Reader for POSC 300

Dr. Christopher Dennis

The Scientific Method in Political Science

Political scientists are interested in a wide variety of different topics. Some political scientists study the causes and consequences of war. Other political scientists study why voters support a particular candidate and what difference it makes which candidate or party is victorious. For political scientists studying these and a vast array of other topics, quantitative application of the scientific method offers one of the most useful approaches to increasing knowledge. While the scientific method is not useful in answering questions posed by normative political philosophers (e.g., What is justice?), it is extremely valuable in understanding how and why political phenomena occur. As a first step in this process, we need to understand what the scientific method is. Let me suggest that science is defined by its methodology, not its subject matter. Thus, it is how one studies something, not what is studied, that determines whether or not the researcher is using the scientific method. Throughout this course I will use the following definition of science: a communicable (can communicate to those who do not "know"), falsifiable (possibility of non-confirmation), and logical (conclusions follow from the facts) method of pursuing knowledge involving the recognition and formulation of a problem, the collection of data through observation and experiment, and the formulation and testing of hypotheses. That is a long and involved definition!

In order to fully understand the above definition of science, it is useful to break it into parts. As stated in the definition, "communicable" means that your study can be understood by those who are not already part of it. Thus, you need to be able to explain your method to other researchers so that they can both check your work for accuracy and apply it to new circumstances. For example, if you were testing a new vaccine, future researchers might wish to replicate (repeat) your work on the same type of subjects and then administer it to a different group. Therefore, other researchers would have to know both the composition of your vaccine and how the tests of it were conducted.

The second portion of the above definition of science concerns "falsifiability." Falsifiability means that there must be some outcome which would countermand what we expect to happen. For example, suppose you said that if God wants you to go to Pittsburgh he will provide you with the plane tickets. In terms of falsifiability, this would not be a scientific test of the existence of God. Regardless of whether you received the tickets or not, you would not doubt the existence of God. No possible outcome would lead to a rejection of the premise that God exists.

The next portion of the definition of science concerns logic. As stated above, your conclusions must follow from the facts. You cannot conclude that the Democratic party favors a greater tax burden on the wealthy than the Republican party when the facts suggest otherwise.

The next sections of the definition of science are rather straightforward. Certainly, you must formulate a problem or else there is nothing to study. Furthermore, you collect data which bear on the problem (or topic) you are studying. For example, you collect data on the differences between the tax proposals offered by the Democratic and Republican parties.

The final portion of the definition of science concerns the formulation and testing of hypotheses. Let me define a hypothesis as follows: a relational statement between two, or more, concepts which is deductively plausible and empirically generalizable. I think it would be useful to begin by defining a concept. A concept is an abstraction representing an object, a property of an object, or a certain phenomena. For example, "poverty" could be a concept. Should we conceptualize "poverty" in "relative" or "absolute" terms? A person with an income of $20,000 could be thought of as impoverished relative to someone who had an income of $500,000. On the other hand one could argue (as conservatives generally do) that poverty is absolute. Thus, the determination of poverty would not concern how much income you had relative to someone else, but rather whether you could attain a particular standard of living (e.g., avoid hunger). My point is that a researcher measuring poverty has to use one of these conceptualizations. Obviously, it matters which one they choose. One of the great advantages of the scientific method is that the researcher must state and defend their choice. By so doing, other researchers can then use the same, or different conceptualizations, and see to what extent the results are affected by the conceptualization employed.

A concept that can assume different values is called a variable. For example, since all governments do not have the same degree of liberalism, governmental liberalism is a variable. Two particular types of variables are fundamental to the scientific method (particularly hypothesis testing). The presumed "causal" factor is referred to as the "independent" (or "predictor") variable and the effect is referred to as the "dependent" variable. Suppose we hypothesize that the degree of governmental liberalism alters the percentage of income going to the poor. Since governmental liberalism is presumed to effect the percentage of income going to the poor, governmental liberalism is the independent variable and the percentage of income going to the poor is the dependent variable. Alternatively, you might think of it this way: the score on the dependent variable depends upon the score on the independent variable (not the other way around).

An operationalization is the measurement of a concept. For example, how do you quantify whatever conceptualization of "income" you are using? If you receive medical benefits from the government does this count as "income"? Should the use of a company car count as "income"? The researcher must explain how they measure "income" and defend why their measure is appropriate.

A relational statement is a causal or associational link between concepts. For example, suppose we hypothesize that the liberalism of the federal government is "positively" associated with the percentage of income going to the poor. The previous statement is relational because it depicts an association between the concepts (liberalism of the federal government and the percentage of income going to the poor). Notice the use of the directional term "positive." A "positive" relationship means that higher scores on one variable are associated with higher scores on the other variable. For example, if the liberalism of the federal government and the percentage of income going to the poor are "positively" related (i.e., associated) it would mean that if the score on federal government liberalism were to increase from say 50% to 75% (e.g., by electing many more Democratic congressmen and senators) the percentage of the national income going to the poor might then

increase from 10% to 12%.

Relationships between variables can also be "negative." A "negative" relationship means that higher scores on one variable are associated with lower scores on the other variable. For example, if the liberalism of the federal government and the percentage of income going to the poor are "negatively" related it would mean that if the score on federal government liberalism were to increase from 50% to 75% the percentage of the national income going to the poor might then decrease from 10% to 8%.

Deductive plausibility means that the researcher deduces (reasons from) something else which is plausible. Thus, if we observe that Democrats tend to be more liberal than Republicans we may reasonably deduce from this that a particular Democratic candidate is likely to be more liberal than their Republican opponent (this is what we hypothesize and will be testing).

Empirically generalizable means that our findings are applicable to much of the observable (empirical) world. We want to generalize as far as we can. From Democrats and Republicans in California to Democrats and Republicans in the United States as a whole. Any theory (a theory is just a more certain hypothesis) is more valuable the wider its applicability. For example, isn't the anti-crime argument for the death penalty (that the death penalty will lower the murder rate) actually just an application of the basic economic theory that the more something costs (here "costs" would refer to the penalty) the less of it will be sold (each murder would be an occurrence, i.e., a "sale")?

Basically, the scientific process is just a continual testing of hypotheses in order to find their limits (i.e., how far they can be generalized) and then to modify the theory in light of the findings. For example, in most western democracies the poor are more supportive of liberal governments than conservative governments. Since it is logical to hypothesize that a government will pursue policies that disproportionately benefit its supporters, it would seem logical to hypothesize that the degree of liberalism of the government is positively associated with the percentage of income going to the poor. Thus, higher scores on our measure of governmental liberalism should be associated with higher scores on our measure of the percentage of income going to the poor. In testing this hypothesis we may find that government today has either a greater, or lesser, impact on the distribution of income than during the 1950s.

Another benefit of the scientific method is that the user must make their model explicit. For example, it is possible that the liberalism of the government has little "direct" effect on the percentage of income going to the poor. Since governments often control policy instruments (e.g., the money supply), as opposed to policy outcomes (e.g., the percentage of income going to the poor), it is likely that much of the effect of government on the percentage of income going to the poor would be "indirect" (i.e., through other factors). For example, a more liberal government could increase the money supply. A larger supply of money lowers interest rates which, in turn, make borrowing less expensive. The reduced cost of borrowing money generally causes plants to expand which, in turn, lowers the unemployment rate. As the unemployment rate decreases, the percentage of income going to the poor typically increases. My point is that the user of the scientific method must explain

which variables effect which other variables (i.e., they must make their "model" explicit).

Because users of the scientific method must make their models and measures explicit, other researchers can replicate (i.e., repeat) and expand on the original study. Over the past two decades, political scientists have tested the governmental liberalism hypotheses I have been mentioning in most all major industrialized democracies in the world. They have used an impressive group of alternative income measures, time periods, and models. For example, in addition to studying the effect of governmental liberalism on the money supply, political scientists have also examined the effects of governmental liberalism on the amount and distribution of the tax burden over various income groups, numerous measures of social welfare spending and the amount of economic growth.

Users of the scientific method usually have two goals in mind. Typically, the first goal of a user of the scientific method is explanation. In our example we are trying to explain why the percentage of income going to the poor varies (i.e., is not always the same - hence a "variable"). Our hypothesis is that variation in the liberalism of the government is what causes variation in the percentage of income going to the poor. A large literature (to which political scientists have greatly contributed) has rather firmly established that governmental liberalism is positively associated with the percentage of income going to the poor. However, while governmental liberalism is likely to positively influence the percentage of income going to the poor, other factors (i.e., independent variables) are also likely to influence the percentage of income going to the poor (e.g., international economic trends). The result of incorporating these additional independent variables in the data analysis is a richer explanation of why the percentage of income going to the poor varies.

A second goal of users of the scientific method is prediction. Applied to our hypothesis this would mean to predict how much the percentage of income going to the poor would increase, or decrease, depending upon a particular amount of change in the liberalism of the government. Often these two goals are related. As our ability to "explain" a process improves, our predictions are likely to become more accurate. However, prediction is more difficult than explanation. The impact of some of the independent variables may change in the future. Consequently, accurate predictions are difficult. Nevertheless, political scientists have formulated relatively accurate forecasts of the share of the vote American political parties will receive (the dependent variable) based upon changes in various economic and non-economic variables (the independent variables). However, typically the major goal of contemporary quantitative political science is explanation.

Research Design

Before continuing, make sure you understand that pages 2-5. The topics dealt with over pages 2-5 are the foundation of every reading in this course. The first quiz (coming the day this reading assignment is due) may well ask you to define a variable, abstract a hypothesis from written material, and/or to explain the difference between a "positive" and "negative" relationship. You will need the information from

the aforementioned lecture on the scientific method for quizzes 1-3 and the final examination.

The main purpose of pp. 6-14 is to discuss the early stages of a quantitative research project. The first decision any researcher must make is what topic to study. A political scientist should be able to defend their choice of a topic on normative grounds. Thus, why is the topic important? For example, why should we study the causes of war? I think one could make an excellent case that war is undesirable and, consequently, that determining why war starts is a logical pre-condition to minimizing its occurrence. Although a normative defense of a topic is important, it is typically handled in several sentences. The central contribution of quantitative research is to explain what takes place and why, not what is "good" or "bad."

In quantitative research (the topic of this course), we are usually testing a theory of behavior. Whether it is the behavior of governments or individuals, we will probably be examining the causes (and/or consequences) of some form of political behavior.

Any quantitative (i.e., empirical) study is trying to perform two fundamental tasks. First, we are trying to test and refute hypotheses. For example, is the liberalism of a government positively associated with the amount of government support for the poor? Second, we are trying to estimate the magnitude of the relationships between the variables (Hanushek and Jackson, Statistical Methods for Social Scientists, pp. 2-3). For example, the replacement of a Republican President with a Democratic President would result in how much more support for the poor?

After formulating the hypotheses (defined on pages 3-5), we need to begin thinking about how we will test them. The strategy by which one tests their hypotheses is called a research design.

While this may be a bit of an oversimplification, there are two basic types of research designs. The first type of research design is called an experimental design. With an experimental design, the researcher can adjust the level (i.e., amount/scores) on each of the independent variables. For example, supposing a biologist formulates a new plant growth additive and wishes to test its effectiveness. The amount of the plant growth additive each plant receives would be the independent variable. The growth rate of the plant would be the dependent variable. The biologist would probably think that factors other than the amount of the plant growth additive would alter the rate of plant growth. Thus, the plant biologist would want additional independent variables. For example, such factors as the type of plant, plant condition, water quality and the amount of sunlight could all affect the growth rate of a plant. Each of these factors, in addition to the growth additive, is an independent variable. The advantage of using an experimental research design is that the researcher can set the level of each of the independent variables. For example, the plant biologist can determine what types of plants will be used, the amount of the growth additive each plant will receive and the amount of sunlight each plant is exposed to. Being able to set the level (i.e., amount) of each of the independent variables is an extremely useful capability. If all conditions (i.e., independent variables) other than the independent variable in which the plant biologist is most interested (the growth additive) are set at the same level (i.e., "controlled" - each plant is of the same type, receives the same amount of sunlight, etc.) and if plants

that receive more of the plant growth additive grow faster, we are on rather sound ground in thinking that the plant growth additive matters. Since the plants do not differ on any factors that could conceivably alter their growth rates except the amount of the plant growth additive, it makes sense to think that the growth additive increased plant growth rates.

By contrast, a political scientist will almost invariably have to use what is termed a nonexperimental research design. With a nonexperimental research design the researcher is not able to set the levels of the various independent variables. The inability of the researcher to set the levels of the various independent variables is important because it is possible (in some circumstances likely) that the independent variables will be related to each other, as well as to the dependent variable. We refer to the situation where the independent variables are strongly related to each other as "multicollinearity."

For example, suppose we are trying to test a model of partisan affiliation (the dependent variable). Thus, our model will be trying to explain why individual voters register as Democrats, Republicans, or Independents. Please note that we have three categories of responses (i.e., Democrat, Republican or Independent) on one dependent variable. Let us say that two of our hypotheses are that the more highly educated a voter is the more likely they are to register Republican and the higher the voter's income the more likely they are to register Republican. Note that both education and income are independent variables. Since occupations requiring a higher level of education generally pay higher salaries than occupations with lower educational requirements, education and income are likely to be related. If it turns out, as is likely, that education and income are strongly related to each other (hence we have "multicollinearity"), and both education and income are related to partisan affiliation, it can be difficult to determine the impact of either education or income on partisan affiliation. In the worst case situation, where all voters with high levels of education have high incomes and vice versa, it would be impossible to determine the contribution of either education or income to partisan affiliation.

A political scientist would like to assign various levels of education to high income voters. Thus, some high income voters would have low levels of education (e.g., through the tenth grade), others would have a somewhat higher level of education (e.g., high school graduate) and others a still higher level of education (e.g., college graduate). As all voters with high incomes would not have the same level of education, this would eliminate the multicollinearity between income and education. Needless to say, "assigning" levels of education is not possible. For example, how could a political scientist remove four years of education from a voter?

While a biologist can often set the level of each independent variable for each observation (i.e., each plant) and hence eliminate multicollinearity, a political scientist is unlikely to be in a similar situation. However, as we will see later, political scientists using nonexperimental research designs can still "control" for the impact of each independent variable on the dependent variable. We just do it statistically rather than by setting the level of each independent variable. Furthermore, suppose no voter in our sample had a doctorate in medicine. While a political scientist might like to study the effect of having a doctorate of medicine on someone's partisan affiliation, unless some members of our study have such a

degree, we will be unable to estimate the impact.

A political scientist frequently encounters one additional problem: Did change in the independent variable precede change in the dependent variable? In order for a change in income to "cause" a change in partisan affiliation (i.e., a voter's income increases from $40,000 annually to over $200,000 so they change from being a Democrat to a Republican), the change in income would have to occur before the change in partisan affiliation. The fact that most voters with an annual income of over $200,000 are Republicans may, or may not, mean that if a Democrat's income changes from $40,000 to over $200,000 they will become a Republican. The assumption of our model is that income change precedes partisan change. While this is plausible, it may not be accurate. It would be preferable to test according to the assumptions of our model. In this case that would literally mean we would have to change a voter's income and then see what, if anything, happened to their partisan affiliation. Obviously, we can not do this. As previously mentioned, the plant biologist is in a preferable situation because s/he can first administer the plant growth additive and then see how fast the plant grows.

The situations I have just described are the crux of the differences between an experimental and a nonexperimental research design. To recap briefly, the previous analysis suggested three weaknesses of the nonexperimental research design relative to the experimental research design: (1) more severe multicollinearity (e.g., voters with high incomes were also likely to have high levels of education); (2) an absence of some possible levels of an independent variable (e.g., no one in our study of partisan affiliation with a doctorate of medicine); and (3) less confidence that change in the level of one of the independent variables preceded change in the level of the dependent variable (e.g., did a voter's partisan affiliation change before, or after, a change in their income?). You might well have gotten the impression that since political scientists typically have to use a nonexperimental research design their findings are not very useful. Fortunately, this is not the case. Furthermore, the situation is improving.

