Cross Tabulation

Probably the simplest method of assessing the association between two, or more, variables (a basic part of hypothesis testing) is cross tabulation. Two examples of cross tabulation appear ahead. In the first example (Tables 1 and 2) we will be assessing whether a person's location (whether they live by the sea coast or live inland - the independent variable) is related to their degree of tolerance (i.e., how supportive an individual is of permitting people to express non-traditional opinions, lifestyles, etc., - the dependent variable - "high" tolerance means much willingness to tolerate such differences). Please note that the common convention in displaying cross tabulation tables is to percentage by the independent variable [e.g., in the tables below notice that the column percentages total 100% - for example, 45% + 55% = 100% - this is because the categories of the independent variable are going across the page - just keep reading] and to include the number of cases (observations) in parentheses.

Table 1
Tolerance by Location

Tolerance	Coastal	Inland
High	45% (180)	19% (97)
Low	55% (220)	81% (403)
	100% (400)	100% (500)

Each possible combination of responses in Table 1 is called a "cell" (just keep reading). For example, all those respondents who both live in a coastal area and are "high" in tolerance are placed in the same "cell." The number in the parentheses for this "cell" is 180. This means that there are 180 respondents who both live in a coastal area and are "high" in tolerance. Additionally, 45% of those living in a coastal area are "high" in tolerance (180 is 45% of 400). Table 1 contains four cells (i.e., coastal/high tolerance, coastal/low tolerance, inland/high tolerance and inland/low tolerance).

It seems that living in a coastal area and tolerance are associated (i.e., coastal residents are more tolerant than inland residents - because 45% is greater than 19%). However, is location the only influence on tolerance? If not, we could conclude that location influences tolerance when another independent variable is actually influencing tolerance. A second independent variable that could influence tolerance is education. Typically, the more educated one is, the more likely they are to be exposed to, and respect, the right of people to differ (whether by appearance, lifestyle or beliefs). Thus, let us "control" (i.e., remove the effect) of education and

see if location and tolerance are still related (how this is done is explained on the next page). Since we want to see if location is related to tolerance, we will remove the influence of education on tolerance (i.e., "control" for education) and see if location and tolerance are still related. To see the effect of education on tolerance, we would "control" for location and see if education is related to tolerance.

Table 2
Tolerance by
Location Controlling for Education

Tolerance	College Graduates		High School Graduates	
	Coastal	Inland	Coastal	Inland
High	57% (170)	57% (57)	10% (10)	10% (40)
Low	43% (130)	43% (43)	90% (90)	90% (360)
	100% (300)	100% (100)	100% (100)	100% (400)

Notice that within each category of education the percentage of those who are "high" on tolerance is the same (57% vs. 57% and 10% vs. 10% - just keep reading). Thus, among college graduates 57% are "high" in tolerance regardless of whether they live by the sea coast or live inland. Furthermore, among high school graduates only 10% are "high" in tolerance regardless of whether they live by the sea coast or live inland. Thus, within each category of education, location does not matter (i.e., within each category of education, there is no difference between those living by the sea coast and those living inland). The original relationship (i.e., Table 1 on page 30) occurred because most college graduates live in coastal areas (300 of the 400 college graduates live in coastal areas). Thus, when we simultaneously examine the effects of both education and location on tolerance, we find that education is related to tolerance (i.e., 57% of college graduates are high in tolerance while only 10% of high school graduates are high in tolerance), while location is not. Alternatively, we can say that the relationship between location and tolerance in Table 1 on page 30 is "spurious" (i.e., existed before the "control" variable - education - was included but disappeared once the "control" variable was accounted for).

Do not confuse "controlling" for an independent variable with "setting the level" of an independent variable. In a nonexperimental research design, the researcher cannot set the levels (i.e., scores) of the independent variables (see pages 5-7). We are using a nonexperimental research design in Table 2 above because we cannot either increase or decrease any respondent's level of education. For example, we cannot add four years of college to a high school graduate and then see if that person's level of tolerance changes. However, even though we cannot set (or change) the level of each person's education, we can "control" for education because we can examine various levels of education where each person has the same amount of education (e.g., each high school graduate has the same amount of

education) and then see if among those who have this same amount (or level) of education their location (coastal or inland) is related to their degree of tolerance. Thus, just because we cannot set the level of an independent variable (e.g., the person's education or location), we can still control for a particular independent variable (as we just did with education).

Tables 1 and 2 were inspired by pages 438-439 of Research Methods in the Social Sciences, third edition, by David Nachmias and Cava Nachmias.

Our second example of cross tabulation concerns the effect of gender (the independent variable) on the speed with which one is promoted at work (the dependent variable). We are trying to assess whether gender discrimination is occurring in the workplace.

Table 3
Year of Promotion by Gender

Year of Promotion	Men	Women
one year	33%	20%
after one year	67%	80%
	100%	100%
	(148)	(192)

There would appear to be discrimination by gender. Men seem to be promoted faster than women. However, as speed of promotion could be affected by many factors we would be more certain of gender-based discrimination if we "controlled" for these other factors. Obviously the table would become rather unwieldy if we tried to simultaneously control for more than one variable (this is a major limitation in using cross tabulation). If I were a lawyer representing women who had either not been promoted, or promoted later than most men, I think I would want to control for productivity so that my opposition could not make the case that the men who were promoted more rapidly were more "deserving."

Table 4
Year of Promotion by Gender Controlling for Productivity

Year of Promotion	High Productivity		Low Productivity	
	Men	Women	Men	Women
one year	37%	30%	28%	8%
after one year	63%	70%	72%	92%
	100%	100%	100%	100%
	(88)	(104)	(60)	(88)

Regardless of productivity men are promoted faster than women (37% is greater than 30% while 28% is greater than 8%). However, productivity is also related to speed of promotion (highly productive men are promoted faster than non-highly productive men - 37% is greater than 28% - the same pattern holds for women). Thus, unlike the previous example, the initial relationship between the independent and dependent variables holds after the control variable (i.e., productivity) is introduced. Furthermore, as the gap in the first year promotion rate is higher between highly productive and non-highly productive women (30% - 8% = 22%) than between highly productive and non-highly productive men (37% - 28% = 9%), productivity seems to matter more for women than men. The results say that you have to work harder to be promoted if you are a women. This appears to be a clear case of gender-based discrimination.

Tables 3 and 4 above were inspired by pages 146-156 and pages 170-171 of Quantitative Methods for Public Administration, 2nd edition, by Susan Welch and John Comer.

Measures of Association

It is often useful to have a summary statistic to show the association between variables. For example, a score of .19 on a measure of association can summarize much of the meaning of a many-celled cross tabulation table. For this reason, it is common for a cross tabulation table to also contain a measure of association. For reasons that will be discussed shortly, regression is by far the dominant analytical tool of modern quantitative political science. However, as you occasionally see measures of association in journal articles, you should be familiar with them. While there are many different measures of association, the only ones that you see with any frequency in political science are: gamma (symbol: γ), Kendall's tau_b (symbol: τ or tau_b) and Pearson's Product Moment Correlation (symbol: r). Pearson's Product Moment Correlation is usually referred to as either Pearson's r or just correlation. In the discussion that appears ahead, do not be concerned with "how" measures of association (i.e., gamma, Kendall's tau_b and Pearson's Product Moment Correlation) are calculated. Rather, be concerned with how measures of association are interpreted.

Suppose you are an international relations scholar and you are trying to see if a nation's political system influences its foreign policy. Specifically, your hypothesis is that since nation's with a democratic political structure are more likely than non-democratic nations to peacefully resolve conflicts within their own nation, they will also be more likely to peacefully resolve conflicts with foreign nations. Let us say that we have a data set of 715 international disputes from over the past 100 years. For each dispute we have scores for each of the nations involved concerning their level of democracy (a 10 point scale from "1" - least democratic, no elected offices no competing political parties, etc. to "10" - most democratic, high percentage of government officials are elected, at least two competing political parties, easy voter registration laws, etc.) and degree of peacefulness of conflict resolution (e.g., a 6 point scale from "1" - least peaceful, war is declared to "6" - most peaceful, no war, no threats of war, no break in diplomatic relations, etc.).

Our hypothesis would be that higher scores on degree of democracy are associated with higher scores on degree of peaceful resolution of conflict. Since there are 10 possible scores on degree of democracy and 6 possible scores on degree of peacefulness of conflict resolution, a cross tabulation table would have 60 cells [i.e., there a 60 possible combinations of scores (10 times 6 = 60) on the two variables - "1" on democracy and "1" on peaceful resolution of conflict is one combination, "1" on democracy and "2" on peaceful resolution of conflict is a second combination, etc.]. If you are confused, just look back at page 28. Didn't Table 1 have four cells because each variable had two categories (i.e., 2 times 2 = 4)? Yes! So, the number of cells is equal to the product (multiplication) of the number of categories of all the variables. Thus, if we have 10 categories on degree of democracy and 6 categories on the degree of peaceful resolution of conflict, then a cross tabulation table with these two variables would have 60 cells [i.e., 10 times 6 = (10) (6) = 60].

A cross tabulation table with 60 cells would be extremely cumbersome and difficult to interpret. Some would show this relationship by reducing the number of categories of the variables. For example, we could code scores on democracy as either "high"(a score from 7 to 10), "medium" (a score from 4 to 6) or "low" (a score from 1 to 3). This would reduce the number of cells from 60 [(10) (6) = 60] to 18 [since we now have only 3 categories on democracy and 6 on peacefulness of conflict resolution the number of cells is 18, i.e., (3) (6) = 18]. However, reducing the number of possible scores on a variable increases measurement error and denies the political scientist the knowledge that those extra categories provide. For example, assuming that the democracy scale was well constructed to begin with, there is a good reason why a nation was coded as scoring "7" rather than "10." However, if we use the "reduced category" approach that I just outlined, both "7" and "10" would be in the "high" democracy category. By treating "7" and "10" as the same score (they would both be considered "high" on democracy) we are less accurately measuring a potentially important variable. Thus, we are increasing the degree of measurement error. This is not desirable. Therefore, let us reject such an approach and use the full 10 categories for degree of democracy and 6 categories for degree of peaceful resolution of conflict.

