measures of association. While helpful for reminding analysts that few relations

involve only one independent variable producing an effect, it does assume that the effects are additive; that is, that each additional variable explains a particular amount of the variation. The co-presence of two factors is not regarded as contributing anything to the value of observations on the dependent variable.

3. **conjoint effects** – additional variables are understood as interacting with one another to produce the result. Mathematically, this is done by designating one of the independent variables as the "first" one and using the others as "additional".

Inferential Data Analysis

So far we have been dealing with data sets giving complete sets of observations (all parties in an election; all midterm elections in a particular period). However social scientists often deal with data sets that include observations on only a sample of the units of analysis that could possibly be observed. This is true with polls, but it can be true of other sorts of data as well. When researchers have a data set about a sample, and want to determine what that data says about the whole population, they use the techniques of **inferential data analysis** based on probability theory to figure out the descriptive statistics that would apply to the whole population from the statistics on the sample.

Inferential data analysis typically uses means, proportions, and variances of sample data to figure out likely means, proportions, and variances of populations. Since it is important to keep track of which statistics refer to the sample and which refer to the whole population, researchers refer to the numbers summarizing the sample data as "sample statistics" or "statistics" and the numbers probably summarizing the whole population as "population parameters" or "parameters." (This is a very specialized use of the word "parameters;" don't confuse it with the usual meanings of the word.)

When trying to infer population characteristics from observed data regarding a sample, researchers add a test of statistical significance to reports of results. Like the "sampling error" mentioned in the discussion of surveys and polls, tests of **statistical significance** indicate the degree of confidence a researcher should place in the idea that the relation observed among the sample observed would also be observed if every unit of the population could be observed. Tests of statistical significance indicate the likelihood of the **"null hypothesis"** – that the relation seen in the sample is the result of randomness rather than any actual relation between the independent (X) and the dependent (Y) variables.

With interval or ratio data, researchers use Pearson's r and r^2 to determine the strength of the relationship in the sample. They then add a **t test** to assess the likelihood that the population would exhibit the same relation. A t test is accurate when three assumptions about the sample and population are true: 1) both are normally distributed, 2) the relation between independent variable and dependent variable is roughly linear, and 3) the data is **homoscedastistic,** a difficult to pronounce term denoting a situation in which the observed values on the dependent variable are spread roughly evenly above and below the regression line. For most purposes, this can be checked by looking at the

scatterplot. If the regression line is located within the data points of the scatterplot like a lane marker running down the middle of a road formed by the plotted data points, it should be safe to use the t test. **Heteroscedastistic** data is the opposite, a set of data that does not spread roughly evenly above and below a regression line. Finding that the data is heteroscedastistic warns that the relation between independent variable X and dependent variable Y may not be linear and encourages checking for some other type of pattern (exponential, S-curve, etc).

Suppose we have data from a random sample of robbery 500 trials in Linno Province indicating that the mean of sentences imposed on persons convicted of robbery there is 27.3 months in jail while the mean robbery sentence in all of Prova is 28.7 months. The Minister of Justice wants to know whether Linno Province is slacking off. Statisticians in the Office of Crime Statistics can help out using standard routines for hypothesis testing. These involve five steps:

1. Checking assumptions about the data: they will first check to be sure that the Linno Province and national means were calculated on random samples, that they have interval or ratio level data (this is ratio data because it is possible to have a sentence of 0 months), and the sampling distribution of all possible sample means is normal in shape.

2. Stating the Null and Research Hypotheses. The Null Hypothesis is that the real Linno Province and national means are not different, and any difference seen at this time stems from random factors. The Research Hypothesis is that the two are different.

3. Establishing the basis of accepting or rejecting the null hypothesis. This involves selecting a confidence level, an odds that the null hypothesis is true. Many social scientists use a 5% chance of being wrong (0.05) as the cutoff; more demanding ones use a 1% chance. If using the 0.05 level, the "critical regions" for t are calculated by $\pm$ 1.96 of the population mean; if using the 0.01 level the "critical regions are calculated by $\pm$ 2.58 of the population mean.

4. Computing the test statistic. This involves knowing the sample mean, the sample standard deviation (s), the total number of units observed in the sample (N), and the population mean (μ) and using them in the following equation:

$$t = \frac{\text{sample mean} - \text{population mean}}{\text{sample standard deviation} / \sqrt{N-1}}$$

5. Determining whether to accept or reject the null hypothesis. The null hypothesis is accepted if t falls between the population mean and the edges of the "critical regions."

To determine whether Linno Province differs systematically (for some non-random reason) from Prova national level data, the Office of Crime Statistics calculates

$$t = \frac{27.3-28.7}{3.7 / \sqrt{500-1}} = \frac{-1.40}{3.7/22.3} = \frac{-1.40}{0.166} = -8.43$$

and then compares this to the "Z score" established by comparing the calculated value of the t test to the value associated with the critical region at the desired level of confidence. Here, the level does not matter because -8.43 is larger than either -1.96 or -2.58. Thus the Office can report that Linno Provinces's sentencing for robbery does differ from the national practice.

This result is a **two-tailed test** that calculates the likelihood of the sample mean being significantly higher or lower than the population mean. Suppose the Minister starts the discussion by expressing an opinion that authorities in Linno Province are being too lenient to robbers. The Office of Crime Statistics can use a modified form of the t-test in a one-tailed test that looks only at whether the sample mean is significantly higher or lower than the population mean. Here, the relevant "critical regions" are found at the 0.05 confidence level by using +1.29 for the upper tail and – 1.29 for the lower, and at the 0.01 confidence level by using + 2.33 or -2.33. Even on a one-tailed test inspired by the Minister's hunch, Linno Province's sentences are shown to be non-randomly different.

Remember that even when a statistical test of data supports concluding that the observed data confirms the research hypothesis, the statistical test does not tell you whether the relation is causal or important enough to be included in theories about political interaction. Hence another maxim for good political scientists and smart citizens to remember: **statistical significance is not the same as real-world significance**.