Let me now address each of the three problems mentioned above. First, in many studies the interrelationships between the independent variables are actually quite low (i.e., multicollinearity is quite low). Additionally, even when multicollinearity is rather high, we can often accurately estimate the impact of the interrelated independent variables. For example, in a study of voting in the U.S. Senate the principle independent variable in which the researcher may be interested (the senator's political philosophy) is highly related to some of the other independent variables (e.g., the senator's partisan affiliation). Nevertheless, the findings concerning the impact of political philosophy are quite strong and reliable. Hence, even though multicollinearity appears to be a major problem, it is not. Furthermore, later in the course we will discuss strategies to deal with severe multicollinearity. A major topic of this course is how we "control" (i.e., set, or hold constant) the level of various independent variables. While our approach to isolating the unique impact of each independent variable on the dependent variable is not as desirable as that offered by the experimental design, it is nonetheless quite useful.

The second problem of a nonexperimental research design is that we may not have observations on some scores for one, or more, of the independent variables. For example, perhaps no voters in our sample have a doctorate of medicine degree. While potentially important, this problem is usually not catastrophic (terrible pun!). With large sample sizes we usually have several cases of each interesting score. In the partisan affiliation study, political scientists typically have samples of 2,500, or more, respondents. Even if we have few medical doctors in such a study, we probably have enough individuals with similar educational backgrounds (e.g., dentists) for useful statistical analysis. Furthermore, in many instances the omission of a particular category is not of critical importance. For example, it may not be important that we have no respondents with zero dollars in income. Even the poor have some income. It is probably not important to be able to generalize one's findings to situations which are extremely unlikely to ever occur.

The third problem of a nonexperimental research design concerns the degree of confidence we can have that change in the independent variable(s) precedes change in the dependent variable. Thus, did a change in the voter's income precede a change in their partisan affiliation? This is a serious problem. However, like the preceding two problems, the situation is far from hopeless. In many practical research situations our theory is strong enough to be reasonably certain that change in the independent variable preceded change in the dependent variable.

Suppose we are interested in the impact of a senator's political philosophy (the independent variable) on the probability that the senator will vote in favor of shifting the federal tax burden more toward higher income earners (the dependent variable). We can feel quite certain that the senator's political philosophy was formed prior to the time they voted on the tax shift. Few senators either enter the Senate without a political philosophy, or significantly change their political philosophy after they begin serving in the Senate. Much prior research has established the preceding point. In such situations, we can be quite confident that the level of the independent variable (e.g., the senator's political philosophy) was established prior to the score on the dependent variable (i.e., the direction of the senator's vote on the tax shift). Additionally, political scientists have minimized this "time of change" problem through the increasing use of time series studies. A time series study means that the data are collected over time (e.g., annually - each year from 1950 to the present). By contrast, a study in which all data are collected at the same time point (e.g., all nations of the world in the year 2008) is called a cross-sectional study.

For example, if we study the effect of political party control of the executive (i.e., whether the president is a Democrat or a Republican) on such economic outcomes as the unemployment rate or the growth rate, we would probably collect our data annually for a number of years. Obviously, we would know what the level of the independent variable was (i.e., the political party of the president) prior to any change in the dependent variable (e.g., the unemployment rate). Thus, with time series data we can often be more confident of our conclusions than with cross-sectional data. A time series style study done by collecting data on the same individuals at several time points is called a panel study. For example a famous panel study of political socialization (i.e., how people learn about politics) was done

interviewing the same people over several decades (M. Kent Jennings and Richard G. Neimi, Generations and Politics). Multiple interviews of the same person at several different time points (e.g., in 1986, 1996 and 2006) can give us a more valid view of the learning process than by interviewing someone as an adult and asking them to "recall" their youth.

Measuring Variables in Political Science

After formulating our research design, we need to measure our variables. Thus, we will need measures for each independent variable and the dependent variable. Our purpose is to classify outcomes on each variable. For example, in an international relations study we may need to measure the balance of power. How does the researcher do this? First, we need a "concept" of power. Second, we need an "operationalization" of power. For this discussion I will assume that we already have both a concept and an operationalization of power (see page 2 on the meaning of "concept" and "operationalization").

Once we have concepts and operationalizations for each variable, we can proceed to assigning mathematical values for each possible category on every variable. For example, if we conceptualize power as money and operationalize military power in terms of the defense budget, then we need categories for each possible outcome. In this case that would likely mean the amount of money (probably in U. S. dollars) spent by each nation over some specified period of time (probably annually). Perhaps, we should have conceptualized power differently. For example, if economic power is part of "defense" (or "offense"!), then perhaps the value of the gross national product would be a better indicator of "power" than military spending. While critical, the answers to such questions are often unique to a particular topic. Our consideration here is to assign mathematical values of outcomes on a particular variable.

There are four different "levels" of measurement. In the presentation that follows, each succeeding "level" retains all of the desirable properties of the preceding "level(s)," but adds some useful properties not contained in the preceding "level(s)." The weakest (i.e., "lowest") level of measurement is called nominal measurement. A nominal measure classifies each possible outcome. For example, the dependent variable in international relations studies is frequently whether or not war occurred. Let us say we scored "peace" as zero and "war" as one. This would be an example of a nominal level measure. Each outcome (peace or war) is mutually exclusive (i.e., can be only one category). Thus, peace can not be classified as war and vice versa. Furthermore, every outcome has a category. Thus, the only possible outcomes are peace or war. Hence, our measure is collectively exhaustive. When a variable has only two categories of responses (e.g., peace or war), it is termed a "dummy" variable. Occasionally, we will have a nominal level measure with more than two categories. For example, suppose we are studying voter attitudes on foreign policy (the dependent variable), one of our independent variables might be the respondent's race. Race would obviously have many more than two categories. Additionally, there is no inherent ordering to the categories. The coding would be entirely arbitrary. For example, would it make any more sense to code Asian-

American as zero, White as one, African-American as two than White as zero, Asian-American as one and African-American as two. No!! Thus race is inherently a nominal level variable.

The ability to rank (i.e., order) categories of responses on a variable is a feature of the second level of measurement, ordinal level measurement. For example, in a study of the foreign policy attitudes of voters (the dependent variable), political party affiliation of the voter might be a logical independent variable. Political party affiliation is often measured by what is termed a Likert scale (named for psychologist Rene Likert). Such a scale of partisan affiliation could be formulated as follows: (1) strong Democrat, (2) weak Democrat, (3) Independent, (4) weak Republican, (5) strong Republican. The preceding scale has an underlying order (hence "ordinal") or continuum. The continuum might be thought of as being from the most Democratic (category #1) to the least Democratic (category #5). Thus, a strong Republican is the least likely to support a Democratic candidate. Alternatively, a strong Democrat is the "most" Democratic orientation. Ordinal measures add the ability to rank (or "order") to the mutually exclusive and collectively exhaustive traits of nominal level measurement. While such an "advance" is useful, we do not know the difference between the categories. For example, is the difference between a strong Democrat and a weak Democrat the same as between a weak Democrat and an Independent? We have no way of knowing.

The third level of measurement, interval level measurement, retains the mutually exclusive and collectively exhaustive properties of nominal level measurement, and the rank (or order) capabilities of ordinal level measurement. However, interval level measures also contain an equal mathematical difference between categories. For example, supposing a political scientist were trying to explain the likelihood of someone voting. Weather might be a useful predictor variable. While some conceptualizations of weather would be nominal (e.g., it either rained or it did not), temperature would be an interval level measure. Temperature is an interval level measure because there is a constant unit of measure (a degree). Thus, the difference between 39 and 40 degrees is the same as the difference between 70 and 71 degrees. This equal unit of measure is lacking in ordinal level measures.

A measure that has all the properties of an interval level measure (e.g., rank ordering and equal mathematical difference between categories) and has the added property that a score of zero indicates the absence of the phenomena being measured, is called a ratio level measure. For example, since a temperature of zero degrees does not indicate the absence of temperature (i.e., a temperature of zero degrees does not mean that there is no temperature but rather a very cold temperature), temperature is not a ratio level measure (just keep reading). However, if we measure income by dollars, a score of zero does indicate the absence of money (i.e., no dollars). Thus, a score of zero dollars indicates the complete absence of dollars. For this reason, income measured by dollars is a ratio level measure whereas temperature is an interval level measure (the next paragraph will make it clearer).

Ratio level measures are even more useful than interval level measures because, as the name implies, we can form ratios. For example, we can say that someone with an income of $40,000 has twice as much income as someone with an income of $20,000 (i.e., a "ratio" of 2 dollars to 1 dollar). The reason we can say this is that a score of zero implies the absence of money (i.e., no money). However, since a temperature of zero degrees does not mean the absence of temperature we can not say that a temperature of 70 degrees is twice as high (or as warm) as a temperature of 35 degrees.

While ratio level measures are quite common in political science, interval level measures are relatively rare. Political scientists often measure variables in either percentages (e.g., percentage of times a nation resolves its' disputes with other nations by peaceful means - where zero indicates no conflict was resolved peacefully) or other scales in which zero indicates the absence of the phenomena (e.g., a person with zero years of education indicates they have no formal schooling).

A basic rule of statistical analysis is that any statistical technique usable with a lower level of measurement can be used with a higher level of measurement, but not the reverse. For example, if a statistical technique is acceptable for use with nominal level measurement, then it can be also be used with ordinal, interval or ratio level measures. However, if a statistical technique requires an interval level of measurement, then it can not be used with nominal or ordinal level measures.

While the difference between each level of measurement is important, the primary distinction is between the interval and ordinal levels. Thus, we may usefully think of measures as either interval (i.e., interval or ratio) or sub-interval (i.e., nominal or ordinal). As you will read in future assignments, the statistical techniques that require at least an interval level of measurement are much more desirable than those that require only nominal or ordinal levels of measurement. The increased precision that interval or ratio level measurement provides is very useful.

Not surprisingly, researchers have tried using techniques designed for interval level data with ordinal level data. This raises an important question: How serious are the consequences of treating ordinal level data as interval level data? As is often the case, the seriousness of the consequences differ depending upon the gravity of the violation. A good rule in this regard is: The more categories of responses and the more uniform the distribution of responses, the less serious the consequences of violating the interval assumption. For example, it would be preferable to have five categories of responses (e.g., strong Democrat, weak Democrat, Independent, weak Republican, strong Republican) as opposed to three categories of responses (e.g., Democrat, Independent, Republican). We can further minimize the severity of violating the interval assumption by having an approximately uniform distribution of responses. For example, with five categories of responses it would be desirable to have each category contain approximately 20% of the responses. Thus, if 50% of our sample selected response "A" but only 5% selected response "E" we could potentially have serious problems in treating ordinal data as interval data. However, if we have several categories of responses and a relatively uniform distribution of responses, it appears that we can relax the interval assumption without grave consequences (Herbert Asher, Causal Modeling, second edition, pp. 37, 90). In such situations, the advantages of interval level techniques probably outweigh the

consequences of violating the interval assumption.

Two important considerations in evaluating a measure of a variable are validity and reliability. Validity assesses how accurately we are measuring what we claim to be measuring. For example, if our measure of the balance of power says that two nations have the same degree of power, is this really true? A second, and related concept, is reliability. A reliable measure is one which, if applied time after time, will yield the same results (assuming no change in the level, or score, on a particular variable). For example, a reliable measure of unemployment will report the same incidence of unemployment in 2007 as in 2008 if indeed unemployment was the same in those two years.

It is useful to be able to distinguish between validity and reliability. Whereas a valid measure is always reliable, a reliable measure may not always be valid. For example, a gas gauge scale that consistently reports that your car has three more gallons of gas than it actually does is reliable, but not valid. Your car always has three less gallons of gas than the gauge suggests. However, since the gas gauge always reports your car as having three more gallons of gas than it actually does, the gauge is reliable.

Generalizing Our Results

As mentioned previously (pages 2 - 5), one of the tenets of the scientific method is to try and generalize our results. Thus, is what occurs in a city similar to what occurs in a state? At this point it would be useful to discuss what is termed the "unit of analysis." The "unit of analysis" is what we collect data on. For example, if we survey individual voters the "unit of analysis" is the individual. On the other hand, if we are using the unemployment rate for the entire United States, the nation is the "unit of analysis." Suppose we are interested in explaining variation in statewide voter turnout rates. As literate individuals are more likely to read political information (and hence be more political "involved"), a reasonable hypothesis might be that literacy (the independent variable) and voter turnout rates (the dependent variable) are "positively" associated. Thus, as literacy increases, voter turnout rates would be expected to increase. Suppose we have the following statewide data on the percentage of adults who are literate in the state and the percentage of the registered voters who voted in the last election:

State	Percent Literate	Percent Voted
California	90%	90%
Kansas	70%	70%

While tempting, you should not interpret the above data to conclusively support a hypothesis that the more literate an individual is, the more likely they are to vote. For example, the 90% "voted" figure for California could have occurred by having 100% of the 10% illiterate in California vote and only approximately 80% of the 90% of adult Californians who are literate vote. In order to infer to the behavior of

individuals the data should be collected on individuals. This would mean that individuals were surveyed in both of the above states and we knew whether or not each individual was literate and whether that same individual voted. The above data are statewide. Inferring from one "unit of analysis" (here statewide) to another "unit of analysis" (here individuals) is called the ecological fallacy.

Make sure you do not confuse the "unit of analysis" with other concepts previously discussed. For example, do not confuse the "unit of analysis" with the "level of measurement." There is no necessary relationship. We could collect nominal level data on either an individual, a state, or a nation. Furthermore, do not confuse the "unit of analysis" with either a "variable" or the "number of observations". If we are collecting data on the education and political philosophy of 100 individuals, there is one "unit of analysis" (the individual), two "variables" (the individual's level of education and their political philosophy) and 100 "observations" (100 scores on each of the two variables). Moreover, do not confuse the "unit of analysis" with the subject of the study. The subject of the study might be the impact of education (the independent variable) on political philosophy (the dependent variable). The "unit of analysis" is still the individual because the data are collected on individuals. Finally, there is no relationship between the "unit of analysis" and the type of research design the researcher uses. Regardless of whether the researcher uses an experimental or a non-experimental research design, there is always a "unit of analysis." In the above example, if the research could set the level of an individual's education, they would be using an experimental research design. If the researcher could not set the level of an individual's education, they would be using a non-experimental research design. In either event, the individual is still the unit of analysis. Since it is highly unlikely a researcher would be able to either add or subtract years of education from an individual, such a study would invariably use a non-experimental research design.

On quizzes I often ask people to write a hypothesis. Do not write something such as: liberal senators do not support the rich. If you only include "liberal senators," then we have a constant instead of a variable because all the senators you studied would be liberals (just keep reading). Remember that in order to be a variable, a concept (such as political philosophy) must be able to assume more than one value or score (such as: liberal or conservative - two different values or scores). Thus, a better formulation would be: liberal senators are less supportive of wealthy taxpayers than conservative senators. The use of "less" (or "more") conveys the notion of probability. While a liberal senator may support wealthy taxpayers, they are less likely to support them as frequently as conservative senators. Better still, think in terms of a continuum (or gradation) of scores. Thus, all liberals are not equally liberal just as all conservatives are not equally conservative. Therefore, a better phrasing of the above hypothesis would be: the more liberal the senator, the less supportive they are of wealthy taxpayers. The degree of liberalism of the senator is the independent variable while the senator's degree of support for wealthy taxpayers is the dependent variable. This phrasing allows for multiple categories of both liberalism (a senator could be very liberal, somewhat liberal, or not very liberal - i.e., rather conservative, etc.) and support for wealthy taxpayers (much support, some support or perhaps no support).

Descriptive Statistics

The purpose of this section is to introduce you to descriptive statistics. While a quiz on this material may involve some calculation, you do not need to memorize any formulas or bring a calculator to class. Do not panic! The small amount of math that is involved is explained step by step. Honestly, if you can add "5" and "3" you should have little trouble.