We are still in the position of having a 60 celled cross tabulation table. In order to present the degree of association between a nation's level of democracy and the degree to which they resolve conflict peacefully in a more readily interpretable fashion, a political scientist might turn to a measure of association. For reasons I will discuss later, political scientists are increasingly moving away from either a cross tabulation table or a measure of association. But for now, let us assume the political scientist opts for a measure of association. While not as useful as the approaches we will later study, a measure of association is more desirable in our current situation than a 60 celled cross tabulation table.

Which measure of association do we use? The choice of a measure of association is largely governed by the level of measurement of the variables we are examining (on levels of measurement review pages 11-12). For example, both gamma and Kendall's tau$_b$ require that our data be at least measured at the ordinal level (i.e., either ordinal, interval or ratio, but not nominal because it does not possess the "ranking" quality that is necessary here, again, see pages 11-12). Both

our variables are probably best thought of as ordinal level measures. Let us examine the democracy variable. We can rank scores from "lowest" to "highest" on democracy. Therefore, democracy is measured at either the ordinal or interval levels. However, the differences between the categories of democracy are not likely to be equal. For example, is the difference between level "2" and level "3" the same as between level "5" and level "6"? Probably not. Therefore, the democracy variable is probably best classified as an ordinal level measure. For the same reasons, the peaceful resolution of conflict variable is also probably best classified as ordinal. Thus, we are trying to see if two ordinal level measures are associated with each other. Since Pearson's r (i.e., Pearson's Product Moment Correlation) assumes that variables are either interval or ratio (i.e., that there is an equal interval between categories), it should not be used with either nominal or ordinal level data (see pages 11-12). However, since both gamma and Kendall's tau_b are designed for ordinal level data, we could use either measure. Gamma will either be the same, or higher, than Kendall's tau_b. Typically, the differences are not great. For example, a score on Gamma might be .29 whereas the figure for Kendall's tau_b might be .22. While one can make a rather convincing case that Kendall's tau_b is preferable to gamma, political scientists are more likely to use gamma. So, let us select gamma. Thus, we ask the computer to calculate the gamma between level of democracy and degree of peaceful resolution of conflict for our 715 observations. As I previously, do not be concerned about the formula and computations the computer uses to calculate gamma. Be concerned with how we interpret gamma. Suppose the computer tells us that gamma is .55. What would this allow us to say?

Interpreting Measures of Association

Gamma, Kendall's tau_b and Pearson's Product Moment Correlation all show both the direction and strength of the association between two variables. All three measures range from +1.0 (strongest positive association) to -1.0 (strongest negative association), with .00 indicating no association. Since the gamma in this example is .55 (and not -.55) we know that there is an association (i.e., the gamma was not .00, or something very close to it) and that the association between degree of democracy and degree of peaceful resolution of conflict is positive. Thus, the more democratic the nation (i.e., the higher a nation's score on democracy) the more peacefully that nation resolves its disputes with foreign nations (i.e., the higher the score on peaceful resolution of conflict). Since we hypothesized a positive relationship, the gamma of .55 supports our hypothesis.

Be sure not to confuse the direction of the association with the strength of the association. For example, a gamma of .55 and -.55 have the same strength, only the direction of the relationships differ. As the above example demonstrates, a positive association means that higher scores on one variable are associated with higher scores on the other variable. However, a gamma of -.55 would indicate that higher scores on democracy were associated with lower scores on peaceful resolution of conflict.

While we now know that the relationship between degree of democracy and degree of peaceful resolution of disputes is positive, we do not know how "strong" this relationship is. In order to interpret the "strength" of the association, we first

need to discuss random measurement error. "Random" means that there is no pattern. For example, say that we had not perfectly measured the level of democracy of the nations involved in a particular dispute. In those cases in which the measure was not correct, let us say that we almost always overstated the degree of democracy (i.e., the score on democracy was invariably higher - closer to 10 -than it should have been). This would be a case of systematic (i.e., non-random) measurement error. Alternatively, if we are as likely to record a nation's score on democracy as being too low as too high, we have a case of random measurement error. For this discussion I am going to deal with random measurement error.

Random measurement error reduces the association between variables. Suppose we had two variables that were measured without error and were perfectly associated with each other (e.g., a gamma of 1.0). If we then introduced random measurement error into one of the variables, the association would be weakened (e.g., from 1.0 to say .70). This is why in the necessary strength of association is lower for variables measured with a "high" degree of random error than variables measured with a "low" degree of random error. The greater the random measurement error the more difficult it is to attain a strong association. As the following diagram indicates, if our variables are measured with a low degree of random measurement error, a gamma of .55 between degree of democracy and degree of peaceful resolution of conflict would constitute a "strong" positive association. Let me mention that the example I have been using, the relationship between a nation's level of democracy and its likelihood of resolving of peacefully resolving disputes with foreign nations has been extensively tested by quantitative international relations scholars. In general, their results are consistent with the hypothetical results I have used.

The degree of democracy measure would probably be best classified as having a "low" degree of random measurement error. By contrast, survey data often has a "high" degree of random measurement error. For example, when we ask voters about their political philosophy (e.g., conservative, moderate, liberal, etc.) their responses are likely to contain a "high" degree of random measurement error. This is because a concept such as "conservatism" has different meanings to different individuals. We can still learn much about voters from asking them about their political philosophy, but we need to be aware that such a measure is likely to have a "high" degree of random measurement error. The practical effect of working with a variable that has a "high" degree of random measurement error is that it is more difficult for us to achieve a "strong" association (e.g., a gamma of .70). The following table provides a guide for interpreting measures of association in relation to the degree of random measurement error.

A Standard to Interpret the Strength of Gamma, Kendall's Tau_b and Pearson's Product Moment Correlation

Variables Containing a High Degree of Random Measurement Error:

plus/minus .01 to plus/minus .15 - weak association
plus/minus .16 to plus/minus .29 - moderate association
plus/minus .30 to plus/minus .49 - strong association
above plus/minus .49 - very strong association

Variables Containing a Low Degree of Random Measurement Error:

plus/minus .01 to plus/minus .25 - weak association
plus/minus .26 to plus/minus .49 - moderate association
plus/minus .50 to plus/minus .69 - strong association
above plus/minus .69 - very strong association

The Changing Nature of Statistical Analysis in Political Science

Since the late 1970s there has been a sharp decline in the use of cross tabulation and measures of association in both political science and the social sciences generally. This trend has occurred for three primary reasons.

First, cross tabulation tables (but not measures of association) almost force the researcher to work with either a small number of variables and/or a small number of categories per variable (just keep reading). For example, using just four variables (e.g., the dependent variable and three independent variables) with only four categories of responses per variable would produce a cross tabulation table containing 256 cells (4 x 4 x 4 x 4 = 256). By contrast, Table 1 on page 30 has only four cells. A table with 256 cells would take several pages to display and would be extremely difficult to interpret. This is why users of cross tabulation typically include only one, or two, independent variables. As you saw on pages 30-32, the relationship between one independent variable and the dependent variable can change considerably if another independent variables is included. By greatly limiting the number of independent variables we can use, cross tabulation is highly likely to produce misleading results. Furthermore, by virtually forcing us to use few categories of responses per variable, cross tabulation considerably increases measurement error. In the example on page 30, we measured an individual's tolerance as being either "high" or "low." All those listed as "high" in tolerance probably do not have the same degree of tolerance. More categories of responses (e.g., ten categories of responses on tolerance instead of just two) would have produced a more valid measure. Similarly, if one of our variables is a percentage, it would have 101 categories of responses (0 plus 1-100). As 101 categories of responses would produce a gargantuan cross tabulation table, users of cross tabulation will typically reduce the 101 categories to, say, 3 categories: 0-33, 34-66

and 67-100. Such a procedure would put a score of 1 in the same category as a score of 33 (i.e., they would both be in the 0-33 category). However, assuming a valid measurement scale, a score of 1 is quite different than a score of 33. Therefore, the practical difficulties of using cross tabulation are highly likely to increase measurement error and produce misleading results.

