# Types of information. Generalization of results of statistical research.

## What are variables.

Variables are things that we measure, control, or manipulate in research. They differ in many respects, most notably in the role they are given in our research and in the type of measures that can be applied to them.

Correlational vs. experimental research. Most empirical research belongs clearly to one of those two general categories. In correlational research we do not (or at least try not to) influence any variables but only measure them and look for relations (correlations) between some set of variables, such as blood pressure and cholesterol level. In experimental research, we manipulate some variables and then measure the effects of this manipulation on other variables; for example, a researcher might artificially increase blood pressure and then record cholesterol level. Data analysis in experimental research also comes down to calculating "correlations" between variables, specifically, those manipulated and those affected by the manipulation. However, experimental data may potentially provide qualitatively better information: Only experimental data can conclusively demonstrate causal relations between variables. For example, if we found that whenever we change variable A then variable B changes, then we can conclude that "A influences B." Data from correlational research can only be "interpreted" in causal terms based on some theories that we have, but correlational data cannot conclusively prove causality.

Dependent vs. independent variables. Independent variables are those that are manipulated whereas dependent variables are only measured or registered. This distinction appears terminologically confusing to many because, as some students say, "all variables depend on something." However, once you get used to this distinction, it becomes indispensable. The terms dependent and independent variable apply mostly to experimental research where some variables are manipulated, and in this sense they are "independent" from the initial reaction patterns, features, intentions, etc. of the subjects. Some other variables are expected to be "dependent" on the manipulation or experimental conditions. That is to say, they depend on "what the subject will do" in response. Somewhat contrary to the nature of this distinction, these terms are also used in studies where we do not literally manipulate independent variables, but only assign subjects to "experimental groups" based on some pre-existing properties of the subjects. For example, if in an experiment, males are compared with females regarding their white cell count (WCC), Gender could be called the independent variable and WCC the dependent variable.

Measurement scales. Variables differ in "how well" they can be measured, i.e., in how much measurable information their measurement scale can provide. There is obviously some measurement error involved in every measurement, which determines the "amount of information" that we can obtain. Another factor that determines the amount of information that can be provided by a variable is its "type of measurement scale." Specifically variables are classified as (a) nominal, (b) ordinal, (c) interval or (d) ratio.

a. Nominal variables allow for only qualitative classification. That is, they can be measured only in terms of whether the individual items belong to some distinctively different categories, but we cannot quantify or even rank order those categories. For example, all we can say is that 2 individuals are different in terms of variable A (e.g., they are of different race), but we cannot say which one "has more" of the quality represented by the variable. Typical examples of nominal variables are gender, race, color, city, etc.

b. Ordinal variables allow us to rank order the items we measure in terms of which has less and which has more of the quality represented by the variable, but still they do not allow us to say "how much more." A typical example of an ordinal variable is the socioeconomic status of families. For example, we know that upper-middle is higher than middle but we cannot say that it is,

for example, 18% higher. Also this very distinction between nominal, ordinal, and interval scales itself represents a good example of an ordinal variable. For example, we can say that nominal measurement provides less information than ordinal measurement, but we cannot say "how much less" or how this difference compares to the difference between ordinal and interval scales.

c. Interval variables allow us not only to rank order the items that are measured, but also to quantify and compare the sizes of differences between them. For example, temperature, as measured in degrees Fahrenheit or Celsius, constitutes an interval scale. We can say that a temperature of 40 degrees is higher than a temperature of 30 degrees, and that an increase from 20 to 40 degrees is twice as much as an increase from 30 to 40 degrees.

Ratio variables are very similar to interval variables; in addition to all the properties of interval variables, they feature an identifiable absolute zero point, thus they allow for statements such as x is two times more than y. Typical examples of ratio scales are measures of time or space. For example, as the Kelvin temperature scale is a ratio scale, not only can we say that a temperature of 200 degrees is higher than one of 100 degrees, we can correctly state that it is twice as high. Interval scales do not have the ratio property. Most statistical data analysis procedures do not distinguish between the interval and ratio properties of the measurement scales.

Relations between variables. Regardless of their type, two or more variables are related if in a sample of observations, the values of those variables are distributed in a consistent manner. In other words, variables are related if their values systematically correspond to each other for these observations. For example, Gender and WCC would be considered to be related if most males had high WCC and most females low WCC, or vice versa; Height is related to Weight because typically tall individuals are heavier than short ones; IQ is related to the Number of Errors in a test, if people with higher IQ's make fewer errors.