The purpose of descriptive statistics is to summarize and illuminate some important characteristics of the data. For example, suppose you listen to a Presidential debate and hear several candidates discuss tax rates. Perhaps one of the candidates mentions tax rates in both the United States and various foreign countries. Perhaps the discussion sparks an interest on your part. So, you decide to assess tax rates in the approximately 160 (or more) nations of the world. What is the average tax rate of all nations of the world? How much variation (difference) is there in tax rates among nations of the world? These are the types of questions that descriptive statistics seek to answer. So the ensuing discussion will make more sense, let me mention that a nation's annual gross national product (GNP) is the total value of goods (e.g., cars, houses, etc.) and services (e.g., teachers' salaries, nurses' salaries, etc.) produced in that nation in a particular year. Descriptive statistics can summarize the data in that they would allow us to make a statement such as the following: In 1995, taxes represented approximately 35% of the value of the gross national product for the average nation. This certainly conveys information in a more useful form than would 160 slips of paper each containing the percentage of the gross national product represented by taxes for a different nation. Thus, we have summarized the data. This becomes even more imperative with larger data sets (e.g., 3000 respondents to a national opinion poll).

A first step in uncovering some salient characteristics of our data is to assess how many and what proportion of the scores are between two points. For example, how many nations have taxes equal to between 20%-29% of their gross national product? Additionally, this number of nations is what proportion of the total number of nations? A frequency distribution is a method of helping us answer such questions. To construct a frequency distribution of national tax rates for 160 nations in the year 2008, we would need to divide the total dollar amount of taxes collected in a nation in 2008, by the dollar value of the gross national product in the same year for the same nation. For example, if all governments in the U.S. collected one trillion dollars in taxes in 2008, and the value of the U.S. gross national product was three trillion dollars in 2008, the tax rate for the U.S. in 2008, would be .33 (1 trillion dollars / 3 trillion dollars = 1/3 = .33). Since .33 is a proportion, we need to multiply it by 100 in order to obtain a percentage (percentage means "per hundred"). As (.33)(100) = 33, this would mean that in 2008 taxes in the U.S. were equal to 33% of our gross national product. Using this same procedure for the additional 159 nations might produce a frequency distribution as follows (if the table on page 15 does not make sense, just keep reading - there is a discussion of the table immediately after the table appears):

Table 1
Frequency Distribution of International Tax Rates
(Hypothetical Data)

National Taxes as a Percentage of National GNP	Percentage of Nations	Number of Nations
70% or more	0%	0
60 - 69	10	16
50 - 59	15	24
40 - 49	40	64
30 - 39	31	50
20 - 29	4	6
0 - 19	0	0
	100%	160

There are several important points to remember in constructing a frequency distribution. First, have a descriptive title. The title should convey to the reader what information the table contains. Second, use appropriately sized categories. For example, suppose the above table had only two categories, less than 30 percent and 30 percent or greater. If so, 96% of the nations would have been in the same category (30 percent or greater). As the data in Table 1 show, such a categorization scheme would have concealed a large amount of variation. Notice the percentage of nations in each category above 39 percent. All this information would have been concealed if one category was used for all nations with taxes equal to 30 percent, or more, of the gross national product. Third, always present both the percentage (column 2) and the frequency (column 3). Percentages "standardize" the data (just keep reading - it will be clear shortly). For example, suppose that we wanted to compare how likely a nation was to have taxes consume between 50% and 59% of that nation's gross national product in 2008 and in 1958. According to the table above, in 2008 there were 24 nations out of the total of 160 nations, or 15%, (24 is 15% of 160) that fit this criteria. Suppose we found that in 1958, 24 nations out of the total of 120 nations, or 20%, (24 is 20% of 120) fit this same criteria. While the number of nations, 24, is the same in both 2008 and 1958, the percentage of nations in which taxes consumed between 50% and 59% of the gross nation product was lower in 2008 than in 1958 (because 15% is lower than 20%). If we did not adjust for the different number of nations in the two time periods (160 vs. 120), we would not have known this. Expressing the number 24 as a percentage placed both 2008 and

1958 on a common measuring scale which then revealed a difference between the two years that would not have been apparent had we just examined the number of nations in the 50% to 59% category in both years. As the reader may want to reorganize the data, it is also important to show the frequency (i.e., number of nations) in each category.

Measures of Central Tendency and Measures of Dispersion

While frequency distributions are important, they do not convey all the salient characteristics of a variable. Two additional questions that would be useful to answer are: (1) What is the average score; and (2) How representative is the average score of all the scores?

We often think in terms of an "average." You may assess how well you scored on a test by comparing your score with "the average." There are several different measures of "the average." The statistics pertaining to the average are termed "measures of central tendency." There are three commonly used measures of central tendency. The appropriate measure depends upon three factors: (1) the question you want to answer; (2) the level of measurement of your data; and (3) the distribution of the scores. Remember from the notes on the level of measurement that while a nominal level measure categorizes the data (e.g., pear, oak and elm are three categories of trees), it neither orders the data (e.g., are pear trees "higher" than oak trees?) nor provides an equal unit of measure between categories (e.g., is the difference between pear and oak the same as between oak and elm?). Given these limitations, the only method of measuring the average for a nominal level measure is to see which category occurs the most frequently. Such a measure is called "the mode." For example, the mode for the following scores: 2, 3, 4, 5, 5, and 6 is 5. If two different scores are equally numerous then we have what is termed a "bi-modal" distribution (just keep reading - the next sentence will make it clear). For example, suppose we had the following scores: 1, 1, 2, 3, 4, 5, 5. Since "1" and "5" both appear twice and no other score appears more than once, both "1" and "5" are modes. As there are two modes, the scores are "bi-modal."

If our data are ordinal (on "ordinal" see page 11), in addition to the mode, we may also employ a second measure of central tendency, the median. The median is the number that divides the distribution into two equal parts. For the following 25 scores: 0, 1, 1, 2, 2, 2, 2, 3, 3, 3, 4, 4, 5, 6, 6, 6, 6, 6, 7, 8, 9, 75, 100, 125 and 300 the median is the "middle" score (i.e., if 25 scores then the 13th score from the left) which is 5. Note also that the mode for this data is 6. If we had an even number of scores the median would be the mid-point between the two middle scores. For example, with ten scores the median would be midway between the fifth and sixth scores.

Look again at the 25 scores that we just used in the example of the median. Notice that the last four scores (75, 100, 125 and 300) are much higher than the other scores. However, if a score is the same (or higher) than the middle score, the median is unaffected (e.g., if the last four scores were 14, 15, 16 and 17 the median would still be 5). It is this feature of the median that makes it usable with ordinal level data. Remember that ordinal level data tell the rank, but not the precise numerical

difference between the scores. Thus, if we just want to select the middle score, all we have to know is the ranking of the scores, not the degree of difference between the scores. However, if we have interval or ratio level data (see pp. 11-12), we also know the numerical difference between the categories. It would seem natural to develop a measure of the average that took account of the numerical differences between the scores. This is the idea behind the mean. The mean of a group of scores is the total of the scores divided by the number of scores. For example, the mean of the following scores: 10, 20 and 60 is 30 (because 10 + 20 + 60 = 90 and 90/3 = 30). Hence, the mean tells us the average per score.

In the calculation of the median, we used the 25 scores listed above. The mean of these 25 scores is 27.44 (because if you add the 25 scores above together they total 686 and 686/25 = 27.44). Obviously, the median (5) is quite different than the mean (27.44) for this data. This great difference between the median and the mean occurs because the highest four scores are so much greater than the other 21 scores. While the size of the numerical difference between scores has no effect on the median, it has a great impact on the mean. Remember that one of the criteria for choosing a measure of central tendency is the distribution of the data. Because the extreme scores distort the mean value of the 25 scores listed on the previous page, I would recommend using the median as the measure of central tendency for the above data set. However, note that if we had not had interval or ratio level data, we could not have calculated the mean. Hence, we would have had to use either the median or the mode as our measure of central tendency. Additionally, if you wanted to know which score occurred the most frequently, you would use the mode as your measure of central tendency.

By providing a notion of "the average," measures of central tendency reveal important characteristics of the data. However, a measure of central tendency does not tell us how representative the average is of all the scores in the distribution. For example, if you were a professor and found that the mean score on an exam was 50, I think you would want to know if the mean of 50 occurred because nearly all students scored approximately 50, or because perhaps half the students scored 0 and the other half scored 100. In both circumstances the mean would be the same, but the ramifications for how you taught the course would be entirely different. This example illustrates the need for what are termed "measures of dispersion." The purpose of measures of dispersion is to assess how representative the average is of all the scores in the distribution.

The simplest measure of dispersion is called "the range." The range is the difference between the highest and lowest score. As the range implies order, we would need at least an ordinal level of measure to compute the range. While the range is simple, it does not convey as much information as we might like. For example, suppose we return the glorious 1980s when American life was good, pure and Donald Trump was in financial ascendence. Relaxing in the plush surroundings of Trump Tower, I joyously calculate Sir Donald's income for the current year. When I compare his income to that for each other American family, I discover that not only is Trump's income higher than that for any other American family, but that no other American family is within $1,000,000 of Trump's income. If we were to use the range as a measure of dispersion, we would simply take the difference between Trump's

income and the lowest family income. However, by using only two scores we have masked the important information that Trump's income was by far the highest. Thus, the range is insensitive to any scores except the two most extreme (highest and lowest). We can minimize this problem somewhat by using several additional data points. For example, in addition to the highest and lowest scores we could include the score wherein 1/4 of the families were above and wherein 1/4 of the families were below. Such a measure is referred to as the "interquartile" range.

While the interquartile range conveys more information than the range, it would be desirable to have a measure of dispersion based upon all the scores in the distribution. Intuitively, we might think that we could measure dispersion by subtracting the mean from each score. The difference between a single score and the mean would indicate the amount of "deviation," or "dispersion" of that particular score from the mean. If we summed (added) each of these "dispersions" we would have the total amount of dispersion present in our data. We could then divide this total by the number of observations in order to obtain a typical deviation (i.e., the deviation per score). While there is a mathematical "problem" in this method, the procedure we have just outlined is the basis of the approach we will ultimately use.

The mathematical "problem" with the approach we just formulated is that it must result in an answer of zero. This is because one of the properties of the mean is that the total amount of "distance" below the mean must be equal to the total amount of "distance" above the mean. For example, the mean of the following scores: 4, 6, and 8 is 6 (because 4 + 6 + 8 = 18 and 18/3 = 6). The "average" of the deviations of these same scores would be "0" [because 4 - 6 = -2, 6 - 6 = 0, 8 - 6 = 2; adding these deviations equals 0 (i.e., -2 + 0 + 2 = 0) and then dividing this total by the number of scores yields "0" (i.e., 0/3 = 0)]. The value of the "positive" deviations from the mean (i.e., scores higher than the mean) equals the value of the "negative" deviations from the mean (i.e., scores lower than the mean) and thus the "total" deviation from the mean must equal zero. Therefore, we are left with the impression that there is no deviation (i.e., no variation) in our data. Thus, we would mistakenly conclude that the mean occurred because every score was the same.

This problem can be rectified by taking the "absolute" value of each deviation from the mean. Hence, a deviation of -2 would be treated as a deviation of 2. As we would be adding a series of positive numbers, the cancellation problem I just discussed would not occur (i.e., in the example above remember that -2 + 0 + 2 = 0 and hence our deviation measure ended up as 0/3 = 0; taking "absolute" values would have instead produced a typical deviation of 1.33 because 2 + 0 + 2 = 4 and 4/3 = 1.33) This revised procedure (i.e., using "absolute" values) is called the average deviation. While such a measure tells us the average amount of deviation from the mean for a typical score, the result is still not as useful as we might prefer. For example, suppose we calculated the average deviation and found that it was 5.7. What statement(s) could we make? We could say that the typical score deviated 5.7 units (in whatever units the variable was measured, dollars, percentage points, etc.) from the mean.

We could further amplify the preceding approach by taking the average deviation as a percentage of the mean (even if you are confused, just keep reading through the end of this paragraph). Thus, an average deviation of 5.7 would seem

small if the mean were 1,000, but large if the mean were 10. Therefore, an obvious approach would be to divide the average deviation by the mean (just keep reading). This would yield a "proportion" which we could then multiply by 100 in order to obtain a "percentage" (just keep reading). For example, an average deviation of 5.7 is 57% of a mean of 10 (5.7/10 = .57 and (.57)(100) = 57). However, an average deviation of 5.7 is only .57% (approximately 1/2 of 1%) of a mean of 1000 (5.7/1000 = .0057 and (.0057)(100) = .57). If the average deviation is 57% of the mean, it tells us that the mean was achieved by many scores being quite far from the mean (e.g., the mean of 0 and 10 is 5 because (0 + 10)/2 = 10/2 = 5 but neither 0 nor 10 is very close to 5). On the other hand, if the average deviation is only 1/2 of 1% of the mean, this tells us that virtually all the scores are quite close to the mean.

While the aforementioned procedure is a definite improvement over the range, we can still do better. A critical question to ask with the result from any statistic is: How can you interpret the answer? In the last paragraph I outlined an approach to interpreting the average deviation. However, we could improve upon my approach if, in addition to taking the average deviation as a percentage of the mean, we could also compare the average deviation to a known distributional formula. For example, suppose the mean of a group of scores was 50 and the average deviation was 5. Using my approach we could say that the average deviation was 10% of the mean [5/50 = .10 and (.10)(100) = 10]. This would suggest that the mean was achieved by most scores falling pretty close to the mean. Put differently, when we consider the average deviation relative to the mean, dispersion seems low. However, it would be even more informative to give a percentage distribution of the scores. For example, it would be desirable to be able to say that approximately 68% of the scores were within one average deviation of the mean. Given our results (mean = 50, average deviation = 5) this would mean that approximately 68% of the scores would fall between 45 and 55. Unfortunately, we can not make such a statement with the average deviation. In order to obtain a percentage distribution of the scores we need to calculate the standard deviation.

While our purpose is to illustrate the calculation and use of the standard deviation, this is a good point to introduce some statistical symbols and mathematical procedures. Do not panic! While a quiz on this material may involve some calculation, you do not need to memorize any formulas or bring a calculator to class. I will provide any formulas which are needed on the quiz. You only must know how to work the formulas on pages 21-24. Just follow how I work the computations on pages 21-24 (i.e., how I calculate the various answers). Finally, you will not have to calculate a square root. The quizzes on this material will have you working largely with single digit numbers. For example, you might have to subtract 3 from 5 (i.e., 5 - 3 = 2). Wasn't that difficult? If we are feeling "adventurous" later on we might actually try multiplying 2 times 2!! Wow!!! Do you really need a calculator to do these computations? I hope not! When people have trouble on a computational quiz it is almost always because they do not know the order of mathematical operations. For example, do you add and then multiply or multiply and then add? A calculator can not help you with the order of mathematical operations. The calculator assumes you know the order of operations. That is why a calculator is of such little value on the type quiz you will take.

The standard deviation is undoubtedly the most often used measure of dispersion in political science. As just mentioned, the standard deviation will permit us to make what I will term "percentage distribution" statements. Just keep reading!! For example, let us say that you are studying international relations. International relations scholars have often quantitatively tested interesting models concerning the factors (i.e., independent variables) that explain why some nations are more likely to resolve their conflicts peacefully than other nations. The dependent variable in this type of study might be the percentage of times a nation resolves its disputes peacefully. Peaceful resolution would mean that a dispute was resolved without violence. Typically, international relations scholars pretty much agree on what constitutes a "dispute." So, through historical records we construct the number of disputes that each nation has been involved in over the last 100 years. While the name and boundaries of many nations have changed over the past 100 years, we could still obtain data for a large number of nations. Using the historical record, (continued on the next page)

NOTE: Odd page breaks will occur when there are formulas. This material was originally written in Word Perfect and the equations do not automatically transfer from one word processing package to another. I don't know how to use equation "boxes" in Microsoft Word. Additionally, in some cases there are drawings that I paid to have made that I want to transfer to this edition of the material. So, just be prepared for an occasional page fragment!

we calculate the percentage of disputes which each of these nations resolved in a peaceful manner. To simplify the presentation, let us examine just two regions of the world. If each of the two regions contained 5 nations, the data might look something like the table below.