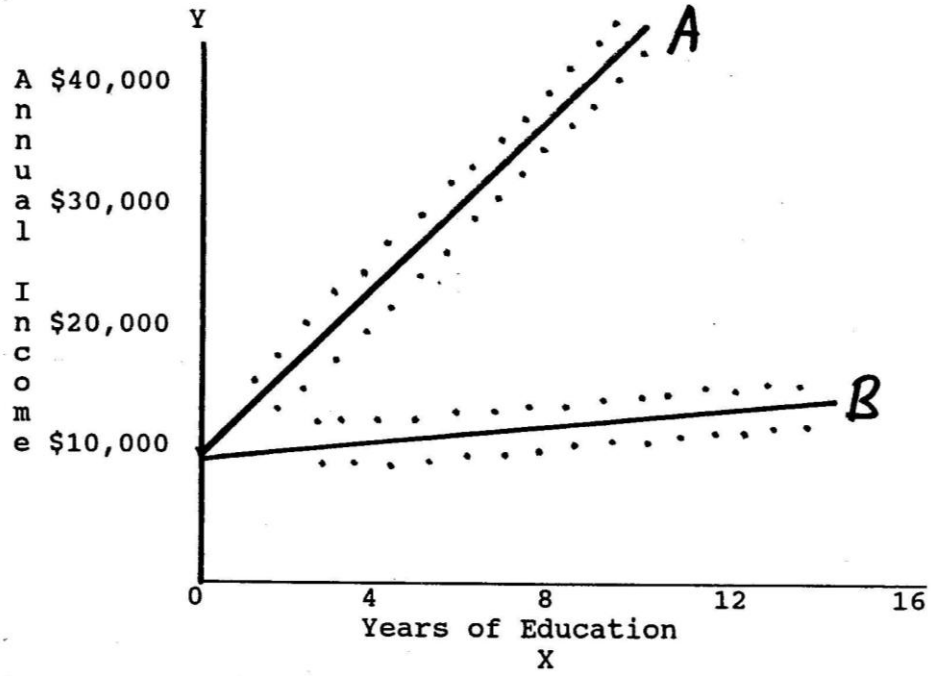
Second, even if we use a cross tabulation table with many variables and many categories of responses per variable the small number of observations in many of the cells make significance testing (your next statistically oriented reading assignment) extremely risky (just keep reading). Suppose our sample contains only one conservative female blue collar worker who is also 30 years old, a Democrat and has an annual income of over $100,000. We would have a "cell" with only one observation. This is simply too few observations for reliable analysis. This is analogous to flipping a coin one time and concluding the coin is biased toward flipping heads. We would desire many more flips of a coin before being very confident that a coin is biased. Each time we add another category in a cross tabulation table, in effect we reduce the number of observations. Look again at Table 1 on page 28. There are only 97 respondents who both live in an "inland" area and are "high" on tolerance. In an effort to improve the accuracy of our measures, suppose we created additional categories of "inland." After all, are all "inland" areas the same? Probably not. Supposing, for example, that we place those living from 5 to 50 miles from the ocean and being "high" on tolerance in one category, those living from 51 to 200 miles from the ocean and being "high" on tolerance in a second category and those living more than 200 miles from the ocean and being "high" on tolerance in a third category. From a measurement standpoint, this is an improvement upon just lumping all three categories together in one cell (i.e., being both "inland" and "high" on tolerance). Assuming that there are some respondents in each of these three new categories, some of the resulting new cells will contain far fewer than the 97 respondents who were listed as being both "inland" and "high" on tolerance. The total number of respondents in all three cells would be 97 but each cell would have fewer than 97. For example, maybe there are only 10 respondents who both live from 5 to 50 miles from the ocean and are "high" on tolerance. If we continue this process very long, we will end up with cells that contain very few respondents. A low number of respondents in a cell means it is difficult to be very confident about the results. What we need is a method that estimates the relationship between variables while preserving the size of our sample. In the current example this would mean holding the sample at 97 and not reducing it 10 through more categories.

Third, neither cross tabulation nor measures of association provide us with a measure of the magnitude of the relationship between the variables (just keep reading). For example, a Pearson's Product Moment Correlation of .7 between an individual's level of education and their annual (yearly) income indicates that the relationship between education and annual income is both "strong" and "positive". Therefore, we know that higher levels of education are associated with higher annual incomes. However, neither Pearson's Product Moment Correlation nor any other measure of association (e.g., gamma or Kendall's tau$_b$) can tell us whether each additional year of education is associated with an expected annual increase in income of $50, $500, $5,000, $50,000, or any other amount. If you were contemplating spending thousands of dollars to obtain an advanced degree, I think

you would want to have a clear idea of how much your annual income might be expected to increase. As one of the central tasks of a quantitative analysis is to estimate the amount (i.e., the magnitude) of change in the dependent variable associated with a specified amount of change in the independent variable(s), such a failure is a critical limitation.

Perhaps a diagram would help make the previous point clearer. The "strength" of the association between variables X and Y is measured by how close the points are to a line drawn to fit them. In the diagram ahead there are two lines. Since the points surrounding each line are equally close to the line they surround, each line would represent a correlation of identical strength (for example, .70). Alternatively, the steepness of the line is the "magnitude" of the association between variables X and Y. The steepness of the line (i.e., the magnitude) tells us how many units of change in Y occur for a particular amount of change in X (e.g., how many dollars of additional annual income you can expect to earn (continued on next page)

for each additional year of education). Clearly, if line "A" is correct, variable X (years of education) has much more impact on variable Y (annual income) than if line "B" is the correct line. Since both lines are consistent with the same strength of association (for example, a Pearson's Product Moment Correlation of .70), the strength of the association does not tell us the steepness of the line (i.e., the magnitude). Remember from page 5 that one of the two goals of a quantitative analysis is to estimate the magnitude of the association between the variables. Since both regression and logit tell us the magnitude of the association between the variables the major emphasis of this course will be on them.



Rather than read the boring assignments for this course, suppose you decide to go to Las Vegas. Let us say that you walk into a casino and just feel "lucky." Instead of playing one of the various games, you locate the floor manager and tell him that you think you will flip heads on each of the first ten tosses of a coin. The floor manager might then ask you what odds you would want and how much you would be willing to wager. You are so confident that you reply that you would expect to be paid ten dollars for every one dollar that you wager and you would be willing to wager up to $1,000. If the floor manager knew much about statistics he would likely accept your offer. The actual probability of tossing ten consecutive heads with a fair coin is slightly less than one in a thousand ($.5^{10}$ = approximately .001). Thus, if the coin is fair, you would lose this bet slightly more than 999 times out of 1,000. You were very generous to accept ten to one odds. Perhaps you should have taken this course before wagering! In any event, the floor manager accepts your terms and wants you to wager $1,000. You nod in agreement, pull out a coin and start tossing. After tossing ten consecutive heads you expect to be paid. However, before the floor manager pays off, he wants to toss the coin himself. In effect, he challenges your belief that the coin is fair. Fortunately, this is an easy request to honor. The floor manager tosses the coin 200 times and heads come up 100 times. So, he concedes that the coin was fair and pays you. By doing so, he is admitting that we have just witnessed an extremely rare event.

Two points in the preceding example are important for our purposes. First, notice that a fair coin could come up heads ten consecutive times. It was just extremely unlikely (but possible). Thus if a survey tells us that the Republican candidate is supported by more potential voters than the Democratic candidate, it is not impossible that there is actually either no difference in their support levels, or that the Democratic candidate is leading. Given our results, it was just more likely that the Republican candidate was ahead. Second, in the coin tossing example, we were able to repeat the experiment. While we will never know the proportion of heads the coin would ultimately produce, the ease with which we could continue to toss the coin meant that it was possible to obtain a large number of tosses. Thus, the floor manager could be quite certain (but never know for sure) that the coin was actually fair. Unfortunately, in most situations a political scientist will not be able to replicate the study. To liken this to the coin tossing example, it would mean that when the floor manager challenged the fairness of the coin, we would not have been able to continue tossing the coin. Thus, the only results we usually have are those from the observations (coin tosses) we could originally obtain. There would be no additional information. Therefore, we would have had to make a judgment about the fairness of the coin with only a few tosses. Given that the probability of a fair coin producing ten heads in ten tosses was .001, our best guess would have been that the coin was unfair. Obviously, we would have been incorrect, but that would have been the most reasonable conclusion given the actual probability and the behavior of the coin over those ten tosses.

The coin tossing example, and the ensuing discussion, deal with one of the most important topics of this course, statistical inference. Two of the most important concepts in statistical inference are a population and a sample. A population consists of all the possible observations on the same unit of analysis (e.g., a person, a city, a nation, etc.) having a particular attribute in common (e.g., being an eligible voter in the United States). A sample is a subset of the population.

A sample is "random" if every member of the population has an equal chance of being selected. Statistical inference is important to study because we almost never know the population result. Hence, we almost invariably infer the population result from a sample. As one might guess, sampling becomes an important topic because the more accurately our sample represents the population, the more accurate our inferences are likely to be.

To continue the coin tossing example for a moment. If possible, we would like to know whether, or not, the coin was fair. As the coin does not wear out, it could be tossed an infinite number of times. This is what is termed an "infinite population." Hence, we could never know for certain whether the coin was actually fair. So, we "infer" what the ultimate (or "population") probability of tossing a head with this coin on the basis of a sample of tosses. The fundamental question of statistical inference is: How likely are the results to be the product of chance? Applied to the coin tossing example, this question could be phrased as follows: How likely would a fair coin flip ten heads in ten tosses? As we know, the probability is less than .001. Therefore, we conclude that the coin is probably unfair. Given what happened in subsequent tosses of the coin, we realize that such a judgment would probably be incorrect (although we are not certain).

Inferring from a sample (the 200 coin tosses) to a population value (the "true" probability of tossing a head for this particular coin) is the process of statistical inference. As you read the following pages try to keep the fundamental question of statistical inference in mind. See how the readings help us answer this question. In the pages immediately ahead, we have a "population" of only ten families. Since we know the income of all ten families we can calculate the "true" population mean income. We then draw samples of two families each and calculate the mean income for each of these samples. In all, there are 45 possible samples. Do not be concerned with how we know there are 45 different samples. That would needlessly detain us. Just take it on faith. The importance of the example is that since we know the "true" population value (i.e., the "true" mean income of the ten families), we can see how the sample estimates of the mean income (i.e., the mean income from each of our two family samples) vary around the "true" population mean. In this way, we can see how close our estimates (each sample mean is one "estimate" of the "true" population mean) are to the actual value we are trying to estimate (i.e., the "true" population mean). We can use this information to assess how far off our estimates are likely to be when we do not know the "true" value in the population (e.g., the "true" popularity of a president among 180 million potential voters). Since we almost never know the "true" population value, assessing the accuracy of our "estimate" is critical.