Why relations between variables are important. Generally speaking, the ultimate goal of every research or scientific analysis is finding relations between variables. The philosophy of science teaches us that there is no other way of representing "meaning" except in terms of relations between some quantities or qualities; either way involves relations between variables. Thus, the advancement of science must always involve finding new relations between variables. Correlational research involves measuring such relations in the most straightforward manner. However, experimental research is not any different in this respect. For example, the above mentioned experiment comparing WCC in males and females can be described as looking for a correlation between two variables: Gender and WCC. Statistics does nothing else but help us evaluate relations between variables. Actually, all of the hundreds of procedures that are described in this manual can be interpreted in terms of evaluating various kinds of inter-variable relations.

Two basic features of every relation between variables. The two most elementary formal properties of every relation between variables are the relation's (a) magnitude (or "size") and (b) its reliability (or "truthfulness").

a. Magnitude (or "size"). The magnitude is much easier to understand and measure than reliability. For example, if every male in our sample was found to have a higher WCC than any female in the sample, we could say that the magnitude of the relation between the two variables (Gender and WCC) is very high in our sample. In other words, we could predict one based on the other (at least among the members of our sample).

Reliability (or "truthfulness"). The reliability of a relation is a much less intuitive concept, but still extremely important. It pertains to the "representativeness" of the result found in our specific sample for the entire population. In other words, it says how probable it is that a similar relation would be found if the experiment was replicated with other samples drawn from the same population. Remember that we are almost never "ultimately" interested only in what is going on in

our sample; we are interested in the sample only to the extent it can provide information about the population. If our study meets some specific criteria (to be mentioned later), then the reliability of a relation between variables observed in our sample can be quantitatively estimated and represented using a standard measure (technically called p-value or statistical significance level, see the next paragraph).

What is "statistical significance" (p-value). The statistical significance of a result is the probability that the observed relationship (e.g., between variables) or a difference (e.g., between means) in a sample occurred by pure chance ("luck of the draw"), and that in the population from which the sample was drawn, no such relationship or differences exist. Using less technical terms, one could say that the statistical significance of a result tells us something about the degree to which the result is "true" (in the sense of being "representative of the population"). More technically, the value of the p-value represents a decreasing index of the reliability of a result (see Brownlee, 1960). The higher the p-value, the less we can believe that the observed relation between variables in the sample is a reliable indicator of the relation between the respective variables in the population. Specifically, the p-value represents the probability of error that is involved in accepting our observed result as valid, that is, as "representative of the population." For example, a p-value of .05 (i.e.,1/20) indicates that there is a 5% probability that the relation between the variables found in our sample is a "fluke." In other words, assuming that in the population there was no relation between those variables whatsoever, and we were repeating experiments like ours one after another, we could expect that approximately in every 20 replications of the experiment there would be one in which the relation between the variables in question would be equal or stronger than in ours. (Note that this is not the same as saying that, given that there IS a relationship between the variables, we can expect to replicate the results 5% of the time or 95% of the time; when there is a relationship between the variables in the population, the probability of replicating the study and finding that relationship is related to the statistical power of the design. See also, Power Analysis). In many areas of research, the p-value of .05 is customarily treated as a "border-line acceptable" error level.

How to determine that a result is "really" significant. There is no way to avoid arbitrariness in the final decision as to what level of significance will be treated as really "significant." That is, the selection of some level of significance, up to which the results will be rejected as invalid, is arbitrary. In practice, the final decision usually depends on whether the outcome was predicted a priori or only found post hoc in the course of many analyses and comparisons performed on the data set, on the total amount of consistent supportive evidence in the entire data set, and on "traditions" existing in the particular area of research. Typically, in many sciences, results that yield pОписание: Описание: Описание: Description: Опис : http://www.statsoft.com/textbook/graphics/lte.gif .05 are considered borderline statistically significant but remember that this level of significance still involves a pretty high probability of error (5%). Results that are significant at the p Описание: Описание: Описание: Description: Опис : http://www.statsoft.com/textbook/graphics/lte.gif .01 level are commonly considered statistically significant, and p Описание: Описание: Описание: Description: Опис : http://www.statsoft.com/textbook/graphics/lte.gif .005 or p Описание: Описание: Описание: Description: Опис : http://www.statsoft.com/textbook/graphics/lte.gif .001 levels are often called "highly" significant. But remember that those classifications represent nothing else but arbitrary conventions that are only informally based on general research experience.