Percentage of Disputes Resolved Peacefully

Region 1		Region 2	
Nation A	50%	Nation F	68%
Nation B	60%	Nation G	69%
Nation C	70%	Nation H	70%
Nation D	80%	Nation I	71%
Nation E	90%	Nation J	72%

Like the average deviation, the standard deviation is based upon how each score deviates (hence "deviation") from the mean (i.e., the average score). Accordingly, before we can calculate the standard deviation, we must first calculate the mean. Do not "freak" when you see "strange" symbols! It is all explained over the next page and a half. So, just keep reading! The symbol and formula for the mean are shown immediately below.

$$\textit{Mean of Variable } X = \bar{X} = \frac{\Sigma X}{N}$$

$\bar{X}$ is the symbol for the mean of variable X. If we had designated this variable as variable Y, then the symbol for the mean would, not surprisingly, be $\bar{Y}$. The Σ symbol above is called a "summation operator." The "summation operator" tells us to add the scores. "N" is the symbol for the "number of scores." To obtain the mean score for region 1, the computation would be as follows:

$$\textit{Mean of Region 1} = \bar{X}_1 = \frac{\Sigma X_1}{N} = \frac{50 + 60 + 70 + 80 + 90}{5} = \frac{350}{5} = 70$$

To obtain the mean score for region 2 the computation would be as follows:

$$\textit{Mean of Region 2} = \bar{X}_2 = \frac{\Sigma X_2}{N} = \frac{68 + 69 + 70 + 71 + 72}{5} = \frac{350}{5} = 70$$

The mean score for both regions is the same, 70. This means that in both regions the average nation resolved disputes peacefully 70% of the time. However, while the mean score for both regions is the same, it is obvious from the data [redacted] that the scores varied more (i.e., were more different) in region 1 than

in region 2. Put another way, the mean in region 2 is more representative of all the scores in region 2 than the mean in region 1 is of all the scores in region 1 (thus the mean of 70 is closer to numbers like 71 and 72 as in region 2 than 70 is to numbers like 50 and 90 as in region 1). The standard deviation is a method of summarizing how much a group of scores differ. For example, how much do the scores in region 1 differ from each other? The standard deviation helps us answer such a question. Do not "freak" when you see the following formula for the standard deviation. Each step will be carefully explained.

$$\text{Standard Deviation of } X = S_x = \sqrt{\frac{\text{Variation of } X}{N}} = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}}$$

Do not worry if you get "lost" on the operations below. Just keep reading! You will see them performed shortly. Here is the order of operations in the above formula:

1. The order of mathematical operations says to perform the computations inside the parentheses first. It is clear that we are to subtract the mean ($\bar{X}$) from something. Obviously, we would first have to know what the mean was in order to subtract it. Therefore, our first step is to calculate the mean. We did this on page 22. We know that the mean score for both regions is 70.

2. The symbols ($X-\bar{X}$) tell us to subtract the mean (70) from each score.

3. After working inside the parentheses, we move outside the parentheses. The symbol "2" in the upper right hand side of the expression $(X-\bar{X})^2$ tells us to square each of the entries in step 2. Since there are five nations in region 1, we would have 5 "entries" for step 2 (i.e., subtracting the mean from nation A, then subtracting the mean from nation B, etc.). We would now square each of these five entries. Squaring means to multiply a number times itself. Thus, take nation A. Look at the table on page 20. The score for nation A is 50. If we are subtracting the mean (70) from this score it reduces to 50-70 which is -20. To square -20 we multiply -20 times -20. This is also stated as (-20) (-20). A negative number multiplied, or divided, by a negative number yields a positive answer. So (-20) (-20) will produce a positive answer, 400. The answer, 400, could be more simply attained by decomposing one of the -20^{s}. Thus, (-20)(-20) = (-20)(-10)(2) = (200)(2) = 400. So, our first squared deviation is 400. We repeat this process for each of the other four nations in region 1.

4. The summation operator tells us to add each of the five square deviations we obtained in step 3. It is critically important that we square before we sum. Watch for a quiz the first day this material is due that will test whether you know to square before you sum.

5. We divide the total we obtained in step 4 by the number of scores we added (i.e., 5 - since there are five nations in each region).

6. Take the positive square root of the answer in step #5. Squares and square roots are opposites. The sign for a square root is: $\sqrt{}$ Thus 4 "squared" is 16 because (4) (4) = 16. Additionally, 4 is the square root of 16. Note that (-4) (-4) also equals 16. Thus, both (4) and (-4) are square roots of 16. In calculating the standard deviation always take the positive square root (i.e., 4 instead of -4).

The table below shows the computations necessary to calculate the standard deviation of region 1 and region 2. The scores in the column marked "X" are the same scores as appeared on page ~~20~~ 22. The symbols at the top of each column are those discussed on pp. ~~20-21~~ 22-23.

Standard Deviation of Region 1

Column 1	Column 2	Column 3	Column 4
Nation	X	$(X - \bar{X})$ the above says subtract the mean (70) from each score in column 2	$(X - \bar{X})^2$ this column is the square of the answer in column 3
A	50	(50 - 70) = -20	400 [because: (-20)(-20) = 400]
B	60	(60 - 70) = -10	100
C	70	(70 - 70) = 0	0
D	80	(80 - 70) = 10	100
E	90	(90 - 70) = 20	400

Σ = 1,000

(Σ means to add the numbers in column 4, thus 400 + 100 + 0 + 100 + 400 = 1,000)

$$\textit{Standard Deviation of Region 1} = S_{X_1} = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}} = \sqrt{\frac{1000}{5}} = \sqrt{200} = 14.14$$

5 nations

Note: 14.14 is the positive square root of 200 because 14.14 is a positive number (i.e., 14.14 instead of -14.14) and (14.14) (14.14) = approximately 200. You will not have to calculate a square root on the quiz.

Standard Deviation of Region 2

Column 1	Column 2	Column 3	Column 4
Nation	X	$(X - \bar{X})$	$(X - \bar{X})^2$
F	68	$(68 - 70) = -2$	4 [i.e., $(-2)(-2) = 4$]
G	69	$(69 - 70) = -1$	1
H	70	$(70 - 70) = 0$	0
I	71	$(71 - 70) = 1$	1
J	72	$(72 - 70) = 2$	4
			$\Sigma = 10$

$$\text{Standard Deviation of Region 2} = S_{X_2} = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N}} = \sqrt{\frac{10}{5}} = \sqrt{2} = 1.41$$

Notice that the standard deviation for region 1 (14.14) is 10 times the size of the standard deviation for region 2 (1.41). One useful method of interpreting the standard deviation is as a percentage of the mean. In region 1 the standard deviation is 20% of the size of the mean (i.e., 14.14/70 = .20 and .20 x 100 = 20). In region 2 the standard deviation is a mere 2% of the mean (i.e., 1.41 is 2% of 70). We could say that the scores were more "dispersed" (i.e., on average further from the mean) in region 1 than region 2. Expressing the standard deviation as a percentage of the mean is called the coefficient of variation. If the standard deviation is 15%, or less, of the size of the mean, we could say dispersion is low (i.e., the typical score is rather close to the mean). If the standard deviation is from 16% to 35% of the size of the mean we could say there is a moderate amount of dispersion. If the standard deviation is greater than 35% of the size of the mean we could say there is a high degree of dispersion (i.e., the typical score is somewhat far from the mean or, alternatively, the mean is not very representative of the typical score).

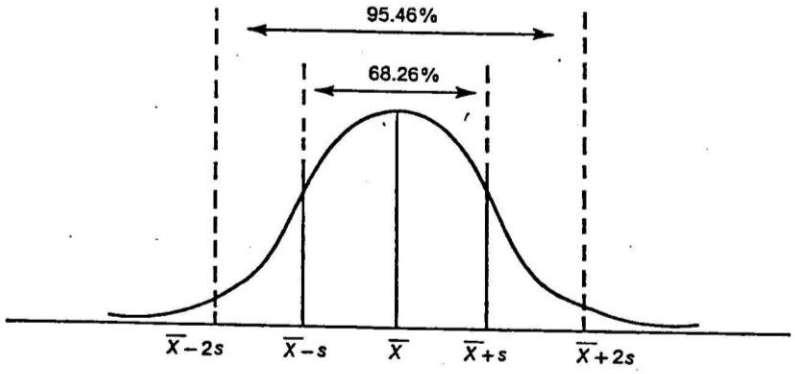
Another measure of dispersion that is very similar to the standard deviation is the "variance." The variance is literally the square of the standard deviation. For example, the variance for region 2 is 2 because 1.41 (the standard deviation of region 2) squared (i.e., 1.41 times 1.41) equals approximately 2.
Put another way, when we calculated the standard deviation, the result of the next to last step (i.e., just before we took the square root) was the variance. The formula for the variance appears below. Except for the square root sign, isn't it identical to the formula for the standard deviation? Yes!

$$\text{Sample Variance of } X = S_x^2 = Var(X) = \frac{\text{Sample Variation of } X}{N} = \frac{\Sigma(X-\bar{X})^2}{N}$$

Since I mentioned that you would not have to calculate a square root and that you would not need a calculator, a possible quiz question might be to give you the formula for the variance and a small group of scores that were easy to work with (e.g., numbers

like 2, 4, etc.) and ask you to calculate the variance.

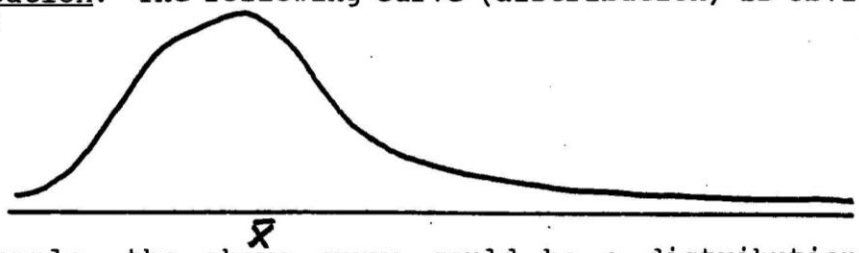
As I mentioned previously, the standard deviation can also be interpreted by using what I termed a "percentage distribution" capability (e.g., 68% of the scores are within - plus or minus - one standard deviation of the mean). Just keep reading it will become clear shortly!!! In order to understand this capability, let us introduce one of the very most important concepts in statistics, the normal distribution. A distribution is a group of scores. For example, because the readings in this course are so exciting (sure!), suppose you decide to go to graduate school in political science. Most doctoral programs in political science require you to take the Graduate Record Examination (GRE). Let us say we took all of the scores on the Graduate Record Examination for a particular year. A diagram of the scores might well look like the drawing below. Note that the height of the line indicates the frequency of occurrence. Thus, in the diagram below, the mean (X) is the most frequently occurring score (i.e., the "mode") because the curve is highest directly over the mean score.



Several other properties of the normal distribution are also important to mention. First, the distribution is "symmetrical" (i.e., the portion to the left of the mean is of the identical shape of the portion to the right of the mean). Second, the mean, median and mode are all at the same point (i.e., same score). Third, approximately 68% of the scores are between one standard deviation above the mean (i.e., "$\overline{X}$ + s" in the above diagram) and one standard deviation below the mean (i.e., "$\overline{X}$ - s" in the above diagram). For example, if the distribution of scores on the GRE was normal (i.e., shaped as the drawing above), then if the mean on the GRE was 1,100 and the standard deviation was 100, approximately 68% of the scores would be between 1,000 and 1,200 (because 1,100 - 100 = 1,000 and 1,100 + 100 = 1,200). Fourth, approximately 95% of the scores would be between two standard deviations below the mean and two standard deviations above the mean. In our example, this would mean that 95% of the scores on the GRE would be between 900 and 1,300 (because 1,100 - 100 - 100 = 900 and 1,100 + 100 + 100 = 1,300). Furthermore, over 99% of the scores would fall within plus or minus three standard deviations of the mean. This is the "percentage distribution" capability of the standard deviation. The average deviation, which was discussed on pages 18-19, does not have such a capability. This is one of the major reasons why the standard deviation is by far the most commonly used measure of dispersion in political science.

Tchebysheff's Theorem

The scores on a variable may not always conform to a normal distribution. The following curve (distribution) is obviously non-normal:



For example, the above curve could be a distribution of wages (e.g., many low - to the left of the "mean," and a few high - to the right of the "mean"). The curve is said to be "skewed" to the right. Note that the total amount of area on each side of the mean is the same. It is just "stretched" more to the right of the mean (or more "dense" to the left of the mean). By contrast, notice that the normal curve (page 26) is symmetrical (i.e., the portion of the curve on each side of the mean has the same shape).

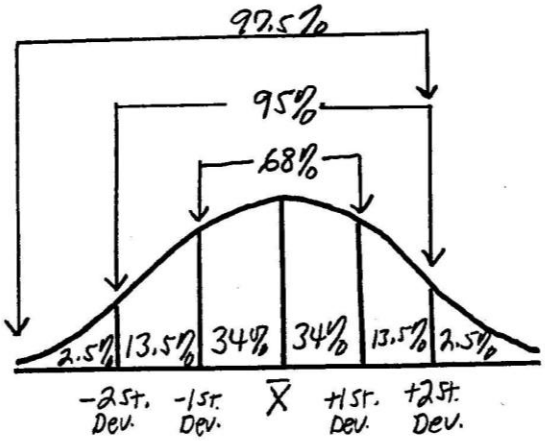
In order to use the percentage distribution capabilities of the standard deviation when dealing with a non-normal distribution, we use Tchebysheff's Theorem. Tchebysheff's Theorem states that 75%, or more, of the scores must be within two standard deviations of the mean and 88%, or more, of the scores must be within three standard deviations of the mean. Thus, regardless of the shape of the distribution of the scores, we can use the percentage distribution capabilities of the standard deviation. As many variables are not normally distributed, Tchebysheff's Theorem makes the standard deviation much more useful. Note that Tchebysheff's Theorem does not imply that a distribution is symmetrical. For example, as in the curve drawn above, we may have more scores below than above the mean. In general, the scores above the mean are further from the mean than the scores below the mean. Thus, few scores further from the mean represent as much "area" under the curve as many scores closer to the mean. Note that the normal curve is consistent with Tchebysheff's Theorem. Tchebysheff's Theorem states that 75%, or more, of the scores must be within two standard deviations of the mean. With a normal curve 95% of the scores are within two standard deviations of the mean (see page 24). Isn't 95% either equal to, or greater, than 75%? Yes, 95% is greater than 75%.

Z (or Standard) Scores

Occasionally, you may want to compare scores from distributions with different means and standard deviations. For example, suppose that you apply to graduate school in political science and one of the schools you apply to gives the option of submitting either a score on the Graduate Record Exam or the Miller Analogies Test. Let us say that you scored 1,200 on the Graduate Record Exam and 50 on the Miller Analogies Test. Which is the "better" score? Thus, which score should you submit? It might be a mistake to "automatically" assume that the "higher" score (i.e., the Graduate Record Exam score) is the "better" score. A good way to see which is the "better" score is to compare each score to the mean score on that particular test. This is the essence of what is called a "Z" (or "standard") score. For example, if scores on the Graduate Record Exam are normally distributed with a mean of 1,100 and a standard deviation of 100, your score of 1,200 is one

standard deviation higher than the mean (i.e., 1,200 = 1,100 + 100). However, if the Miller Analogies Test has a mean of 40 and a standard deviation of 5, your score of 50 is <u>two</u> standard deviations above the mean (i.e., 50 = 40 + 5 + 5). Let us think of this in terms of the normal distribution. As we discovered on page 24, approximately 68% of the scores in a normal distribution are within either one standard deviation above or below the mean. Since, the normal curve is symmetrical (meaning each half of the curve is of the same shape), half of this 68% of the scores (or 34%) are below the mean and half are above the mean. Thus, 34% of the scores are between the mean and one standard deviation above the mean. Remember that your score on the Graduate Record Exam was also one standard deviation above the mean.