Assume we were interested in the income levels of the parents of children participating in a free breakfast program. For simplicity's sake let us assume we have a population of 10 children with their parents' incomes as follows: $3,000, $4,000, $5,000, $6,000, $7,000, $8,000, $9,000, $10,000, $11,000 and $12,000. The mean income of these ten families is $7,500 (because $3,000 + $4,000 + $5,000 + $6,000 + $7,000 + $8,000 + $9,000 + $10,000 + $11,000 + $12,000 = $75,000 and $75,000/10 = $7,500 (example from Research Methods in the Social Sciences, third edition, by David Nachmias and Cava Nachmias). Suppose we tried to estimate the population mean (which we now know is $7,500) by drawing a sample of two families from our population of 10 families. The lowest possible estimate of the mean income

we could attain by choosing two of the ten families would be $3,500 (the lowest two family incomes were $3,000 and $4,000 which, when added, total $7,000 and $7,000/2 = $3,500). Similarly, the highest possible estimate (by taking a sample of two families) is $11,500 ($11,000 + $12,000 = $23,000 and $23,000/2 = $11,500). In each instance our sample estimate was either $4,000 lower, or $4,000 higher, than the "true" mean of $7,500 ($3,500 - $7,500 = -$4,000 and $11,500 - $7,500 = $4,000). Any other possible sample (i.e., picking any two incomes other than the two lowest or the two highest) would have produced an estimate that was less than $4,000 away from the "true" population mean of $7,500. In all, 45 different samples of two could be drawn from these 10 family incomes (i.e., $3,000 + $4,000 is one sample, $3,000 + $5,000 is a second sample, $3,000 + $6,000 is a third sample, and so on). The important question is: How do these 45 sample estimates of the mean income distribute themselves around the "true" mean income of the population (i.e., $7,500)? The sample estimates will be distributed closely to the normal distribution that we studied previously. For example, the sample means that occur the most frequently are those closest to the "true" population mean of $7,500. For example, 5 of the 45 possible samples have the same mean as the population (i.e., $7,500). While the next sentence may be difficult to understand, just keep reading (it will become clearer as we proceed). Second, the mean of the sample means is the same value as the population mean (i.e., $7,500). Thus, as there are 45 different samples, there are also 45 sample means (i.e., we can calculate a mean from each sample). If we add up these 45 sample means and then divide this total by 45 (remember, to calculate a mean we add up the scores and then divide by the number of scores we added) the resulting "mean of the sample means" will equal the population mean (which we know is $7,500). Third, the sample means that are furthest from the "true mean" (i.e., $3,500 and $11,500 are the furthest from $7,500 of any possible sample means) are the sample means least likely to occur (i.e., only one of the 45 samples has a mean of $3,500 and only one sample has a mean of $11,500). The closer to the "true mean" the sample mean is the more likely it is to occur. Since $6,000 is closer to $7,500 than $3,500, more samples have a mean of $6,000 than a mean of $3,500.

In the previous example we had such a small population (10 families) that we could actually know the income of each family in the population. Thus, we could calculate the "true" population mean (i.e., add up the income of all 10 families and divide this total by 10). However, typically a political scientist is working with such a large population that they can not possibly obtain a score for each member of the population. For example, if a political scientist is studying the impact of a government policy on the income of American families, s/he could not possibly find out the income of each American family. Therefore, a political scientist must sample from the population of interest. A very important question then becomes: How representative is our sample of the population it was drawn from? The importance of our previous example was that since we could know the "true" population mean ($7,500) and also draw samples from this population, we could assess how close the sample means were to the "true" population mean. While it is possible that the mean from any one sample of two families could be as much as $4,000 lower or higher than the "true" population mean of $7,500 (i.e., the sample mean could be as low as $3,500 or as high as $11,500), typically, the sample mean is fairly close to the "true" population mean. Most of the sample means are within approximately $1,500 of the population mean of $7,500 (i.e., most of the area under the curve is between $6,000

and $9,000).

Since a political scientist can usually only draw one sample, let us see if we can assess how accurate this one sample we draw is likely to be. As a political scientist typically has a sample size of more than 30, they usually are working with what statisticians call "large" sample properties. Make sure you do not confuse the size of the sample with the number of samples taken. What I just said was that a political scientist typically has more than 30 observations in the one sample that they are able to study. This might mean having the income of each of 30 families. Such a situation would be one sample of size 30, not 30 samples.

Instead of the low income families of the school breakfast program, suppose we draw a random sample (i.e., every member of the population has an equal chance of being selected) of 100 families from the approximately 180 million American families and find that the mean income for this sample is $37,000. Remember from pages 39-40 that the standard deviation shows how far the scores deviate (i.e., differ) from the mean. Applying the formula and computations that we did when we examined the standard deviation, suppose we find that the standard deviation of our sample is $7,000. This would tell us that incomes varied considerably among these 100 families because the standard deviation is approximately 19% of the mean (i.e., $7,000 is approximately 19% of $37,000). Thus, the sample mean income of $37,000 did not occur because most every family in the sample earned approximately $37,000.

Now, we are in a position to answer the question I posed before: How representative is our sample of the population? The next few sentences are likely to be confusing. As always, just keep reading! Over the next several paragraphs, the discussion will start to make sense. Just keep reading! From our discussion of the normal curve we know that if we have a normal distribution, approximately 68% of the cases (i.e., in this instance family incomes) are within (i.e., plus or minus) 1 standard deviation of the mean and approximately 95% of the cases are within 2 standard deviations of the mean. Let us assume that the scores in our sample are normally distributed. Since the sample mean is $37,000 and the sample standard deviation is $7,000, approximately 68% of the families should have incomes between $30,000 and $44,000 (i.e., $37,000 - $7,000 = $30,000 and $37,000 + $7,000 = $44,000). Furthermore, approximately 95% of the families should have incomes between $23,000 and $51,000 (i.e., $37,000 - $7,000 - $7,000 = $23,000 and $37,000 + $7,000 + $7,000 = $51,000).

If we make one simple change, we can apply the information we have from our sample (i.e., the mean and the standard deviation) to estimate how representative our sample is of the population. Our sample mean is the mean income of the 100 family incomes that we randomly selected from the approximately 180 million American families. The population mean income is the mean income of all 180 million American families. Our question is: How close is our sample mean of $37,000 likely to be to the "true" population mean of the 180 million American families? Since we do not know the standard deviation of family income for the 180 million American families, we have to estimate it from our sample. We already know that the standard deviation in our sample is $7,000. The next sentence will be confusing. Just keep reading! Let us divide this sample standard deviation by the square root of the sample size minus 1. Since our sample size is 100, the sample size - 1 is 99 (i.e., 100 - 1 = 99). The square root of 99 is 9.94 (because 9.94 times 9.94 is approximately

equal to 99). If we then divide the sample standard deviation by the square root of the sample size - 1 we have $704.2 (i.e., $7,000/9.94 = $704.2). Do not be concerned with either "why" we needed to make the above "adjustment" to the sample standard deviation or "how" the formula for this "adjustment" was derived. That would needlessly detain us and not be particularly insightful. Just follow the discussion ahead to see "what" this "adjustment" will permit us to do.

Since we have a large sample (i.e., a sample size of over 30 - our sample size is 100, easily larger than 30), we can use the percentage distribution capabilities of the normal curve to see how closely our sample mean corresponds to the population mean. That last sentence was long and difficult, let us apply it. The "adjusted" sample standard deviation (i.e., $704.2) can be used to show how accurate our sample mean income (i.e., $37,000) is of the population mean income of all 180 million American families.

If we have a normal distribution, approximately 68% of the cases (i.e., in this instance family incomes) are within (i.e., plus or minus) 1 standard deviation of the mean and approximately 95% of the cases are within 2 standard deviations of the mean. The next sentence may be confusing, just keep reading!!! Using the sample mean income, $7,000, and the "adjusted" sample standard deviation computed on page 43, $704.2, we can say that our estimate of the population mean, $37,000, is accurate within plus or minus $704.2, approximately 68% of the time. Thus, we have approximately a 68% probability that the "true" population mean is within $704.2 (plus or minus) of our sample estimate of $37,000. In other words, given our sample estimate of $37,000, a sample size of 100, and an "adjusted" standard deviation of $704.2, there is approximately a 68% chance that the "true" mean income in the population of 180 million American families is between $36,296 ($37,000 - $704.2 = approximately $36,296) and $37,704 ($37,000 + $704.2 = approximately $37,704). Furthermore, we can say that there is approximately a 95% probability that the "true" population mean income of the 180 million American families is between $35,592 ($37,000 - $704.2 - $704.2 = approximately $35,592) and $38,408 ($37,000 + $704.2 + $704.2 = approximately $38,408). Equivalently, we can say that if we drew 100 random samples of 100 persons each, the mean income from approximately 95 of these 100 samples would be between $35,592 and $38,408. Political scientists typically say that interval from $35,592 to $38,408 represents a 95% "confidence interval." Thus, given these results, we would be "95% confident" that the "true" mean income of the 180 million American families (the population of interest) was between $35,592 and $38,408.

Remember that the only information we have is the 100 family incomes from our one sample. The above example demonstrates a critically important statistical property: we can tell how possible sample means would vary from each other (e.g., 95% of the samples of size 100 would have a mean between $35,592 and $38,408) even though we can actually obtain data from only one sample. While proving this assertion is beyond the scope of this course, the school breakfast example provided good evidence of this capability. Since the "population" was so small (10 families) we could obtain the family income for all members of the population, draw samples from this population, and then see how the sample estimates of the mean family income differed from the "true" population mean (which we knew to be $7,500). Statisticians employing powerful computer programs have used the same procedures we did with much larger populations and have proven the assertion I

made above. Since a political scientist typically has information (i.e., data) from only one sample, it is extremely fortunate that we can know how other samples that we can not actually attain would likely vary (i.e., differ) from the one sample that we have.