Statistical significance and the number of analyses performed. Needless to say, the more analyses you perform on a data set, the more results will meet "by chance" the conventional significance level. For example, if you calculate correlations between ten variables (i.e., 45 different correlation coefficients), then you should expect to find by chance that about two (i.e., one in every 20) correlation coefficients are significant at the p Описание: Описание: Описание: Description: Опис : http://www.statsoft.com/textbook/graphics/lte.gif .05 level, even if the values of the variables were totally random and those variables do not correlate in the population. Some statistical methods that involve many comparisons, and thus a good chance for such errors, include

some "correction" or adjustment for the total number of comparisons. However, many statistical methods (especially simple exploratory data analyses) do not offer any straightforward remedies to this problem. Therefore, it is up to the researcher to carefully evaluate the reliability of unexpected findings. Many examples in this manual offer specific advice on how to do this; relevant information can also be found in most research methods textbooks.

Strength vs. reliability of a relation between variables. We said before that strength and reliability are two different features of relationships between variables. However, they are not totally independent. In general, in a sample of a particular size, the larger the magnitude of the relation between variables, the more reliable the relation (see the next paragraph).

Why stronger relations between variables are more significant. Assuming that there is no relation between the respective variables in the population, the most likely outcome would be also finding no relation between those variables in the research sample. Thus, the stronger the relation found in the sample, the less likely it is that there is no corresponding relation in the population. As you see, the magnitude and significance of a relation appear to be closely related, and we could calculate the significance from the magnitude and vice-versa; however, this is true only if the sample size is kept constant, because the relation of a given strength could be either highly significant or not significant at all, depending on the sample size (see the next paragraph).

Why significance of a relation between variables depends on the size of the sample. If there are very few observations, then there are also respectively few possible combinations of the values of the variables, and thus the probability of obtaining by chance a combination of those values indicative of a strong relation is relatively high. Consider the following illustration. If we are interested in two variables (Gender: male/female and WCC: high/low) and there are only four subjects in our sample (two males and two females), then the probability that we will find, purely by chance, a 100% relation between the two variables can be as high as one-eighth. Specifically, there is a one-in-eight chance that both males will have a high WCC and both females a low WCC, or vice versa. Now consider the probability of obtaining such a perfect match by chance if our sample consisted of 100 subjects; the probability of obtaining such an outcome by chance would be practically zero. Let's look at a more general example. Imagine a theoretical population in which the average value of WCC in males and females is exactly the same. Needless to say, if we start replicating a simple experiment by drawing pairs of samples (of males and females) of a particular size from this population and calculating the difference between the average WCC in each pair of samples, most of the experiments will yield results close to 0. However, from time to time, a pair of samples will be drawn where the difference between males and females will be quite different from 0. How often will it happen? The smaller the sample size in each experiment, the more likely it is that we will obtain such erroneous results, which in this case would be results indicative of the existence of a relation between gender and WCC obtained from a population in which such a relation does not exist.

Example. "Baby boys to baby girls ratio." Consider the following example from research on statistical reasoning (Nisbett, et al., 1987). There are two hospitals: in the first one, 120 babies are born every day, in the other, only 12. On average, the ratio of baby boys to baby girls born every day in each hospital is 50/50. However, one day, in one of those hospitals twice as many baby girls were born as baby boys. In which hospital was it more likely to happen? The answer is obvious for a statistician, but as research shows, not so obvious for a lay person: It is much more likely to happen in the small hospital. The reason for this is that technically speaking, the probability of a random deviation of a particular size (from the population mean), decreases with the increase in the sample size.

Why small relations can be proven significant only in large samples. The examples in the previous paragraphs indicate that if a relationship between variables in question is "objectively" (i.e., in the population) small, then there is no way to identify such a relation in a study unless the research

sample is correspondingly large. Even if our sample is in fact "perfectly representative" the effect will not be statistically significant if the sample is small. Analogously, if a relation in question is "objectively" very large (i.e., in the population), then it can be found to be highly significant even in a study based on a very small sample. Consider the following additional illustration. If a coin is slightly asymmetrical, and when tossed is somewhat more likely to produce heads than tails (e.g., 60% vs. 40%), then ten tosses would not be sufficient to convince anyone that the coin is asymmetrical, even if the outcome obtained (six heads and four tails) was perfectly representative of the bias of the coin. However, is it so that 10 tosses is not enough to prove anything? No, if the effect in question were large enough, then ten tosses could be quite enough. For instance, imagine now that the coin is so asymmetrical that no matter how you toss it, the outcome will be heads. If you tossed such a coin ten times and each toss produced heads, most people would consider it sufficient evidence that something is "wrong" with the coin. In other words, it would be considered convincing evidence that in the theoretical population of an infinite number of tosses of this coin there would be more heads than tails. Thus, if a relation is large, then it can be found to be significant even in a small sample.