<u>To assess how well you scored on the Graduate Record Exam we could use the following analysis: (1) since you scored above the mean, we know that at least 50% of the scores were lower than yours (i.e., the 50% that were either at the mean level or below); (2) as 34% of the scores are between the mean and one standard deviation above the mean, then approximately 84% of the scores were below yours (i.e., the 50% that were at the mean level and below plus the 34% that were between the mean and one standard deviation above the mean</u>. Since your score on the Miller Analogies Test was two standard deviations above the mean, the analysis would be as follows: (1) since you scored above the mean, we know that at least 50% of the scores were below yours; (2) since 95% of the scores are between two standard deviations above and below the mean and the normal distribution is symmetrical, then approximately 47.5% of the scores (i.e., half of 95%) are between the mean and two standard deviations above the mean; (3) therefore, approximately 97.5% of the scores on the Miller Analogies Test are below your score (i.e., the 50% that were at the mean level and below plus the 47.5% that are between the mean and two standard deviations above the mean). The diagram below pictures the aforementioned analysis.



Although your "raw" score on the Graduate Record Exam (1,200) is much higher than your "raw" score on the Miller Analogies Test (50), the Miller Analogies Test was your "better" score relative to those who took each test. The formula for the "Z" (or standard score) utilizes the reasoning we just followed. A "Z" score is simply how many standard deviations a score is above the mean (if the Z score is positive) or below the mean (if the Z score is negative). The formula is shown below.

$$Z = \frac{X - \overline{X}}{\textit{Standard Deviation of } X}$$

In this example, your Z score on the Graduate Record Exam was 1.0 [because (1,200 - 1,100)/100 = 100/100 = 1] and your Z score on the Miller Analogies Test was 2.0 [because (50 - 40)/5 = 10/5 = 2]. Since 2.0 is greater than 1.0, your Miller Analogies score is your better score. Here are some important points to remember: a positive Z score means the score is above the mean (i.e., above average); a Z score of zero means the score equals the mean (i.e., average); a negative Z score means the score is below the mean (i.e., below average). Just keep reading, examples are ahead.

Suppose, for example, that your Miller Analogies score was 35 (instead of 50). This would have resulted in a Z score of -1.0 [because (35 - 40)/5 = -5/5 = -1] and placed you in the 16th percentile [i.e., in the diagram on page 26 notice that only 16% (2.5% + 13.5% = 16%) of the area under a normal curve is further than one standard deviation below the mean]. A score of 40 on the Miller Analogies Test would have yielded a Z score of 0 [(40 - 40)/5 = 0/5 = 0)] and placed you at the 50th percentile (i.e., half scored higher and half scored lower - sounds like the median - which is also the same point as the mean and mode in a normal distribution). A score of 45 on the Miller Analogies Test would have resulted in a Z score of 1.0 [(45 - 40)/5 = 5/5 = 1] and would have placed you in the 84th percentile (above both the 34% between your score and the mean score and the 50% who score at, or below, the mean: 34% + 50% = 84%).

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.

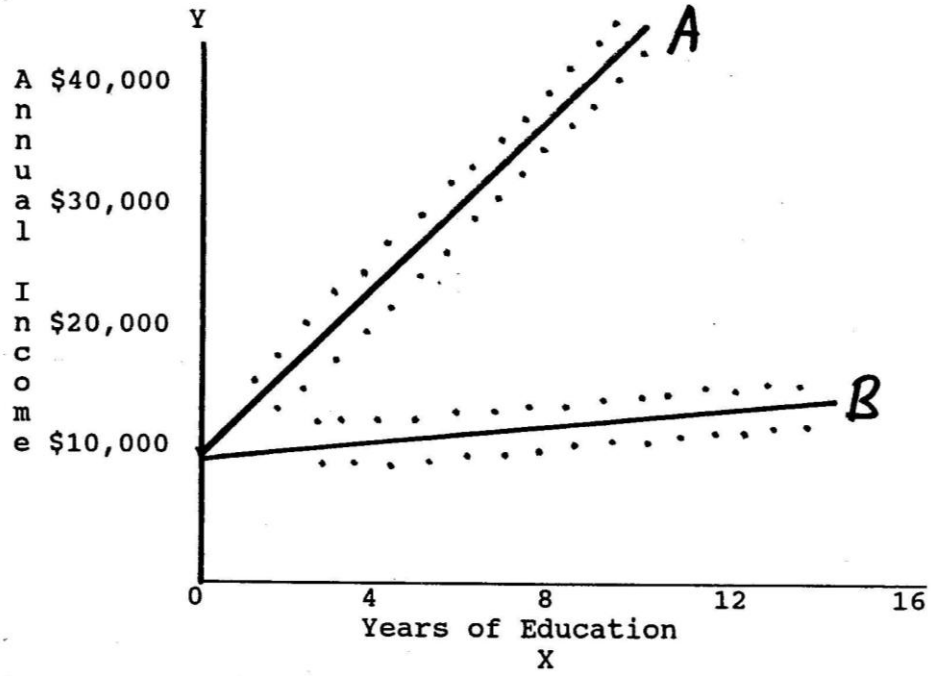
Second, even if we use a cross tabulation table with many variables and many categories of responses per variable the small number of observations in many of the cells make significance testing (your next statistically oriented reading assignment) extremely risky (just keep reading). Suppose our sample 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!

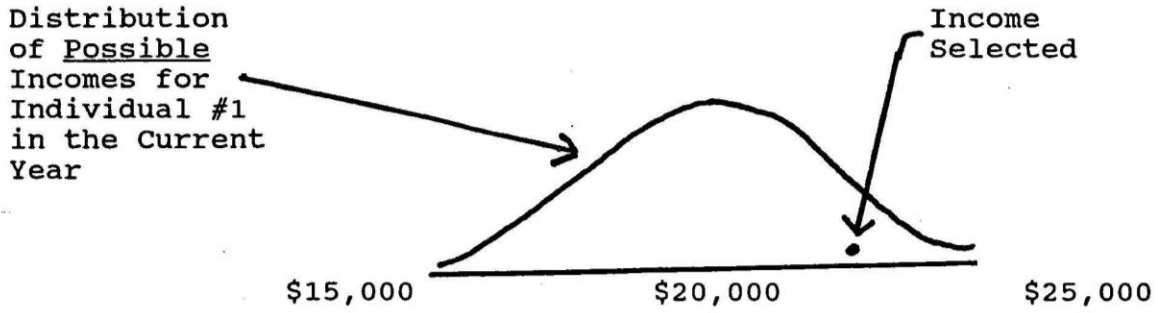
> We continually refer to the set of all possible outcomes as the population and to the processes underlying the outcomes as the population model. If we are interested in people's incomes and the relationship between incomes and education, ... the possible outcome (income) for any individual in any year 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.

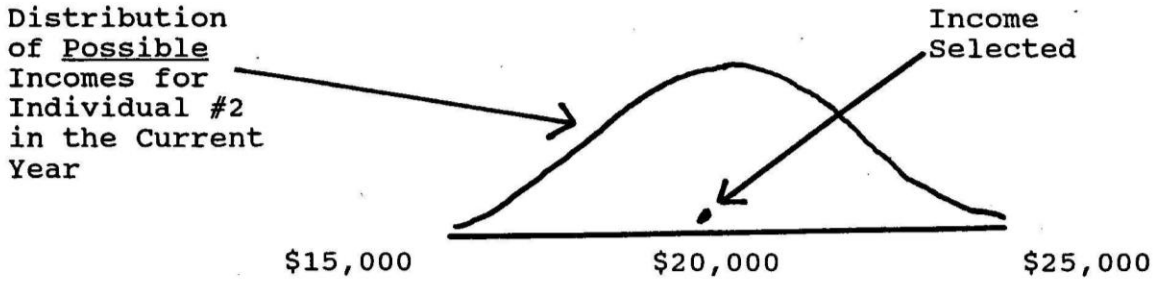
In most all research situations we have data on a sample, not a population. Even when we think we have data on all members of the population we usually have a sample because typically the "population" is infinite. For example, if we have data on the 100 U. S. Senators who served in a particular year, 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

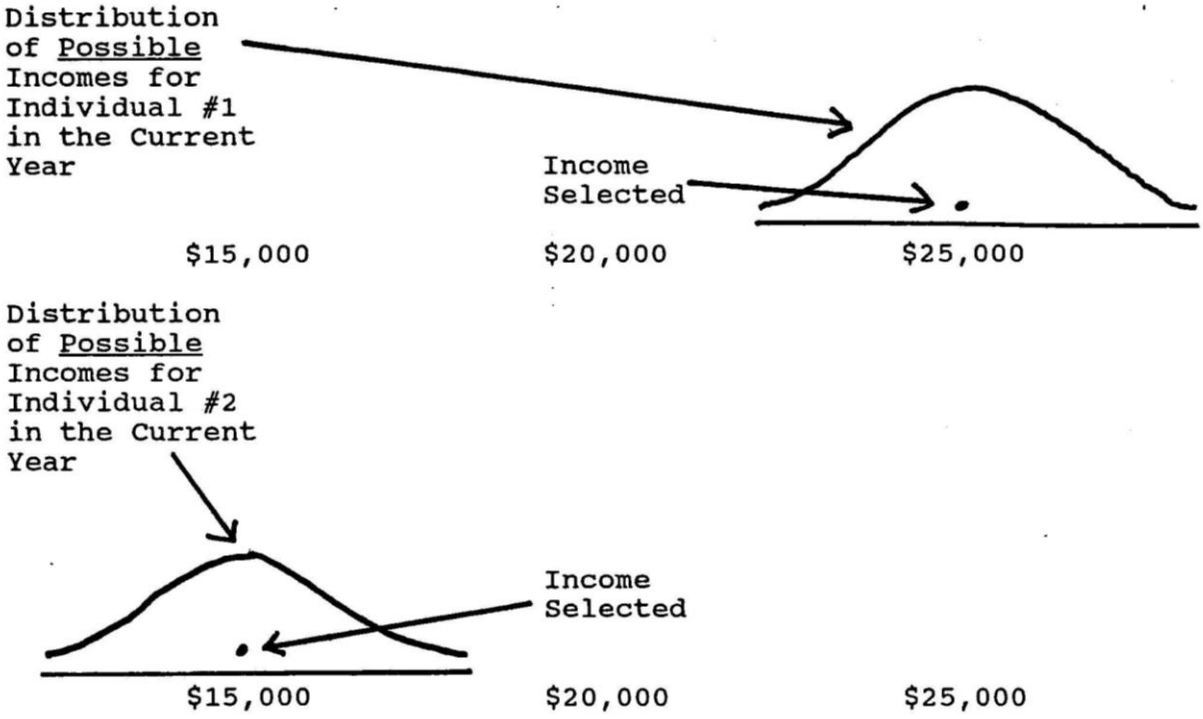




While the income selected (i.e., that actually occurred) for individual #1 (approximately $23,000) is higher than for individual #2 ($20,000), the distribution of income for both individuals is the same. Thus, the two distributions have the same mean income ($20,000) and the same standard deviation. 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."

Multivariate Analysis

When political scientists build a model to explain and/or predict behavior they are typically trying to accomplish two tasks: (1) formulate an accurate model of the behavior in question; and (2) estimate how much impact each independent variable has on the dependent variable. For example, if we want to examine why some senators vote more in favor of tax changes benefiting moderate and low income households than other senators, we need to think through what factors are likely to impact a senator's votes on this issue. The dependent variable in such a model (i.e., what we are trying to explain or predict) might be measured by the percentage of times the senator voted in favor of reducing after-tax income inequality. Reducing after-tax income inequality would consist both of voting in favor of tax changes where over 50% of the benefits go to households with incomes equal to, or less than, the median income (i.e., if we took a 101 families ranked ordered from highest income – household #101 – to lowest income – household #1 this would mean over 50% of the benefits go to households 1 through 51) and voting against tax changes where over 50% of the benefits go to households with incomes above the median (i.e., to households 52 through 101). In the analysis that follows this variable appears as "tax."

More specifically, "tax" represents the percentage of times a senator votes in favor of the immediate self-interest of households with incomes equal to, or less than, the median income. Thus, if the computer reads a score of 62 for tax this means that the senator voted in favor of the economic interest of households with incomes equal to, or less than, the median household income 62% of the time. With the help of a public finance professor, I calculated scores for each senator on a major tax reform. This discussion involves those scores.

Now we need to think through what factors (i.e., independent variables) might influence how frequently (i.e., the percentage) of times a senator would vote for tax changes primarily benefiting households with incomes equal to, or less than, the median. Three factors come to mind: (1) the philosophy of the senator; (2) the party affiliation of the senator; and (3) the wealth of the state the senator represents. Since conservatism is associated with a strong belief in allowing the market to determine relative living standards (e.g., resistance to creating or increasing the minimum wage) and specifically with the government establishing economic rights for citizens (e.g., a right to health care), a reasonable first hypothesis is that the more conservative the senator the lower their support for tax changes where over 50% of the benefits go to households at, or below, the median income (i.e., a "negative" association – *higher* scores on conservatism associated with *lower* scores of tax changes primarily benefiting middle and low income groups). In the analysis to be presented later, we will measure a senator's conservatism by the percentage of times they vote in favor of positions taken by a conservative interest group, the Americans for Constitutional Action. Scores can range from 0% to 100% with higher scores indicating a more conservative voting record. In the analysis that follows this variable is appears as "scons" (for senator conservatism).

Because households with incomes below the median income makeup a greater percentage of the vote of the Democratic party than the Republican party, a reasonable second hypothesis is that Democratic senators will support tax

changes where over 50% of the benefits go to households with incomes at, or below, the median household income a greater percentage of the time than will Republican senators. In the analysis ahead we will measure a senator's political party affiliation with what political scientists refer to as a "dummy" variable (i.e., can only take on two values – in this case Democratic senators are coded "1" and Republican senators "0"). This variable appears as "party."

As state economic self-interest might be a factor in why senators vote in the manner they do, a reasonable third hypothesis would be that the higher the median income level in a state (i.e., the wealthier the state) the less likely the senator will vote in favor of tax changes where over 50% of the benefits go to households at, or below, the median income. In the analysis ahead, state median income is in thousands of dollars (i.e., if the computer reads a score of 35.2 it means in that particular state half the households had incomes greater than $35,200 and half the households had incomes less than $35,200). This variable appears as "medinc" (for median income).

Before discussing techniques for testing our hypotheses, let's review some concepts from the early reading assignments that our important to understanding the data we will be working with (and will be covered on the final examination).

<u>First, what type of research design is used to collect the data</u>? Reviewing pages 5-7, we recall that in an experimental research design the researcher can set the levels of the independent variables. In our current situation that would require the research to be able to adjust how liberal each senator is, change their party affiliation and change the amount of income the median household earn in each of the 50 states. Obviously, I can't do any of these things (e.g., inject a senator with a serum to increase their conservatism). Therefore, as is typically the case in political science, economics and sociology, we are using a nonexperimental research design. In some instances (e.g., media studies) the researcher can set the level of at least some of the independent variables (e.g., determine the order in which the viewer see a series of news reports and/or how a particular event is described), but this is relatively rare in political science.

<u>Second, what is the "unit of analysis</u>"? From page 13 we know that the "unit of analysis" is what we collect data on. It is <u>not</u> a variable. For example, the unit of analysis is *not* tax, scons, party or medinc. Since we are collecting data on individual senators, a senator is the unit of analysis. *Not*, the U.S. Senate as a whole, but rather an individual senator. Suppose our research question were: Why did the Senate pass, or not pass, a particular bill? In this case the unit of analysis would be the U.S. Senate because we would be examining the behavior of the Senate collectively (e.g., Why were there more "yes" votes than "no" votes?). However, that is *not* what we are doing in the current analysis. We are *not* trying to explain why a particular tax bill was either passed or defeated. Rather, we are trying to explain why *individual* senators voted as they did on tax legislation (regardless of whether the legislation passed or was defeated). Therefore, we collect data on individual senators.

Since one of the variables, state median household income, is collected on states I could understand you thinking that a state is the unit of analysis. However, this wouldn't be correct. Each senator represents one state. We are interested in the median income of the senator's constituents, which happens to

be a state. If senators did not represent states then we would need an income measure of whatever their constituency was. Thus, it's coincidental that a senator's constituency consists of a state. Our interest in senators, not states. Therefore, a senator is the unit of analysis.