How does the accuracy of the estimate of the population mean vary according to the size of the sample? The larger the sample the closer the sample mean is likely to be to the "true" population mean. For example, if the size of our random sample had been 1,000, the 95% "confidence interval" would have been from $36,557 to $37,443 (i.e., minus or plus $443 from $37,000). The 95% "confidence interval" from the 1,000 person sample ($36,557 to $37,443) is considerably "narrower" than the 95% "confidence interval" from the 100 person sample ($35,592 to $38,408). The "narrower" the 95% "confidence interval," the closer the typical sample mean is likely to be to the "true" population mean.

You have probably seen polling results reported on either television and/or in the newspaper. Let us use the presidential popularity question that pollsters typically ask: Do you believe President (then the last name of the current president) is doing a good job as president? Typically, respondents can answer "yes," "no" or "no opinion/decline to state." Suppose a political scientist is trying to obtain a random sample from Long Beach voters to estimate the president's popularity in Long Beach. The next sentence may be confusing, just keep reading! An important question would be: How large a random sample do I need to be 95% confident that my sample results are within say plus or minus 3% of the "true" figure for the city of Long Beach? Thus, if 57% of the respondents in my randomly drawn sample of the eligible voters in Long Beach think that the president is doing a good job, how large would my sample need to be in order for me to conclude that there is a 95% probability that the president's popularity among all eligible voters in Long Beach is between 54% and 60% (i.e., within minus or plus 3% of 57%)? Assuming that the president's popularity is "around" 50% (which is not that different from 57%), Table 7-2 below tells me that since Long Beach has a population between 100,000 and 500,000, I would need a sample of approximately 1,056 respondents to be 95% confident that my sample results were within minus or plus 3% of the "true" support level for the president among the eligible voters of Long Beach.

Sample Size Necessary for 95 Percent Confidence

Size of Population	+/- 1 percent	+/- 3 percent
2,000	Entire Population	696
100,000	8,763	1,056
500,000 +	9,423	1,065

Source: Adapted from H.P. Hill, J.L. Roth and H. Arkin, Sampling in Auditing

Please note that in the above example the "population" of interest is not all citizens of Long Beach, but rather all eligible voters of Long Beach. Since children can not vote, they are not part of the "politically relevant" population. Remember that the "population of interest" is composed of all those who share some particular characteristic (i.e., being an eligible voter in Long Beach). This is not the same as all people in Long Beach. Remember also that a population can be something other than people. For example, a population of coin flips, states (not the people in the states), wars between nations (again, not the people in the nations), etc.

Further note that for a population of 500,000 or more (e.g., the entire adult U.S. population), you need only 9 more respondents than for a population of 100,000 (1,065 instead of 1,056) to have a 95% probability that our estimate is within 3% (plus or minus) of the "true" value (i.e, a 3% error margin). To repeat our previous example, if we randomly surveyed 1,056 adult residents of Long Beach and found that 57% of them approved of how the president was handling his job we would have a 95% probability that our estimate was within 3 percent of the true popularity of the president in Long Beach. Thus, we have a 95% chance that the president's true popularity in Long Beach is between 54% and 60% (i.e., minus or plus 3% from 57%), with our best estimate being that it is 57%. Remember that this means that there is also a 5% chance that the president's true popularity in Long Beach is not between 54% and 60% (i.e., either lower than 54% or higher than 60%).

To achieve the same accuracy for the entire adult U.S. population we would need to randomly survey approximately 1,065 respondents. This is only 9 more people than our Long Beach survey of 1,056. However, we could not just "add" 9 respondents to our random sample from Long Beach and accurately generalize to the entire U.S. adult population. Obviously such a sample would not even approach randomness (over 99% of the sample would be from Long Beach while Long Beach represents less than 2/10s of 1 percent of the U.S. population). Nevertheless, despite the fact that the entire adult U.S. population is many times larger than the adult population of Long Beach, the necessary sample size (assuming it is randomly drawn) is almost identical (1,065 vs. 1,056). Also, notice that for a population of only 2,000 you would need a random sample of 696 to have a 95% chance of having an estimate that is within plus/minus 3% of the true figure. This would mean that the sample would be approximately 35% of the size of the population (696 is approximately 35% of 2,000). To achieve the same accuracy for a population of 500,000 would require that the sample be approximately .2% (two tenths of one percent) of the population. This illustrates an important aspect of sampling. It is the absolute size of the sample, not the sample as a percentage of the population that is the critical factor. With a random sample of approximately 1,100 people we can fairly accurately generalize to about any size population.

The next time you see a national poll on television or in the newspaper, notice in the "fine print" that the sample size will usually be approximately 1,100 and that it will have a 95% probability of a plus/minus 3% error margin. The main reason that most pollsters do not strive for a lower error margin than plus/minus 3% is the cost. Notice in the Table page 47 that for a population of 500,000 (or more) in order to lower the error margin from plus/minus 3 percent to plus/minus 1 percent would require an increase in the sample size from 1,065 to 9,423. The increased precision is simply not worth the additional cost.

Statistical Inference and Hypothesis Testing

The importance of sampling is that it allows us to estimate values in the population of interest (e.g., the mean score in the population of interest). This is particularly important when we try to test a hypothesis. The strategy by which we test a hypothesis is called a research design. A good research design is one that eliminates plausible alternative explanations (i.e., alternative to the independent variable) for the effect, if any, that is being observed on the dependent variable. One alternative is simply chance, since samples will vary from their population by chance alone, as we have seen. For example, in the school breakfast example, not every sample had the same mean family income. Procedures for establishing statistical significance are a way to define the likelihood of chance as an explanation when randomness can be assumed, such as when observations have been selected at random. Just keep reading!! The following example was inspired by Susan Welch and John C. Comer, Quantitative Methods for Public Administration, 2nd ed., pp. 48-52.

Since many of you are interested in public law, let us use an example that a lawyer might face: jury selection. In a community that is 50 percent women and 50% men, what is the likelihood that no men will serve on a particular jury? We will make the following assumptions: (1) the jury is composed of 12 people; (2) the selection of each juror is an independent event (i.e., that choosing any one person does not affect the chance of any other particular person being selected – thus, if a spouse is selected it would not be an independent event because the second person was selected because they were married to the first person selected); and (3) the city has an equal number of women and men.

With the assumptions above, what is the probability of having a jury entirely composed of women? Without doing the math, it is approximately .0002 (i.e., only 2 times in 10,000 would this occur by chance). Thus, the laws of probability tell us that in only 2 times out of 10,000 (.0002) would a jury be all women (or all men) if random selection were used to pick 12 jurors from a population that was 50 percent women and 50 percent men. A critically important result is that an evenly divided jury (i.e., 6 women and 6 men) would occur only about 23% of the time. Therefore, we can expect to have an unequal jury selected (i.e., either more women than men or vice versa), even though the selection process was fair, over 75% of the time. So, a reasonable question might be as follows: how much of a departure from a 6 women, 6 men jury will we accept before we think the jury selection process is biased in favor of either women or men? For example, would an 8 woman, 4 man jury be insufficiently different than 6 women and 6 men, or if we obtain an 8 woman, 4 man jury should we reselect the jury on the basis that the selection process wasn't fair?

Having a full list of the probabilities would be useful, so let me provide it: 12 women – 0 man (.0002); 11 women – 1 man (.0029 or about 3 times in 1,000); 10 women – 2 men (.016 or about 1.5%); 9 women – 3 men (.0537 or about 5%); 8 women – 4 men (.1208 or about 12%); 7 women – 5 men (.1934 or about 19%) and 6 women - 6 men (.2256 or about 23%). Since women and men are an equal percentage of the population is this particular city, the probabilities for majority male juries are the same as for majority female juries (i.e., the probability of 12 men – 0 women is .0002).

What is termed the "null" hypothesis is a hypothesis of no effect. For example, a null hypothesis would be that the balance of power between two nations

has no impact on the probability those nations will go to war with each other. Thus, if the null hypothesis is true, if nation A had 1.5 times the military power of nation B and this ratio suddenly changed to 2.0 (i.e., nation A now had twice the military power of nation B) the probability that war would breakout between these two nations would be unchanged.

Applied to our jury selection example, the null hypothesis is that the jury selection process is "fair" (i.e., unbiased) and that any deviations from a 6 woman, 6 man jury is strictly the result of chance (i.e., like a "fair" coin coming up "heads" 6 straight time rather than 3 heads and 3 tails). Remember that the long run probability may not occur in the short run. This is exactly what a gambler is counting on: that over the series of bets that they make they will win more frequently than the laws of probability say they should (e.g., if they are betting on "heads" that even though the coin is "fair" it will flip more than 5 heads in the next 10 flips). Thus, in our jury selection example the question is this: if there is an unequal number of women and men selected to the jury, did this occur because the jury selection process was unbiased or was the selection process biased in favor of the gender that is a majority of the jury?

To help answer this question statisticians refer to what is called the "region of rejection." The region of rejection is a group of outcomes that are so different from what the null hypothesis predicts that we conclude that the null hypothesis is probably false (although we do not know for sure - there is still a small chance the null hypothesis is true). In the jury selection example there is less than a 10% chance (the actual figure is 7.2%) that the jury selection process is unbiased if 3 or fewer women are selected (i.e., the probability of 0 women is .0002, 1 woman is .029, 2 women is .015 and 3 women .053: .0002 + .0029 + .016 + .0537 = .0721 = 7.2%).

If we are willing to run a 10% chance of rejecting the null hypothesis that the jury selection process is unbiased in favor of the alternative hypothesis that the jury selection process is biased when in fact the jury selection process is unbiased, then we would reject the null hypothesis if 3 or fewer women are selected for the jury. If we do this, 90% of the time the null hypothesis is incorrect. Thus, there is a 90% probability that if 3 or fewer women are selected for the jury there is gender bias in the jury selection process. Alternatively, there is a 90% probability that the null hypothesis is false. However, this also means that there is a 10% chance that the null hypothesis is actually true (i.e., there is still a 10% chance the jury selection process is unbiased if 3 or fewer women are selected).