Third, what is the level of measurement of each of the variables? Two of the four variables (tax and scons) are percentages. From pages 11-12 we know that a percentage is a ratio level measure because it meets all the criteria. The scores on a percentage variable form a continuum from highest to lowest (i.e., 58% indicates more of the trait being measured than does 57%). Additionally, there is an equal interval between the scores (i.e., the difference between 58% and 57% is the same as the difference between 93% and 92%). Furthermore, a score of 0% indicates the absence of the trait being measure (i.e., a score of 0% on tax indicates that the senator never voted in favor of the interest of those with incomes at, or below, the median income). Therefore, a percentage variable meets all the requirements of the highest level of measurement, the ratio level.

Since state median family income meets all of these same criteria (e.g., $12,000 is greater than $11,000, the difference between $12,000 and $11,000 is the same as the difference between $11,000 and $10,000 and a score of "0" would indicate that the median – or middle – household had no income) it is also a ratio level measure.

The last variable, party affiliation, is more difficult to classify. Since there are only two categories of responses (i.e., 0 and 1) it is difficult to determine whether is an equal interval between all categories. By definition there is if you only have two categories. But, it's not all that reassuring. Additionally, there may, or may not, be a continuum. We could make a good case that if Republican senators are scored as "0" and Democratic senators scored as "1" then the continuum is in terms of increasing liberalism (i.e., going from 0 to 1 means a move in a liberal direction). Perhaps. It would be an easier assessment if there were more possible categories of responses.

As a "first look" at our data let's look at some of the descriptive statistics we examined earlier in the semester. The statistical package we will use, Stata, is probably the most commonly used statistical package political scientists. Previously (pages 17-24) we discussed measures of central tendency (e.g., the mean, median and mode) and dispersion (e.g., range and standard deviation). The Stata output immediately below shows these statistics for our data.

Variable	Obs	Mean	Std. Dev.	Min	Max
tax	100	46.54	28.73193	7	97
scons	100	35.11	31.24258	0	100
party	100	.62	.4878317	0	1
medinc	100	9.205	1.524174	6.1	12.4

All the entries in the "Obs" column are 100 because we have data on all 100 senators for each variable. The mean score for tax, 46.5 means that the average senator supported tax changes primarily benefiting households with incomes at, or below, the median household 46.5% of the time. This also means that the average senator supported the economic interest of households with incomes above the median 53.5% of the time (100 - 46.5 = 53.5). Immediately to the right of the mean on the tax variable, 46.5 is the standard deviation of 28.7.

Previously, I mentioned that a very useful way to interpret the standard deviation is in relation to the mean. Additionally, as mentioned in that same discussion, if the standard deviation is at least 50% of the size of the mean (which it is in this instance because 2 x 28.7 = 57.4 which is greater than the mean of 46.5 – i.e., the standard deviation must be at least half as large as the mean) then the mean was obtained by scores quite different than the mean. For example, if we had only two senators and one scored 52 while the other scored 48 they would have a mean of 50 but both scores would be quite similar to the mean. If so, there is little dispersion of scores (i.e., the scores tend to be concentrated close to the mean). However, a mean of 50 for two senators could also result for one senator scoring 100 and the other senator scoring 0. This is why we like to know the standard deviation: it gives us a good idea how concentrated, or dispersed, the scores are. Concerning the variable tax, since the standard deviation is over 50% of the size of the mean it tells us that scores very different from 46.5 averaged to 46.5 (e.g., 90, 72, 25, 10, etc.) rather than the mean of 46.5 occurring because the bulk of the scores were close to the mean (e.g., 52, 44, 49, etc.). This tells us a lot! It means that there is much disagreement on one of the most important policy areas: taxation.

To the right of the standard deviation you see "min" (the minimum or lowest score) and "max" (the maximum or highest score). Previously, we learned that the difference between the highest and lowest scores is called "the range." So, the range for tax is 86 (93 - 7 = 86). Given that the total possible range of the scale is 100 (i.e., from 0 to 100), this is a pretty high range. The fact that the standard deviation was well over 50% of the size of the mean meant that we were likely to have a high range. Since the range is equal to 86% of the possible difference in the scores (i.e., the scores range from 0 to 100 and 86 is 86% of 100), the range is large relative to the width of the scale.

Before leaving this discussion of descriptive statistics, let me mention how the mean on tax was calculated. Each senator had the opportunity to vote 76 times on the particular tax legislation I examined. Each senator's tax score was the percentage of times they voted in the interest of those households at, or below, the median income. Realistically, some votes shifted more money between income groups than other votes. Unfortunately, the Congressional Budget Office did not provide estimates about how much money each vote entailed. Therefore, all votes were treated as equal.

Here is something important to keep in mind about how the mean for state median household income (medinc) is calculated: all states are treated the same. Thus, the computer added up the median household income in each state and divided this total by 50 (i.e., the number of states). This means that each state had an equal impact. Since the smaller states tend to be poorer, the mean of the 50 states is likely lower than if the value of each state's median income was weighted by the size of the state's population (i.e., if California has 8 times the population of Ohio then California's median household income would "count" 8 times as much as Ohio's median household income). Looking at the previous page, you see that the mean on median income is 9.2. Since the data are in thousands of dollars this means that the average state had a median household income of $9,200. This figure is so low because the data are from the 1970 census. Currently, the median household income in the United States is around

$50,000. After adjusting for inflation, the median household income is higher today than in 1970, but not nearly as much as the difference between $9,200 and $50,000 suggests.

As mentioned previously, once we have selected the dependent variable (in this case the percentage of times each senator votes in favor of tax changes where over 50% of the benefits go to households earning at, or below, the median income) and the independent variables (a senator's political philosophy, party affiliation and the median income in the state they represent) likely to explain variation (i.e., some senators support lower income households on 90% of their votes on tax changes while others support lower income households on only 20% of their votes on tax changes – hence "variation"), we need a statistical procedure that will tell us the amount of impact that each independent variable has on the dependent variable.

Earlier in the semester (pages 30-40 of this reader) you read about cross tabulation and measures of association. Let's apply those approaches to trying to answer the fundamental question: How much impact does each of the independent variables have on the dependent variable? I will first try to answer this question using cross tabulation. Read the second paragraph on page 37 (i.e., "First, cross tabulation …") *before* continuing.

In a cross tabulation table a "cell" represents one possible combination of scores on the variables used in the analysis. In our situation, a cell might represent a senator who scored 83 on "scons" (i.e., is rather conservative), who was a Republican (i.e., score "0" on "party"), represented a state whose median household income was $41,511 and who supported households making at, or below, the median household income 17% of the times they voted on tax legislation. If *any* of those scores change then we need another cell. For example, since it is possible for a senator to score 83 on scons, be a Republican (i.e., score "0" on party), represent a state whose median household income is $41,511 (i.e., all three scores identical to the first senator) but support tax changes primarily benefiting households with incomes equal to, or less than, the median income 16% of the time (instead of 17% as with the first senator) we need another cell to represent this second possible combination of scores.

We need an additional cell for each possible combination of scores *even if* there are no senators who have this combination of scores. For example, since a score of "0" on scons would mean the senator had no conservatism (i.e., was a liberal as the scoring mechanism would allow) no Republican senator is likely to score "0" on scons. However, a cross tabulation table would still need to construct a cell to represent a possible set of scores for such a senator (e.g., a Republican senator who scored "0" on scons, represented a state with a median household income of $35,917 and voted in favor of households with incomes equal to, or less than, the median income 12% of the time). The total number of cells in a cross tabulation table is equal to the product (multiplication) of the number of categories of responses of the variables. Since state median household income can take on a virtually infinite number of values, I can't calculate how many cells our analysis would take. However, if we only use the variables "scons" (101 possible scores – 0 plus 1 through 100), "party" (2 possible scores – 0 and 1) and "tax" (101 possible scores – 0 plus 1 through 100) we would need a cross tabulation table containing 20,402 cells (101 x 2 x 101 =

20,402). Needless to say this would be a monstrosity! Our statistical package, Stata, won't produce such a table.

Even if Stata would generate such a table, think of what this would mean for significance testing. Read the second paragraph on page 38 (i.e., "Second, even if …") *before* continuing. Suppose we have only one senator who is a Republican, has a conservatism score of 93%, whose state has a median household income of $11,411 (in 1970) and voted in favor of tax changes primarily benefiting households with incomes at, or below, the median 11% of the time. Applying the second paragraph on page 38 to our situation, this would be analogous to flipping a coin one time and trying to determine if the coin is biased. If the coin comes up heads, what should we conclude? Unless the coin landed on it's side, it would have to have come up either heads or tails. One flip is simply too little evidence to draw any firm conclusion as to whether, or not, the coin is biased in any particular direction (i.e., toward heads or tails). In our current situation, we need a statistical technique which preserves the sample size (i.e., is based on all 100 senators) rather than an approach which sub-divides the sample into so many pieces that there are very few senators in each possible outcome (i.e., one particular set of scores on all the variables).

A proponent of cross tabulation could reply that a remedy to this problem is to reduce the number of cells. For example, if we converted tax and scons from percentages (i.e., from a range of 0 to 100) into three categories each (i.e., 0%-33% = 1, 34%-66% = 2 and 67%-100% = 3) we would have reduced the number of cells dramatically. True. However, conversion comes at a high price: we lose all the information within each of the new categories. For example, under this revised scoring system, a senator whose tax score was 7% (i.e., only voted 7% of the time in the interest of those with incomes at, or below, the median) would receive the same score (1) as would a senator whose tax score was 33%. These senators behaved very differently but received the same score. That's not good measurement. We should be trying to use all the information we have rather than omitting information we already possess. Additionally, even if we used this revised scoring system for the percentage variables and converted state median household income into 3 categories (e.g., high, medium and low state median household incomes) we would have three variables with three categories each (tax, scons and state median household income) and one variable with two categories (party) resulting in a cross tabulation table with 54 cells (3 x 3 x 3 x 2 = 54). The table would still be much too large and wouldn't tell us the precise magnitude of the relationships between the variables (i.e., how much does each additional percentage point in a senator's degree of conservatism impact their degree of support for tax changes primarily benefiting households with incomes at, or below, the median).

The number of cells, 54, would be high enough to likely cause the user of cross tabulation to reduce the number of independent variables. For example, if you cross tabulated the revised scons and tax variables you would have only 9 cells (3 x 3 = 9). However, as you will see shortly, this is a terrible solution. The impact of one independent variable on the dependent variable often changes when *other* independent variables are included. The impact of a senator's conservatism on their support for tax changes primarily benefiting low and moderate income households would likely change if we included the senator's

political party affiliation and their state's median household income in the analysis. Using conservatism as the *sole* independent variable precludes such a possibility and means that we are likely to obtain a less valid measure of the impact of a senator's conservatism on their votes on tax legislation than we would if we had the other two independent variables (party and state median household income) in the analysis.

Up to the mid-to-late 1970s political scientists often used simple methods such as cross tabulation as the primary method of statistical analysis. Frequently a cross tabulation table was supplemented with a measure of association. It would be a good idea to reread pages 34-37 *before* continuing. Of the measures of association discussed previously (pp. 34-37), correlation would be the best choice to use because it requires interval or ratio level data (i.e., unlike gamma or Kendall's tau the calculation procedure for correlation makes use of the fact that there is an equal mathematical interval between adjacent categories – e.g., that the difference between 19% and 20% is the same as between 72% and 73% - that both interval and ratio level data possess). The correlation between scons and tax is -.80. Using the interpretation table on page 37, a -.80 correlation represents a very strong negative association between a senator's conservatism and their degree of support for tax changes primarily benefiting low and middle income households (i.e., the more conservative the senator the lower their support for tax changes primarily benefiting low and middle income citizens). A correlation of + or - .80 is very strong. Since the strongest possible correlation is +1.0 or -1.0, -.80 is 80% as strong as the correlation could have been. Keep in mind that a correlation of -.80 is of identical strength with a correlation of .80. Only the direction (positive or negative) is different.

Unfortunately, the -.80 correlation between a senator's conservatism and their support for tax changes primarily benefiting middle and low income households is a bivariate correlation (i.e., between these two variables alone). As previously discussed, the impact of one independent variable on the dependent variable is likely to change if we have additional independent variables in the analysis. In order to examine this, I asked Stata to provide a partial correlation between conservatism and support for tax changes primarily benefiting middle and low income households. A partial correlation removes the impact of the other independent variables. Put another way, the partial correlation between conservatism and support for middle and low income households on tax legislation is answering this question: if two senators had the same party affiliation (e.g., both Republicans) and the median household income in their states were the same (e.g., both $11,400 in 1970), what is the relationship between their conservatism scores and their support for tax changes primarily benefiting middle and low income households? The partial correlation between a senator's conservatism and their support for tax changes primarily benefiting middle and low income households is -.65. Notice while the direction of the relationship remains "negative," the relationship is not quite as strong as when we did not use the senator's party affiliation and state median family income (-.65 vs. -.80). Now you see why we want to use all the independent variables that theory suggests we should use: the results change. The partial correlation represents a more valid assessment of the relationship between a senator's

conservatism (i.e., scons) and their support for tax changes primarily benefiting middle and low income households (i.e., tax) than either the cross tabulation analysis or the correlation analysis. However, as the diagram on page 40 makes clear, the partial correlation does *not* provide us with an assessment of the *magnitude* of the relationship between a senator's conservatism and their support for the tax changes primarily benefiting middle and low income households. What we really want to know is the answer to a question such as the following: after removing the impact of all other independent variables theory suggests, each one percentage point increase in a senator's conservatism is associated with what percentage point decrease in their support for tax changes primarily benefiting middle and low income households?

As the figure on page 40 vividly shows, the *strength* of the association between two variables does *not* tell us the *magnitude* of the association between two variables. Think back to the example on page 40. A person's education and income could be very strong correlated but the impact of each additional year of education on someone's income could be rather small. As the diagram on page 40 clearly shows, a correlation of .70 between a person's education and their income would, since it is positive (i.e., .70 and not -.70), tell us that the more highly educated a person became the higher their income. However, does each additional year of education, on average, increase someone's annual income by $1,000, $7,500, $20,000 or some other amount? The correlation between a person's level of education and their income cannot tell us the answer because it only tells us the strength and not the magnitude of the association.

In order to obtain the magnitude of an association between variables we need to use regression. Since theory almost invariably suggests that a dependent variable is influenced by more than one independent variable, the magnitude of the association between a particular independent variable and the dependent variable is likely to change if the other independent variables are included in the statistical analysis, the bulk of contemporary quantitative research in political science uses multiple regression (i.e., regression with more than one independent variable) or a similar technique (e.g., probit, logit, cox regression, etc).

The most straightforward approach to discussing multiple regression is to display the multiple regression printout for our model and interpret it. Remember that our model seeks to explain why senators vary in the percentage of times the vote in favor of the interest of households with incomes equal to, or less than, the median household income (i.e., "tax"). From the discussion on pages 60-61, three independent variables that might logically influence how frequently a senator would vote for tax changes primarily benefiting households with incomes equal to, or less than, the median income are: (1) the political philosophy of the senator (scons); (2) the party affiliation of the senator (party); and (3) the median income of the state the senator represents (medinc). The Stata multiple regression output for out model appears immediately ahead.

```
      Source |       SS       df       MS              Number of obs =     100
-------------+------------------------------           F(  3,    96) =   65.44
       Model |  54886.5757     3  18295.5252           Prob > F      =  0.0000
    Residual |  26840.2643    96  279.586087           R-squared     =  0.6716
-------------+------------------------------           Adj R-squared =  0.6613
       Total |    81726.84    99  825.523636           Root MSE      =  16.721

------------------------------------------------------------------------------
         tax |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       scons |  -.6447205   .0756005    -8.53   0.000    -.7947863   -.4946547
       party |   11.20792   4.675335     2.40   0.018     1.927454    20.48839
      medinc |  -.5600809   1.283164    -0.44   0.663     -3.10714    1.986979
       _cons |   67.38277   15.11393     4.46   0.000     37.38186    97.38368
```

Probably the most important numbers in the above printout appear in the "Coef." column. "Coef." Is an abbreviation for coefficient. The coefficients tell the magnitude of the impact of each independent variable on the dependent variable after the impact of each other independent variable has been accounted for. <u>For example, the -.644 coefficient for "scons" is interpreted as follows: If a senator' party affiliation and the median household income in the senator's remain constant (e.g., the senator is a Democrat and remains a Democrat, the median household income is $41,357 and remains at $41,357), for each one percentage point *increase* in the senator's conservatism score (e.g., the senator's conservatism score increases from 57% to 58% - a 1% increase) the percentage of times the senator votes in favor of tax changes where 50%, or more, of the benefits go to households with incomes equal to, or less than, the median household income *decreases* (notice that the coefficient is -.644, not .644 – thus a negative association), on average, by approximately .6% (i.e., 6 tenths of 1%, not 6.4% or 64%)</u>. When interpreting coefficients, think of what is realistically possible. If you interpreted the -.644 coefficient as meaning that a 1% increase in conservatism is associated with a 64% decrease in support for egalitarian tax changes, this would mean that a 2% increase in conservatism (e.g., a senator increasing their conservatism from 59% to 61%) would result in a 128% decrease in support for egalitarian tax changes (2 x 64 = 128). If true (which it isn't) this small 2% point increase in conservatism would result in a decrease in support for tax changes primarily benefiting middle and low income households by more than the entire length of the scale (i.e., the scale is from 0 to 100 so a decrease of 128 would be greater than the entire span of the scale). Such a small increase in conservatism producing such a correspondingly large decrease in support for tax changes primarily benefiting households with incomes equal to, or less than, the median isn't plausible. This is one reason you always want to pay close attention to how the variables are measured.

Nevertheless, the -.644 impact is large. For example, if both party affiliation and state median household income of a senator remain the same, replacing a senator whose conservatism score is 10% (i.e., is rather liberal) with a senator whose conservatism score is 90% would *decrease* support for tax changes primarily benefiting households at, or below, the median income by approximately 55.5% percentage points (90 – 10 = 80 and 80 x -.644 = - 51.5). Political philosophy matters!

<u>The coefficient for party affiliation is interpreted as follows</u>: If a senator' conservatism and the median household income in the senator's remain constant (e.g., the senator has a conservatism score of 72% it remains at 72%, the

median household income is $41,357 and remains at $41,357), replacing a Republican senator with a Democratic senator (i.e., going from a score of "0" on party to a score of "1" – remember that party affiliation is not measured in percentage terms – thus a "unit" of increase in political party affiliation is not a percentage point), the percentage of times the senator votes in favor of tax changes where 50%, or more, of the benefits go to households with incomes equal to, or less than, the median household income *increases* (notice that the coefficient is 11.207, not -11.207 – thus a positive association), on average, by approximately 11%. Since party and conservatism are fairly strongly correlated (-.57 - thus Democratic senators are less conservative than Republican senators) in estimating the impact of a senate election in California the impact of party should be added to ideology.

Over the last 40 years, or longer, Democratic senate candidates are roughly 80%, less conservative than Republican candidates. Therefore, the replacement of a Democratic senator with a Republican senator in California would likely reduce support for tax changes primarily benefiting households with incomes equal to, or less than, the median household income by approximately 66% (the 80 points more conservative the Republican candidate is than their Democratic opponent would lower support by roughly 55% and the separate impact of being a Republican further reduces support by an additional 11%| 55% + 11% = 66%). Elections matter!

<u>The coefficient for state median household income is interpreted as follows</u>: If a senator's conservatism and party affiliation remain constant (i.e., the same - the senator has a conservatism score of 72% it remains at 72% and the senator is, and remains, a Democrat) for each $1,000 increase in median household income in the senator's state (remember from page 60 that state median household income is measured in *thousands* of dollars, not dollars – thus a one unit increase in state median household income – i.e., instead of 34.2 the computer reads a score of 35.2 - is an increase of $1,000, not $1– always pay close attention to the units of measure), the percentage of times the senator votes in favor of tax changes where 50%, or more, of the benefits go to households with incomes equal to, or less than, the median household income *decreases* (notice that the coefficient is -.560, not .560 – thus a negative association), on average, by approximately .5% (i.e., 5 tenths of 1% or one-half of 1%, not 5.6% or 56%).

As with the previous results, the finding for state median household income is important. While the negative coefficient is what we expected (i.e., the wealthier the state the less supportive a senator from that state is of tax changes primarily benefiting those with incomes equal to, or less than, the median household income), the magnitude of the impact across the United States is small. On the bottom of page 62 notice that the highest state median household income is $12,400 (12.4 in the "max" column on page 62) and the lowest score is 6.1 (see the "min" column on page 62). This is a difference of $6,300 ($12,400 - $6,100 = $6,300). However, holding both a senator's political philosophy and party affiliation constant, going from representing the wealthiest state in the union (i.e., the state with a median household income of $12,400 in 1970 dollars) to a state with a median household income of only $6,100 result in only a 3.5% reduction in support for tax changes primarily benefiting households with

incomes equal to, or below, the median. For example, transplanting a senator from the wealthiest state in the United States in 1970 (e.g., California or Alaska) to the poorest state (Mississippi) would only result in a decrease in support for tax changes primarily benefiting middle and low income groups by approximately 3.5%. Conversely, the replacement of Democratic senator from California with a Republican senator from California would reduce support by approximately 66% (see previous calculation), or about 19 times as much (66/3.5 = 18.85). Thus, political philosophy and party affiliation are much more important than state economic self-interest. That tells you a lot about how our political system operates.

The data in this study are actual, not hypothetical, data. This was the United States Senate in action. The questions the methods in this reading assignment are used to answer are "big" questions. When we examine the literature in international relations, comparative politics and public law, we will use these same statistical techniques to help answer "big" questions in those sub-fields of political science.

The final number in the coefficient column on page 68 is 67.382. The is the coefficient for what is termed the "y intercept" (cons stands for "constant"). To keep the discussion short, let me mention that the value of the y intercept is the predicted value for the dependent variable if all the independent variables have a score of 0. With our dataset this means the following: if a senator has a score of 0 on conservatism (i.e., has no conservatism – thus, as liberal as a senator could be on our scale), is a Republican (i.e., scored "0" on party) and the median household income in the senator's state is $0, *then* our results predict that this senator would vote in favor of tax changes primarily benefiting households with incomes equal to, or below, the median approximately 68% of the time. As is often the case, the value of the y intercept is *not* of paramount importance to us because it frequently depicts a state of the world which is highly unlikely to occur. For example, for a senator to score 0 on both conservatism and party affiliation means that a senator who has no conservatism (i.e., is as liberal as possible) would be a Republican. Why would such a liberal senator be affiliated with the Republican party? The short answer is, they wouldn't! Additionally, consider what a 0 score on state median household income means. The only way a state could have a median household income of $0 would be if either half of the state's households literally earned $0 or some households had negative incomes (Were they were paying their employers for the privilege of working for them?). Not only are any of these conditions extremely unlikely to occur, for the y intercept to be an accurate depiction of the state of the nation, they would all have to occur simultaneously!

As is frequently the case, the value of the y intercept is to use it in conjunction with "real world" scores on the independent variables to predict scores on the dependent variable. Let's try an example from our current study. Senator #1, former Democratic Senator Howell Heflin of Alabama, had the following scores on the variables we are using: scons 26; party 1; state median household income 7.4 and tax 54. Thus, Helfin was a fairly liberal senator (a conservatism score of only 26%), a Democrat (i.e., "1" on party), represented a very poor state (median household income of $7,400 in 1970) and voted 54% of the time in favor of tax changes primarily benefiting households with incomes

equal to, or less than, the median. From the results on page 68 we know that the coefficient values for the independent variables are as follows: cons -.644; party 11.207 and state median household income -.560. To predict Helflin's support for middle and low income groups on tax legislation we multiply each coefficient times Heflin's score on that particular independent variable and add the y intercept (67.4). Therefore, the calculation is:

67.4 + (-.644) (26) + (11.207) (1) + (-.560) (7.4)

which becomes: 67.4 + (-16.744) + 11.207 + (-4.144)

which becomes: 67.4 - 16.744 + 11.207 - 4.144

which = 57.7

Thus, Heflin is predicted to support tax changes primarily benefiting households with incomes equal to, or less than, the median household income 57.7% of the time. Since Heflin actually voted in this direction 54% of the time (i.e., his score on "tax" is 54) our model's prediction is "off" by 3.7% (57.7 – 54 = 3.7). That's very, very good! There are other senators in this study (who will remain nameless) for whom the predictions of our model were "off" by as much as 40%. The primary use of the y intercept is to perform this type of calculation.

<u>The prediction calculation immediately above is of tremendous importance in explaining more fully how the regression model works</u>. We now know that the prediction for Senator Heflin was "off" by 3.7%. In statistical terminology the value of the error term, referred to as "e," for observation #1 (in our case, Senator Heflin) is 3.7. Since the formula for the value of "e" for a given observation is the actual score minus the predicted score, the value of "e" for Senator Heflin is – 3.7 (i.e., 54 – 57.7 = - 3.7). What the computer now does is to "square" this error value. In symbols this would be e^2. For Senator Heflin the value of e^2 is 13.69 (-3.7 x -3.7 = 13.69). The computer now performs the same operation on the remaining 99 senators. Thus, for each senator the computer generates a prediction, subtracts this prediction from the senator's actual score on tax (i.e., the dependent variable) and squares the difference. After performing this operation on all 100 senators, the computer then adds up the total of these "squared prediction errors."

Now turn back to page 68 and look toward the top of the printout under the "SS" column. If you look down the "SS" column and to the right of "Residual" (named for "error" – residue, hence residual – i.e., not accounted for by our independent variables) you should see the value 26840.2643. Omitting the decimal and adding a comma, this value is 26,840. What this value represents is the sum of the 100 (since we have 100 senators) squared error scores. If you divided 26,840 by 100 the result is 268.4. This means that for the average senator, the value of their "squared error" is 268.4. Obviously, Heflin's squared error score of 13.69 is much smaller than for the average senator. Put another way, our model's prediction was much less inaccurate for Senator Heflin than for the typical senator.

The values in the coefficient column (i.e., -.644 for scons, 11.207 for party, -.560 for state median household income and 67.382 for the y intercept) were chosen by the computer to *minimize* (i.e., obtain the lowest) total squared errors. Put another way, if any of the values in the coefficient column were changed, the total of the squared prediction errors would be *greater* than 26,840. By basing the calculation of each of the coefficients on *squared* errors, as opposed to the absolute size of the error, the assumption is that large errors are more "costly" than smaller errors.

For example, let's say that the coefficients used by model 1 yield 4 prediction errors of 1 each. This would result in a total prediction error of 4. Since 1^2 is also equal to 1, the total squared errors for model 1 is also 4. Contrast this with the outcomes from model 2 using the same scores on the variables but different values for the coefficients: 4 predictions in which model 2 predicted the exact score on the dependent variable 3 times (e.g., the error for predictions 1-3 was 0 because the model predicted the actual score on the dependent variable) but on the 4^{th} prediction model 2 made a prediction error of 4. For model 2 the total squared prediction errors is 16 [i.e., 0 + 0 + 0 + (4)(4) = 0 + 0 + 0 + 16 = 16]. The *total* prediction error is the same for both models (i.e., 4 since 1 + 1 + 1+ 1 = 0 + 0 + 0 + 4 = 4). However, the *total squared* prediction errors are 4 times as great for model 2 (16) as for model 1 (4). According to the least (or lowest) total squared error principle, the one large error of 4, even when accompanied by 3 perfect predictions, is *less* desirable than no perfect predictions but 4 errors of 1 each. Thus, a few large errors are less desirable than a larger series of smaller errors. Therefore, the computer would report the coefficients from model 1, not model 2.

Hopefully, you remember the discussion of statistical inference from pages 41-59 of this reader. Remember the fundamental question of statistical inference: How likely are the results to be the product of chance? Applied to the results in the coefficient column on page 68 this question could be translated into: How likely are we to obtain a coefficient value for conservatism as large as -.644 when the "true" impact of the coefficient for conservatism is .000? If the true value of the coefficient for conservatism is .000 this would mean that an increase in a senator's conservatism would have *no* effect on their voting on the percentage of times the senator votes in favor of tax changes mostly benefiting household with incomes equal to, or less than, the median household income. Needless to say, this is an important question to answer.

As you read previously, if we reject the null hypothesis that a senator's degree of conservatism has no impact on their support for tax changes primarily benefiting middle and low income households in favor of the alternative hypothesis that the more conservative the senator the less they will support tax changes primarily benefiting middle and low income groups (which is what the -.644 coefficient for senator conservatism indicates) when, in fact, a senator's degree of conservatism has no impact on their voting on this legislation we commit a type I error (i.e., rejecting the null hypothesis when the null hypothesis is true). Obviously, we would like to avoid such a mistake.

The results on page 68 tell us the probability that if we reject the null hypothesis the null hypothesis is actually true. Remember from pages 26 & 28

that if we have a normal distribution of scores, 95% of the scores will be within plus, or minus, 2 standard deviations of the mean. Thus, if a group of scores are normally distributed, the mean of the scores is 50 and the standard deviation is 5, 95% of the scores will be between 40 and 60 (50 – 5 – 5 = 40 and 50 + 5 + 5 = 60). Fortunately, by virtue of what is called the central limit theorem, if we could replicate (i.e., repeat the study with a different 100 senators) many times, the estimates of each coefficient would approximate a normal distribution. If you turn to page 68 and look at the column to the right of the coefficient column (i.e., to the right of "Coef.") you should see "Std. Err." This column contains what are called "standard errors."

The relationship of the coefficient to the standard to which it is attached (e.g., the coefficient for scons of -.644 is "attached" to the standard error of .0756) is the same as the relationship of the mean to the standard deviation in a normal distribution: if we replicate the study 95% of the estimates of the coefficient will be within 2 standard errors of the reported coefficient. Therefore, 95% of the estimates of the coefficient for conservatism (-.644 is the only estimate we have) will be between -.794 (-.644 + .075 + .075 = -.794) and -.494 (-.644 - .075 - .075 = -.494). Since .000 does *not* lie between -.794 and -.494 there is less than a 5% chance that coefficient for conservatism would be reported as -.644 when it's "true" value is .000. What we just did was to calculate what is termed a 95% confidence interval (i.e., 95% of the time the "true" value of the coefficient lies within this interval). Turning to the regression results on page 68, notice that the entries in the furthest two columns on the right for "scons" show exactly this 95% confidence interval (-.794 to -.494) that we just calculated.

***Reread paragraphs 2-3 on page 50 very carefully*. Given the results just presented we can say that since the null hypothesis (i.e., that the "true" value of the coefficient for a senator's conservatism is .000) is true less than 5% of the time we will reject the null hypothesis and accept the alternative hypothesis that the more conservative the senator the less supportive they will be of tax changes primarily benefiting households with incomes equal to, or less than the median household income.**

There is a much easier approach to determining whether, or not, a regression coefficient is statistically significant at .05 level (if we reject the null hypotheses there is a 5%, or less, chance that we are wrong – i.e., thus, a 5%, or less chance that if we reject the null hypothesis we commit a type I error – see paragraphs 2-3 on page 50). If the absolute value (i.e., positive or negative) of what is referred to as the "t ratio"- which is the regression coefficient divided by it's own standard error - has an absolute value of 2.0, or greater, the coefficient is statistically significant at the .05 level (just keep reading). Let's apply the t ratio calculation to senator conservatism. From the immediately preceding paragraph we know that the coefficient for senator conservatism is -.644 and the standard error for the coefficient for senator conservatism is .075. This result means that the t ratio for senator conservatism is -8.53 (-.644/.075 = -8.53 – differences due to rounding). Since -8.53 has an absolute value greater than 2.0, we know that there is less than a 5% that the null hypothesis is true. So, we will reject it. If you turn back to the results on page 68 you should notice that to the immediate right of the "Std. Err." column is the "t" column (for t ratio). Notice further that the entry for "scons" in the "t" column is -8.53.