If we reject the null hypothesis and the null hypothesis is actually true, we will have committed what is called a "type I" error. Rarely, if ever, will we know if the null hypothesis is true. What we will know is the *probability* that the null hypothesis is true. Thus, given our findings, there is a 10% chance that the null hypothesis is true, it does not mean the null hypothesis is actually true, just that there is a 10% *chance* that the null hypothesis is true. It is a probability, not a certainty! If we use the 10% "region of rejection" it means that we will reject any outcome that has a 10% or less probability of occurring by chance. In the jury selection example this would mean rejecting the null hypothesis that the jury selection process is unbiased if 3 or fewer (i.e., 3, 2, 1 or 0) women are selected. The level of significance is equal to the region of rejection. Thus, if the "region of rejection" contains any outcome that has a 10%, or less, probability of occurring by chance then we are using a level of significance of 10%.

Here are some important equalities: the region of rejection is equal to the level of significance which is equal to the probability of committing a type I error (i.e., rejecting the null hypothesis when the null hypothesis is actually true). Thus, if we use the 10% level of significance it means that we will accept the null hypothesis as being true if the result is something that would occur more than 10% of the time by chance (e.g., selecting a jury with more than 3 women) and reject the null hypothesis if the result would occur 10% or less of the time by chance (e.g., selecting a jury with 3 or fewer women). Therefore, our decision rule using the 10% level of significance in the jury selection example would be to reject the null hypothesis if a jury with 3 or fewer women is selected and run a 10% chance that the null hypothesis is actually true. Keep in mind, if we reject the null hypothesis it does not necessarily mean that we commit a "type I" error. The null hypothesis may be false (indeed there is a 90% chance it is false). We only commit a "type I" error if we reject the null hypothesis and the null hypothesis is true. If we reject the null hypothesis and the null hypothesis is false we made the correct decision. Since we rarely, if ever, know whether the null hypothesis is actually true, when we reject the null hypothesis we almost never know if we have committed a "type I" error. All we know is the probability that we have committed a "type I" error (10% in this example).

While you will occasionally see a political science article use a 10% level of significance, the general standard is a 5% level of significance. Thus, political scientists typically only reject the null hypothesis if the null hypothesis has a 5% or less probability of being true. If you read a political science article and it says that the results are either "statistically insignificant" or "not statistically significant" it means that the null hypothesis has greater than a 5% chance of being true. Therefore, we would not reject the null hypothesis. For example, if we use the 5% (i.e., .05) level of significance (which political scientists typically do) and our results say that the null hypothesis has a 7% chance of being true, we would not reject the null hypothesis (because 7% is greater than 5%).

If the results are statistically significant at the .05 level it means the following: (1) we will reject the null hypothesis 100% of the time; (2) 95% of the time we will have made the correct decision because the null hypothesis will be false 95% of the time; (3) 5% of the time we will have committed a type I error because we will have rejected the null hypothesis when the null hypothesis is true; (4) we will never know for certain if the null hypothesis is false.

Why do political scientists typically use the 5% level of significance? Because we are very afraid of committing a "type I" error (i.e., rejecting the null hypothesis when the null hypothesis is true). We are very concerned that we will conclude that variable X influences variable Y when it actually does not. For example, we will want to avoid concluding that the balance of power effects the probably war will occur if the balance of power actually has no effect on the probability that war will occur.

The lower you set the level of significance, the harder it is to reject the null hypothesis. This is because the lower you set the level of significance the more different the results have to be from what would occur if the null hypothesis were true (just keep reading!). Take the jury example we have been working with. From the probabilities provided on page 48, if we use the .10 level of significance, we would reject a jury of 9 men and 3 women as being selected from a biased selection process. However, if we use the 5% significance level (i.e., the .05 level), we would not reject the 9 men/3 woman jury as being chosen from a biased selection process.

Instead, we would accept the null hypothesis that the 9 man/3 woman jury was selected from an unbiased process. Using a 5% level of significance, a 9 man/3 woman jury is not sufficiently different than the 6 man/6 woman jury specified by the null hypothesis to cause us to plausibly rule out an unbiased selection process. It would have taken a jury with 10 or more men (i.e., 2 or fewer women) to conclude that the jury selection process was biased using the 5% (i.e., .05) level of significance (see the probabilities on page 48). Thus, the lower the level of significance, the more difficult it is to reject the null hypothesis.

The opposite of a "type 1" error is a "type II" error: accepting the null hypothesis as true when the null hypothesis is actually false. While the "type II" error is important, political science literature almost never discusses it. Virtually all of the attention in political science (and most social sciences) is on the "type I" error. Why is this so? One answer to this question is as previously mentioned, political scientists are very concerned with committing a "type I" error. As previously mentioned, the lower you set the level of significance (e.g., .05 is lower than .10), the more difficult it is to reject the null hypothesis. The more difficult it is to reject the null hypothesis the less likely you are to commit a "type I" error. However, since lowering the level of significance means that you are less likely to reject the null hypothesis, it also means that you are more likely to retain (or not reject) the null hypothesis. Since a "type II" error is to retain the null hypothesis when we should reject it, this means that the lower we set the level of significance, the less likely we are to commit a "type I" error (rejecting the null hypothesis as false when the null hypothesis is true), but the more likely we are to commit a "type II" error (accepting the null hypothesis as true when the null hypothesis is actually false). Thus, our desire to minimize the possibility of a "type I" error means we will have to place less emphasis on (i.e., run a greater risk of) committing a "type II" error. Put somewhat differently, if we reject the null hypothesis we are making a statement of knowledge (i.e., that X does influence Y) whereas if we do not reject the null hypothesis, we are not making a statement of knowledge (i.e., we are not saying that X influences Y). If we make a "statement of knowledge" (i.e., reject the null hypothesis) we want to be very sure we are correct. The concern with a "type II" is more prevalent in public policy than in political science. For example, the cost of retaining a false null hypothesis such as that a vaccine has no effect of the disease it is intended to prevent (an hence the vaccine isn't distributed) could have potentially fatal consequences.

It is important to realize that if we do not commit a "type I" error it does not mean that we have automatically committed a "type II" error. If we do not reject the null hypothesis and the null hypothesis is true, we made the correct decision. We only commit a "type II" error if we do not reject the null hypothesis when the null hypothesis is actually false.

A second reason why political scientists are typically not greatly concerned about a "type II" error is that political science theory (as with theory in most social sciences) usually does not supply the information necessary to definitively calculate the probability of committing a "type II" error. Let me use a brief example from comparative politics and you will quickly understand what I am talking about. In recent years there have been a number of studies by scholars in comparative politics that test theories concerning factors (i.e., independent variables) that influence how long a government in a parliamentary system lasts. Remember that many foreign

countries (e.g., Great Britain) have an election if the ruling political party or ruling coalition (if no party has a majority of the seats in the legislature) does not prevail on a vote in the national legislature.

Let us say that you are a comparative politics scholar and you want to see what effect the number of political parties (the independent variable) has on the duration of time before the ruling party will fail on a vote in the legislature (the dependent variable). One plausible hypothesis might be that the greater the number of political parties the less likely one party can rule effectively, hence, the shorter the likely duration of time between elections. Thus, we would probably hypothesize a negative relationship between the number of political parties and the duration of time between elections. However, our theory does not specify how much each additional political party is likely to shorten the period of time between elections. For example, on average, is each additional political party expected to reduce the time period until the next election by 1 month, 2 months, 10 months, or what? It is extremely unlikely that any reputable comparative politics scholar would have a theory that would yield a specific amount of time that each additional party is likely to shorten the time until the next election. Unless our theory was strong enough to specify an exact amount of time that each additional political party would likely shorten the time before the next election (e.g., 10 months) we can only crudely estimate the probability of committing a "type II" error. This is invariably the situation in political science, economics, psychology and sociology. This is one reason these disciplines do not pay much attention to the probability of committing a "type II" error.

Most comparative scholars would probably agree that, all other factors being equal, the more political parties the more conflict and the less time any one party or particular coalition of parties will stay in power (i.e., the shorter the time between elections). As a practical matter, what a comparative politics scholar would be trying to do is to see if the evidence is strong enough that we could plausibly reject the null hypothesis that, all other factors being equal, the number of political parties is unrelated to the time between elections with a 5% or less chance that the null hypothesis is true. This is a concern with a "type I" error, not a "type II" error.

In the jury selection example there were situations where we would "accept" the null hypothesis that the jury selection process was unbiased (e.g., if the jury was composed of say 7 women and 5 men). In political science we almost never "accept" the null hypothesis as being true. The nature of the scientific process is such that we never make a "final" judgement. We only make tentative judgements such as: given the current state of the evidence this is what we believe occurs.

It is important to realize that the hypothesis the political scientist tests is called the "alternative hypothesis" or simply "the hypothesis" (i.e., that X effects Y), not the null hypothesis. For example, a political scientist would test a hypothesis such as: the more liberal the government the greater the share of income going to the poor. The null hypothesis would be that the liberalism of the government has no effect on the share of income going to the poor. If the evidence against the null hypothesis is not statistically persuasive (i.e., the null hypothesis has greater than a 5% chance of being true), the political scientist will simply conclude that the evidence is insufficient to reject the null hypothesis. This means that the evidence in favor of accepting the "alternative hypothesis" (or "the hypothesis") is simply insufficient. This does not mean we "accept" the null hypothesis as true. We just could not reject the null hypothesis. If the null hypothesis has less than a 5% chance

of being true, the political scientist will reject the null hypothesis and accept the "alternative hypothesis" (or "the hypothesis").