Returning to the results on page 68, notice that to the right of the t column is a column entitled "P>|t|." The "P" in the column title stands for "probability." The > sign means "greater than." When combine with "P" the left side of the expression mean "Probability greater than." The expression |t| means the absolute (i.e., irrespective of positive or negative sign) value of the t ratio. In the regression results for senator conservatism (scons) the entry in "P>|t|" column is 0.000. This means that since the t ratio has an absolute value of 8.53, there is less than a 1 in 1,000 (theoretically 0 – which is less than 1) chance that the null hypothesis is true. Alternatively, you could say that given a coefficient value of -.644 and a standard error of .075, if we reject the null hypothesis (which we will) there is less than a 1 in 1,000 chance we will commit a type l error (rejecting the null hypothesis when the null hypothesis is true).

Looking at the t ratios for the senator's political party affiliation (party) and the median household income in the senator's state (medinc), notice that party is statistically significant at the .05 level (because 2.4 is greater than 2.0) while state median household income is not statistically significant at the .05 level (because -.44 has an absolute value less than 2.0).

Frequently political scientists want to know how well their model has performed. In our case, this means how well do the senator's conservatism, party affiliation and the median household income in the senator's state explain the percentage of times the senator votes in favor of tax changes primarily benefiting households with incomes equal to, or less than, the median household income? The results for the statistic R-squared will tell us the answer. From the regression results on page 68, notice that the value of R-squared (often referred to as "R^2") is .6716. The number is interpreted as follows: variation in the senator's conservatism (i.e., all senators are not equally conservative), party affiliation and the median household income in the senator's state explain 67% of the variation in the percentage of times a senator's votes in favor of tax changes primarily benefiting households with incomes equal to, or below, the median household income.

To be judgmental, 67% is a fairly high percentage of the variation in the dependent variable to explain. Thus, our model "works" pretty well. Since we could explain 100% (or all) the variation in a senator's support for tax changes primarily benefiting households with incomes equal to, or less than, the median household income, the R-squared of .67 means that 33% of the variation remains unexplained by the three independent variables we have. Perhaps there are independent variables that theory suggests that have not been included in our model. Alternatively, perhaps we have the correct independent variables but measurement error is reducing the percentage of the variation our model explains.

Fortunately, there is a rather intuitive method of understanding the logic behind R-squared. If we have no independent variables and have to predict the percentage of times each senator will vote in favor of tax changes primarily benefiting households with incomes equal to, or less than, the median household income our best solution is to predict that each senator will score the mean value (in our case, 46.5 – see page 62, the mean value of "tax"). In terms of how well our model performs, the question now becomes: does the inclusion of the three independent variables we have (senator conservatism, party affiliation and state

median household income) produce more accurate predictions (or reduce the amount of our prediction errors) than we obtain by predicting the mean score for each senator? Recall that Senator Heflin voted in favor of tax changes primarily benefiting middle and low income households 54% of the time. Since the mean score for "tax" is 46.5, if it had not been for our knowledge of Senator Heflin's scores on the independent variables (i.e., that he has a conservatism score of 26, is a Democrat and that the median household income in his state was $7,400 in 1970) we would have predicted his score to be the mean value of 46.5.

However, based on Senator Heflin's scores on the independent variables, we changed this prediction to 57.7 (see the computations on page 71). If we predicted the mean score of 46.5 for Senator Helflin the difference between his actual score, 54, and our prediction, 46.5, is 7.5 (we subtract the predicted score from the actual score: 54 - 46.5 = 7.5). However, by incorporating the knowledge of Senator Heflin's scores on the independent variables, our new prediction, 57.7 is closer to his actual score of 54 than the mean value (54 – 57.7 = -3.7 which has a smaller absolute value than 7.5). Thus, knowledge of our three independent variables increased the accuracy of our prediction.

The "variance explained" interpretation of R-squared that I mentioned on the previous page is the standard interpretation of R-squared [i.e., variation in the senator's conservatism (i.e., all senators are not equally conservative), party affiliation and the median household income in the senator's state explain 67% of the variation in the percentage of times a senator votes in favor of tax changes primarily benefiting households with incomes equal to, or below, the median household income.]. Alternatively, we could say that the variation in senators' conservatism, party affiliation and state median household income reduce the squared prediction errors (i.e., multiplying each prediction error times itself) 67% from what they would have been by predicting the mean score on the percentage of times a senator supports tax changes primarily benefiting households with incomes equal to, or below, the median.

Political scientists are often interested in the relative importance of the independent variables. For example, we might ask: How important is a senator's philosophy relative to their political party affiliation in explaining how often the senator supports tax changes primarily benefiting households with incomes equal to, or less than, the median household income? If we look at the coefficient values on page 68, it is tempting to think that since the coefficient for party affiliation is 11.207 and the coefficient for conservatism is -.644 that party affiliation is approximately 17 times as important in explaining senator's support for middle and low income households on tax legislation as the senator's conservatism (11.207/.644 = 17.4). As we will soon discover, this would be a serious mistake. The reason we cannot directly compare the coefficient values is that the variables the coefficients are aligned with, party affiliation and conservatism, are measured on very different scales with vastly different means and standard deviations.

We faced a similar situation over pages 27-29 when we compared the Miller Analogies Test with the Graduate Record Examination. The scores weren't directly comparable because the tests have very different means and standard deviations. Our solution, conversion to Z scores, was to subtract the mean score on the test from each student's individual score and divide the difference by the

standard deviation for that particular test. This process was used to find how well a particular student did relative to the average on the test.

If we apply the logic of the Z score to the regression coefficients, we can find out how important each independent variable is relative to the other independent variables. The process we use will produce what are called "standardized coefficients." Note the similarity to Z scores (which are often termed "standard scores"). The formula is to multiply each coefficient on page 68 (what are termed "unstandardized coefficients) by the ratio of the standard deviation of the independent variable the coefficient is associated with to the standard deviation of the dependent variable (just keep reading – it's simple).

From page 62 we know that for senator conservatism (scons) the mean is 35.11 and the standard deviation is 31.2. Additionally, from page 62 we also know that the standard deviation of senatorial support for tax changes primarily benefiting households with incomes equal to, or below, the median is 28.7. From page 68 we know that the coefficient for senator conservatism is -.644. Putting those number into the formula I mentioned above results in a standardized senator conservatism coefficient of -.695 [-.644/(31.2/28.7) = -.644/1.08 = -.695]. If we apply the same formula to the other two independent variables, then we can compare one standardized coefficient with another and obtain the relative importance of the independent variables. The array below provides some useful results:

Independent Variable	Unstandardized Coefficient	Standardized Coefficient
Senator Conservatism	-.644	-.695
Party Affiliation	11.207	.190
State Median Household Income	-.560	-.029

From the results in the standardized coefficient column we can see that senator conservatism is approximately 3.6 times as important in explaining senatorial support for tax changes primarily benefiting middle and low income households as is party affiliation (-.695/.190 = -3.65 – only absolute values matter here). In thinking about senatorial conservatism and party affiliation consider what a one unit increase in each variable means. For senatorial conservatism, a one unit increase is a one percentage point increase (e.g., from 38% to 39%). This is not much change. However, the only unit of party affiliation would be in completely changing parties (i.e., changing from a Republican to a Democrat – i.e., 0 to 1 or vice versa). A one unit change in party represents much more change than a one unit change in senatorial conservatism. As in the case of senator conservatism and party affiliation, standardized coefficients are often quite different than unstandardized coefficients. A "quick" method of roughly gauging the relative importance of the independent variables is to take a ratio of t ratios (just keep reading). For example, using the absolute value of the t ratios

for senator conservatism and party affiliation would predict that senatorial conservatism is approximately 3.6 times as important as party affiliation (from page 68: 8.53/2.4 = 3.55). That's pretty close to the actual figure of 3.65! Since political scientists are typically most interested in the impact of an independent variable on the dependent variable, as opposed to the relative importance of the independent variables, they typically use unstandardized coefficients.

Karl Marx would probably be disappointed with our results. Given an economic class warfare perspective, Marx would have thought that state median household income would have been the most important independent variable. It is, by far, the least important. Additionally, state median household income is the only independent variable that is statistically insignificant. Thus, we can't rule out the possibility that differences in state median household income have no effect on the percentage of times a senator votes in favor of tax changes primarily benefiting households with incomes equal to, or less than, the median household income.

If Marx was statistically oriented, he might raise the following point: the reason that state median household income is statistically insignificant is that the variation that it explains is also explained by the other independent variables. This is the multicollinearity problem that was previously discussed on page 7. To suggest a pictorial example, think of an eclipse. If Marx's variable, state median household income, is like the sun in an eclipse, it is being blocked from view (i.e., from having an impact on the dependent variable).

While it is difficult to correct for multicollinearity if we have it, it is easy to test for it. First, we only worry about multicollinearity for statistically *insignificant* independent variables. From the results on page 68, we know that the t ratios for both senator conservatism and political party affiliation have an absolute value greater than 2.0 (-8.53 for senator conservatism and 2.40 for party affiliation). Thus, we do not need to be concerned about multicollinearity for either a senator's conservatism or their party affiliation. In order to see if senator conservatism and/or party affiliation are preventing state median household income from having a statistically significant impact on senatorial voting on tax legislation, we need to run another regression in which state median household income is the *dependent* variable and the senator's conservatism and party affiliation are the independent variables. The R-squared from this equation will tell us what percentage of the variation in state median household income is explained by a senator's conservatism and their party affiliation. The R-squared from this equation is .26. Since only 26% of the variation in state median household income is explained by senatorial conservatism and party affiliation, Marx would not be justified in thinking that high multicollinearity is the likely reason state median household income is statistically insignificant in the results displayed on page 68.

Before abandoning Marx's theory, let me mention one other possibility: the impact of state median household income on the percentage of times a senator votes in favor of tax changes primarily benefiting middle and low income households is *indirect*: state median family income effects a senator's political philosophy (e.g., the poorer a state the more likely a senator from that state will adopt a liberal political philosophy and/or be a Democrat), and political philosophy and/or party, in turn, effect senatorial voting on tax legislation. In this

model, the impact of state median household income on senators' voting on tax legislation would be through median household income's impact on conservatism and party. This is what is termed a causal model. Since state median household income only explains about 10% of the variation in either a senator's conservatism or party affiliation, there probably isn't sufficient reason to convert to a causal model. I should also mention that state median household income is negatively associated with senator conservatism (i.e., opposite what Marx would have thought, the wealthier the state the less conservative the senators the state elects). So far, Marx doesn't appear "to have an out"!

One remaining possibility for Marx is what is called an interactive model. In this model the impact of one independent variable on the dependent variable is affected by the level of another independent variable (just keep reading). For example, it might be that the impact of state median household income on a senator's support for tax changes primarily benefiting middle and low income households depends upon the philosophy of the senator. Thus, perhaps more conservative senator's almost entirely respond to the wealthy, regardless of how high or low the state's median household income is where more liberal senators representing states with low median household incomes are very supportive of tax changes primarily benefiting middle and low income residents whereas more liberal senators from high median income states are much less supportive of tax changes primarily benefiting middle and low income residents. In order to test this model we need to create an interaction term by multiplying a senator's conservatism score times the median income of their state. Just so this process is clear, let's return to Senator Heflin. From page 71, we know that Senator Heflin's conservatism score is 26 and the median household income in his state (in 1970) is 7.4 (i.e., $7,400). So, having the computer multiply these scores, the score on the newly created interaction term between senator conservatism and state median household income for Senator Heflin is 192.4 (26 x 7.4 = 192.4).

As with the analysis of multicollinearity and causal models, an interaction term is not helpful to Marx. When the interaction term is included along with the three independent variables in the equation on page 68, the interaction term is not nearly statistically significant. If the interaction term replaces state median household income (i.e., the independent variables become senator conservatism, party affiliation and the interaction term between senator conservatism and state median household income) the interaction term is not close to being statistically significant. Marx is just having a rough time! His probable response to this would be that a false consciousness is causing the poor to misperceive their self-interest. He might be right. Many models in political science are interactive. In class you will see several interactive models from international relations, comparative politics, public law and American politics. Typically, the interaction terms will contain the variable names with an "x" (for multiplication) between them (e.g., conservatism x median income).

It is not uncommon for the dependent variable in a political science model to have only two or three possible responses. For example, in the international relations literature the dependent variable is often whether or not a dispute ended in war (e.g., 1 = war occurred; 0 = war did not occur). When the dependent variable has roughly 5, or fewer, possible responses regression is not an appropriate technique. Fortunately there are techniques, such as probit and

logit, which deliver the benefits of regression in situations where the dependent variable has few possible categories of response. There will be several such analyses presented in class.

One of the goals of this course is that you can interpret basic statistics. Because multiple regression is the most fundamental statistical tool of quantitative research, expect several quizzes and the final exam to ask you to interpret multiple regression results. A good way to prepare is to take the following practice quiz. The multiple regression results appear immediately below. After writing your answers, turn to the next page to compare your answers with the key.

Practice Quiz on Multiple Regression

Interpret the following regression results where:

Dependent Variable = the percentage of seats gained or lost by the president's party in the House of Representatives (i.e., if the computer read a score of -5 it would mean that the president's party lost 5% of the seats in the House of Representative in the last congressional election – since there are 435 seats, this would mean a loss of about 21 – i.e., 21 is about 5% of 435)

Independent Variable #1 = percentage change in real income per capita (meaning that income data have been adjusted for inflation and calculated the change on a per person basis – so if the computer read a score of 1.5 it would be mean that after removing the effects of inflation, income per person increased one and one-half percent since the last congressional election)

Independent Variable #2 = the percentage of the public that approves of the job the president is doing (i.e., if the computer reads a score of 38 it means that 38% of those surveyed approved of how the president was performing his job).

Multiple Regression Results:

Y Intercept: -17.7

Independent Variable #1: coefficient = 1.29; standard error of independent variable #1 = .29

Independent Variable #2: coefficient = -.25; standard error of independent variable #2 = .19

R–square = .47

Interpret the above results and then look at the next page.

Answers to Practice Quiz on Multiple Regression

Y Intercept: If the change in real income per capita is zero (i.e., the average person gained or lost nothing) and zero percent of the public approved of the president's job performance, the president's party would be predicted to lose 17.7% of the seats (about 70 seats) in the next congressional election. Note: the dependent variable is not the number of seats the president's party has in the House of Representatives but the percentage change in seats from the last election. So, -17.70 doesn't mean a prediction that if change in real per capita income and presidential popularity are both 0% the president's party would have -17.7% of the seats in the House. Rather, it means that the president's party would be predicted to lose 17.7% of the 435 seats in the House. For example, if the president's party had 50% seats after the last congressional election and lost 17.7% of the total seats in the next election they would then have 32.3% of the total seats in the House of Representatives (i.e., 50% - 17.7% = 32.3%). Although not necessary, you could make this more informative by converting the percentages into seats. Thus, the president's party would go from approximately 218 seats (218 is approximately 50% of 435) to about 141 seats (i.e., 141 is approximately 32.3% of 435).

Independent Variable #1: If the president's approval remained the same (e.g., 40% of the public approved of the president's job performance and it remained at 40%), for each one percentage point increase in real income per capita, on average, the president's party would gain approximately 1.29% of the seats in the House of Representatives (i.e., about 5 seats – 5 is about 1.29% of 435). Since the t ratio has an absolute value of over 2.0 (i.e., 1.29 is well over twice the size of .29) we reject the null hypothesis that the change in real income per capita has no effect on the change in the percentage of seats in the House of Representatives held by the president's party since the null hypothesis is true less than 5% of the time.

Independent Variable #2: If the change in real income per capita remained the same (e.g., it was running at 1% and remained at 1%) for each one percentage point increase in the percentage of the public that approves of the president's job performance, on average, the president's party would lose (remember, it's -.25 not .25) about one-fourth of one percentage point of the seats in the House of Representatives (since one-fourth of a percentage point of 435 is approximately 1 seat, you could mention this in addition to one-fourth of one percentage point decrease and make the answer a bit more informative – not a requirement however). Since the t ratio has less than an absolute value of 2.0 (i.e., -.25 is less than twice the absolute value of .19) we do not reject the null hypothesis that presidential approval is unrelated to the change in the share of seats the president's party has in the House of Representative because the null hypothesis is true greater than 5% of the time.

Interpretation of R-squared: Variation in the change in real per capita income and the president's approval rating together explain 47% of the variation in the percentage change in seats in the House of Representative held by the president's party.