Further Discussion of Samples and Populations

While the readings have distinguished between a "sample" and a "population," they have not distinguished between a "finite" and an "infinite" population. "Finite" means that there is a limited number of outcomes. For example, there are only six possible outcomes from rolling a die (i.e., 1, 2, 3, 4, 5 or 6). By contrast, "infinite" means that the number of outcomes is unlimited. For example, while there is only one income a person actually earns in a particular year, there is an unlimited (i.e., infinite) number of different incomes a person might have earned in that year. Thus, if we "rerun" the same year 100 times, an individual will probably earn 100 different incomes. The following quotation applies the distinction between finite and infinite to statistical inference:

A population can be defined as the totality of all possible observations on measurement or outcomes. Examples are incomes of all people in a certain country in a specific period of time, national income of a country over a number of periods of time, and all outcomes of a given experiment such as repeatedly tossing a coin. A population may be finite or infinite. A finite population is one in which the number of all possible observations is less than infinity. However, the distinction between finite and infinite populations is more subtle than may at first appear. For instance, a series of national income figures for the United States for a number of years, e.g., 1948-1977, represents a finite collection of thirty observations and thus might seem to be a finite population. But this would be a very narrow interpretation of historical events, since it would imply that the thirty measurements of national income were the only possible ones, i.e., that there is only one course that history might have taken. Now there are obviously not many people who would take such an extremely fatalistic view of the world; most people would admit that it was not impossible for some other, even if only slightly different, values of national income to have occurred. This latter view underlies virtually all policy-oriented research in economics and econometrics (and political science) and will be used throughout this book. Thus a population of national incomes in a given time interval includes not only the actual history represented by the values that were in fact observed but also the potential history consisting of all the values that might have occurred but did not. The population so defined is obviously an infinite one. Similarly, the population of all possible outcomes of coin tosses is also infinite, since the tossing process can generate an infinite number of outcomes, in this case "heads" and "tails." Most of the populations with which we deal with in econometrics (and political science) are infinite. (emphasis added)

Source: Jan Kmenta, Elements of Econometrics, 2nd ed., pp. 3-4.

The following quotation is also useful concerning statistical inference. Although the quotation will probably seem confusing, keep reading. I have a truly "brilliant" visual example to follow!

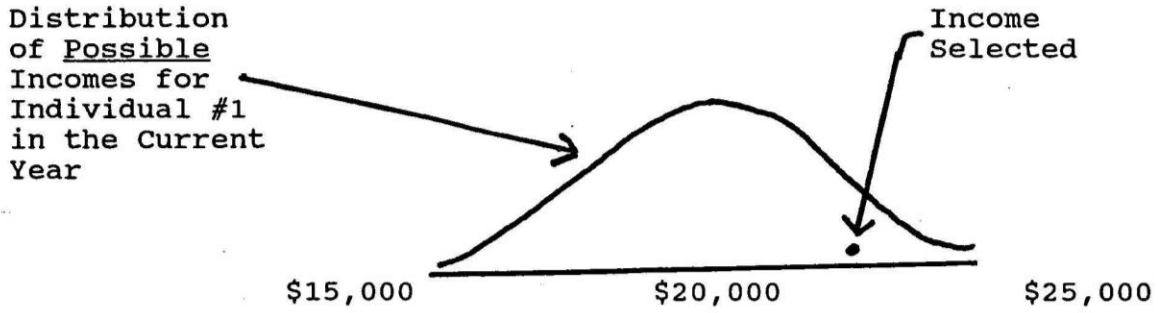
> We continually refer to the set of all possible outcomes as the population and to the processes underlying the outcomes as the population model. If we are interested in people's incomes and the relationship between incomes and education, ... the possible outcome (income) for any individual in any year may take on an infinite number of values, i.e., any positive number, with some outcomes being more likely than others. If we were to record the incomes of all individuals in a region, country, or even the universe in a given year we would not have the entire population of outcomes (even though we have the population of individuals) because each person's income in that year is simply one outcome, or value, from the entire set of possible outcomes for that person. Our presumption in relating incomes to education ... is that the distribution of possible incomes varies for each individual, and that these variations in distributions are related to the educational ... characteristics of the individual. The purpose of statistical analysis is to use the set of observed outcomes (incomes) for each individual to estimate how these variations are related to education. (emphasis added)
>
> Source: Eric A. Hanushek and John E. Jackson, Statistical Methods for Social Scientists, pp. 325-326.

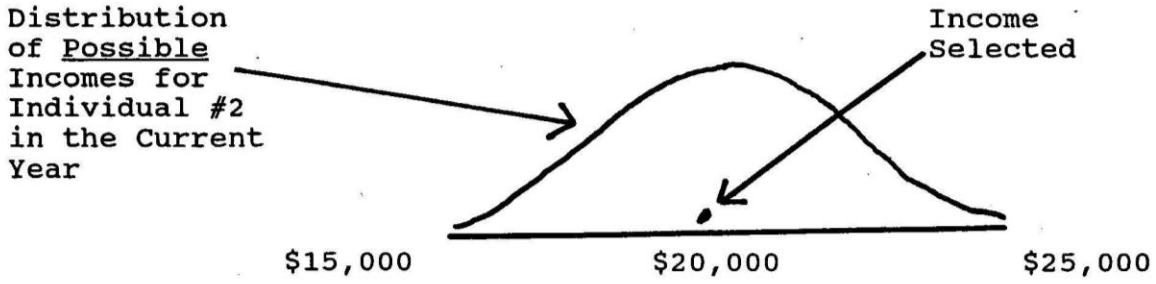
In most all research situations we have data on a sample, not a population. Even when we think we have data on all members of the population we usually have a sample because typically the "population" is infinite. For example, if we have data on the 100 U. S. Senators who served in a particular year, we have data on only one of the infinite number of U. S. Senates that could have been elected. If we were to "rerun" history (i.e., the last national election) and one senator who had been elected now lost, the Senate that resulted from our "rerun" history would have a slightly different membership than resulted from the first election. Although we cannot "rerun" history, this illustration is important because we want to generalize our findings not only to the 100 senators who actually served (i.e., one Senate - composed of those 100 senators), but to an infinite number of possible Senates which could have been elected. Thus, the actual Senate is just one "sample" from an infinite number of possible Senates that could have been elected.

Since there is no limit to the number of different outcomes that could occur in an infinite population, we can never know the "true" value of the mean (or any other statistic) for an infinite population. We can only "sample" from an infinite population. We use significance tests to assess the likelihood (i.e., probability) of various magnitudes of relationships in the population (usually an infinite population) of interest. The following discussion and diagrams should help clarify the education and income example mentioned in the previous quotation from Hanushek and Jackson.

Let us stipulate that individual #1 is a college graduate whereas individual #2 is a high school graduate. Therefore, individual #1 has a higher level of education than individual #2. If in a given year individual #1 has a higher income than individual #2, we would want to know which of the following two scenarios better represents the truth. Remember that our income figure for each individual is just one selection from an entire distribution of income for that particular individual in that particular year. Thus, in this particular year individual #1 could have earned an infinite number of different incomes. According to scenario #1, these possible incomes for individual #1 average $20,000 and are distributed as follows. In scenario #1, individual #2 has the same income distribution as individual #1.

Scenario #1

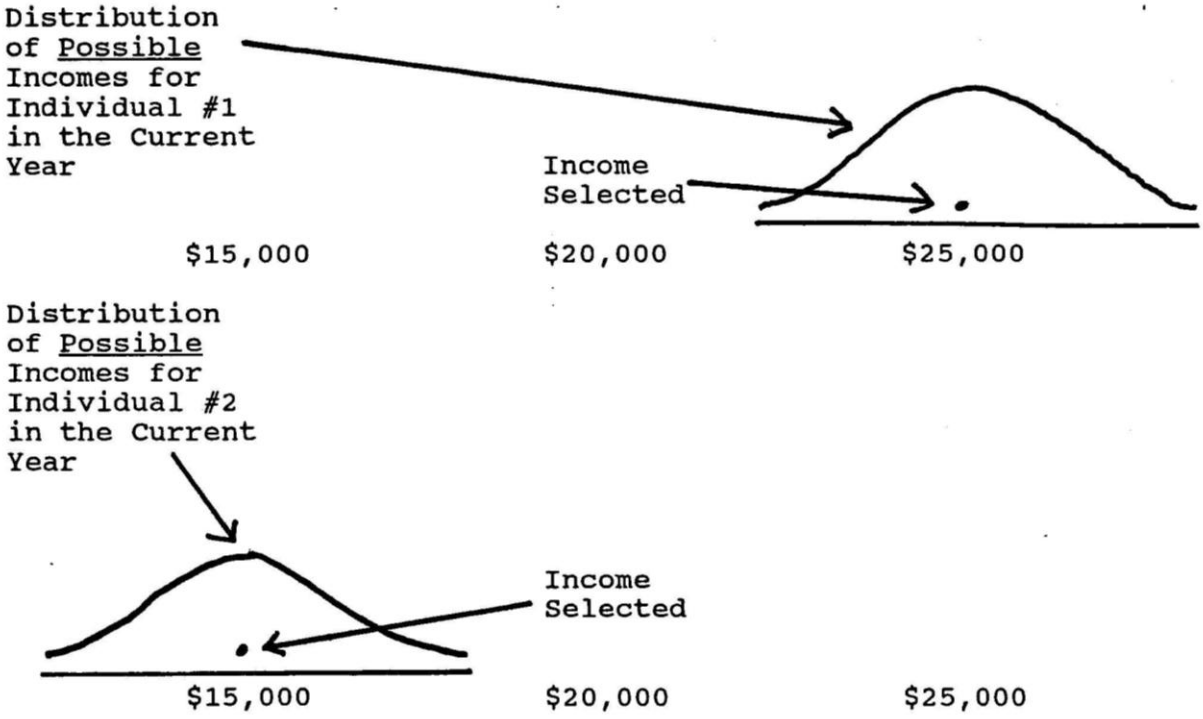




While the income selected (i.e., that actually occurred) for individual #1 (approximately $23,000) is higher than for individual #2 ($20,000), the distribution of income for both individuals is the same. Thus, the two distributions have the same mean income ($20,000) and the same standard deviation. Hence, the null hypothesis of no impact of education on income is true, the difference reported is only due to sampling variation and not to a different income distribution (or "profile") for each individual. Perhaps individual #1 had a higher income because they won $3,000 in the state lottery. This is a "fluke." The higher level of education of individual #1 most likely had no influence on how lucky they were in the state lottery. If we continued to select incomes from each of these two distributions, they would average the same amount (i.e., the two distributions have the same mean). Furthermore, if we "rerun" history, perhaps individual #2 would have won $3,000 in the state lottery. If so, individual #2 would

have had an income $3,000 higher than individual #1. If the world works according to scenario #1, education has no effect on income. The "true" population values are the two means (i.e., $20,000), which are the same. If our results suggest that education does effect income it would merely be the product of chance (i.e., it occurred in the sample, but not in the population of interest). Let us now examine scenario #2.

Scenario #2



The income selected for individual #1 is higher than for individual #2 because the distribution of income for individual #1 has a higher mean ($25,000) than for individual #2 ($15,000). The standard deviations of the two distributions (both normal curves) are the same. If we "rerun" history, individual #1 would almost always earn a higher income than individual #2. This is a "true" difference and not the result of sampling variation. If we find that income and education are positively associated (as both scenarios suggest), significance tests are useful in trying to estimate how often we could reject scenario #1 (the "null" hypothesis - which would mean our results were merely the product of sampling and not because education and income are related in our population of interest) in favor of scenario #2 (that education and income are positively related in our population of interest) when scenario #1 is actually true (this would be committing a "type 1 error"). Even if we had an income figure for each member of the population of interest we still have only a sample (because we have selected only one income for each individual from the entire distribution of income for each individual in that year).

The purpose of this section is to introduce a test for statistical significance and apply the material you have recently studied. The level of statistical significance is equal to the probability of committing a "type I error." A "type I error" is rejecting the null hypothesis when the null hypothesis is actually true. If we reject the null hypothesis that X is unrelated to Y in favor of the alternative hypothesis that X is related to Y and our results are statistically significant at the .05 level, it means thatwe have a 5% (or less) chance of committing a "type I error."

Suppose we are testing a hypothesis that could be derived from international relations theory: the balance of military power between two nations (variable X) is negatively associated with the probability that an ongoing conflict between these nations will escalate (variable Y). Thus, since higher scores on balance of power (i.e., a more equal balance of power) are hypothesized to be associated with lower scores on conflict escalation (i.e., less escalation/less of a conflict) we are expecting a negative association. The test of statistical significance we will use is the chi square test. Do not worry about how the various probabilities below were calculated. Suppose the results are as follows:

	Probability that Conflict will Escalate	
Equal Balance of Power	20%	(25)
Unequal Balance of Power	45%*	(85)
		110

*significant at .05
**significant at .01
***significant at .001

Since the probability that conflict will escalate is lower when power is equally balanced (20%) than when power is unequally balanced (45%), the hypothesis is supported. We have not "proved" the hypothesis is "true." The hypothesis could be still be "false." All we can say is that the data are consistent with (or "support") the hypothesis. Alternatively, it is "likely" that the hypothesis is true. Never say that you have "proved" a hypothesis to be "true." As long as the probability that conflict will escalate is lower when power is equally balanced than when power is not equally balanced, the hypothesis is supported. For example, if the probability that conflict will escalate had been 60% (power equally balanced) vs. 85% (power unequally balanced), the hypothesis would still be supported because the critical factor is the direction and amount of difference between the probabilities (i.e., that the probability conflict will escalate with equally balanced power is lower and by how much), not the level where the differences occur (i.e., 20% vs. 45% as opposed to 60% vs. 85%).

The fundamental question of statistical significance is: How likely are the results the product of chance? Applied to our situation this question can be phrased as follows: How likely are we to find a 25% difference (45%-20% = 25%) in the probabilities that conflict will escalate in our sample when the "true" difference in the population (all nations at all conflictual times) is zero percent? If the actual difference is zero percent the null hypothesis is true. Since we can not know the difference in the population (i.e., all nations at all conflictual times), we will never

know for sure whether the null hypothesis is actually true. Given our sample size (110 - see the table on page 57) the chi square test indicates that this 25% difference in probabilities is statistically significant at the .05 level (note the single asterisk - "*" in the table on page . Therefore, if we reject the null hypothesis, we have a 5% (or less) chance of committing a "type I error." Hence, while the null hypothesis could be "true," it is rather unlikely to be "true." We reject the null hypothesis when (as in our situation) the probability of committing a "type I error" is 5% (or less).

Like all significance tests, the chi square test is based upon the following two criteria. First, how great is the relationship between X and Y? As I just mentioned, what we might call the "size of the difference" in our case is 25% because the difference in the probability that conflict will escalate between our two categories (i.e., equal balance of power and unequal balance of power) is 25% (i.e., 45% - 20% = 25%). Assuming the same sample size, if the "size of the difference" was greater than 25%, the results would be even more statistically significant.

Instead of the cross tabulation table that appears on the previous page, suppose we had calculated a gamma between the balance of power and the probability that conflict will escalate. Suppose the gamma was -.37. Assuming our variables were measured with little random measurement error, we know from page 34 that a gamma of -.37 indicates a moderate negative association between the variables. So, if we were using a gamma, instead of the "size of the difference" the first criteria would be the -.37 association. Assuming the same sample size, had the gamma been -.57, instead of -.37, it would make the result more statistically significant (remember from page 37 that larger negative numbers, like larger positive numbers, mean a stronger relationship).

Instead of either a cross tabulation table or a measure of association (such as gamma), suppose we estimated the magnitude of the relationship between the balance of power and the probability that conflict will escalate. For example, look at the two line slopes on page 40. Line "A" is noticeably steeper than line "B." Put another way: There is a greater increase in Y for each increase in X with line "A" than with line "B." In the example on page 37, X is years of education and "Y" is income. Clearly, each additional year of education is associated with a greater increase in income with line "A" than with line "B." If the sample size remains the same, the steeper the line, the more statistically significant the result. Thus, if the sample size remained the same, line "A" would produce more statistically significant results than line "B." Thus, which ever method by which you are estimating the relationship between X and Y (i.e., by cross tabulation, a measure of association - such as gamma or by the slope of a line - as we will later on with regression), if the sample size remains the same, the greater the relationship between X and Y, the more statistically significant the result.

The second principle of any test of statistical significance concerns how many observations (in our case 110) are used in estimating the relationship between X and Y. The greater the relationship between X and Y, the smaller the number of observations you need to achieve statistical significance. For example, if you toss a coin 10 times, and all 10 tosses are heads, you can be quite sure that the coin is biased. Although the number of observations is small (10), the size of the difference is great (100% heads instead of the 50% heads we would expect if the coin were unbiased). In this situation, if we reject the null hypothesis we have less than a 1 in

1,000 chance of committing a "type I error."

Alternatively, with very large samples (e.g., 3,000) even very small differences will be statistically significant. For example, if you toss a coin 100 times and heads come up 51 times, how sure would you be that the coin was biased? Since only one less head would have produced an unbiased result (i.e., 50 heads and 50 tails), you would probably not be very sure that the coin was biased. An unbiased coin is like a "null" hypothesis (i.e., no difference between the probability of heads and tails). However, if the coin comes up heads 51,000,000 times out of 100,000,000 tosses (as previously, 51% heads), this 1% difference (51% obtained vs. 50% expected) would be statistically significant (because of the extremely large sample). Thus, just because a relationship is statistically significant, it is not necessarily substantively important. A statistically significant finding that the probability conflict would escalate had decreased only 1% if power were equally balanced would not be strong support for our hypothesis.

Make sure you do not confuse statistical significance with support for the hypothesis. Suppose you hypothesize that a coin will flip more heads than tails. If you flip the coin 10 times and get 6 heads and 4 tails the results support the hypothesis (because 6 is greater than 4). However, since the coin was only flipped 10 times with a resulting 6/4 split, the results would not be statistically significant. However, if you flipped the coin 10 times and all 10 flips are tails, this is opposite to the hypothesis (because we hypothesized more heads than tails) and would be statistically significant (only 1 time in 1,000 would the null hypothesis - that the coin flips an even number of heads and tails - be true). This is strong evidence against the hypothesis.

Political scientists invariably reject the null hypothesis if the null hypothesis has less than a 5% chance of being true. Thus, if our results are statistically significant at the .05 level, we reject the null hypothesis that X has no effect on Y in favor of the alternative hypothesis that X does have an effect on Y. In this situation, there is a 5% chance that we will commit a "type I error" (i.e., a 5% chance the null hypothesis is actually true).

If our study had either more observations than 110 and/or a greater "size of the difference" than 25%, our results might have been statistically significant at the more demanding (i.e., more difficult to achieve) .01 (only 1 time in a 100 would we commit a "type I error") or .001 (only 1 time in a 1,000 would we commit a "type I error") level. Obviously, if our results were statistically significant at either the .01 or .001 level, they would also be significant at the .05 level (because the .05 level is easier to achieve than either the .01 or .001 level). So, if our results are statistically significant at either the .01 or .001 level we would reject the null hypothesis. The advantage of achieving statistical significance at either the .01 or .001 level, as opposed to the .05 level, is that we have a smaller chance of committing a "type I error."