Can "no relation" be a significant result? The smaller the relation between variables, the larger the sample size that is necessary to prove it significant. For example, imagine how many tosses would be necessary to prove that a coin is asymmetrical if its bias were only .000001%! Thus, the necessary minimum sample size increases as the magnitude of the effect to be demonstrated decreases. When the magnitude of the effect approaches 0, the necessary sample size to conclusively prove it approaches infinity. That is to say, if there is almost no relation between two variables, then the sample size must be almost equal to the population size, which is assumed to be infinitely large. Statistical significance represents the probability that a similar outcome would be obtained if we tested the entire population. Thus, everything that would be found after testing the entire population would be, by definition, significant at the highest possible level, and this also includes all "no relation" results.

How to measure the magnitude (strength) of relations between variables. There are very many measures of the magnitude of relationships between variables which have been developed by statisticians; the choice of a specific measure in given circumstances depends on the number of variables involved, measurement scales used, nature of the relations, etc. Almost all of them, however, follow one general principle: they attempt to somehow evaluate the observed relation by comparing it to the "maximum imaginable relation" between those specific variables. Technically speaking, a common way to perform such evaluations is to look at how differentiated are the values of the variables, and then calculate what part of this "overall available differentiation" is accounted for by instances when that differentiation is "common" in the two (or more) variables in question. Speaking less technically, we compare "what is common in those variables" to "what potentially could have been common if the variables were perfectly related." Let us consider a simple illustration. Let us say that in our sample, the average index of WCC is 100 in males and 102 in females. Thus, we could say that on average, the deviation of each individual score from the grand mean (101) contains a component due to the gender of the subject; the size of this component is 1. That value, in a sense, represents some measure of relation between Gender and WCC. However, this value is a very poor measure, because it does not tell us how relatively large this component is, given the "overall differentiation" of WCC scores.

Consider two extreme possibilities:

a. If all WCC scores of males were equal exactly to 100, and those of females equal to 102, then all deviations from the grand mean in our sample would be entirely accounted for by gender. We would say that in our sample, gender is perfectly correlated with WCC, that is, 100% of the observed differences between subjects regarding their WCC is accounted for by their gender.

If WCC scores were in the range of 0-1000, the same difference (of 2) between the average WCC of males and females found in the study would account for such a small part of the overall differentiation of scores that most likely it would be considered negligible. For example, one more subject taken into account could change, or even reverse the direction of the difference. Therefore, every good measure of relations between variables must take into account the overall differentiation of individual scores in the sample and evaluate the relation in terms of (relatively) how much of this differentiation is accounted for by the relation in question.

Common "general format" of most statistical tests. Because the ultimate goal of most statistical tests is to evaluate relations between variables, most statistical tests follow the general format that was explained in the previous paragraph. Technically speaking, they represent a ratio of some measure of the differentiation common in the variables in question to the overall differentiation of those variables. For example, they represent a ratio of the part of the overall differentiation of the WCC scores that can be accounted for by gender to the overall differentiation of the WCC scores. This ratio is usually called a ratio of explained variation to total variation. In statistics, the term explained variation does not necessarily imply that we "conceptually understand" it. It is used only to denote the common variation in the variables in question, that is, the part of variation in one variable that is "explained" by the specific values of the other variable, and vice versa.

How the "level of statistical significance" is calculated. Let us assume that we have already calculated a measure of a relation between two variables (as explained above). The next question is "how significant is this relation?" For example, is 40% of the explained variance between the two variables enough to consider the relation significant? The answer is "it depends." Specifically, the significance depends mostly on the sample size. As explained before, in very large samples, even very small relations between variables will be significant, whereas in very small samples even very large relations cannot be considered reliable (significant). Thus, in order to determine the level of statistical significance, we need a function that represents the relationship between "magnitude" and "significance" of relations between two variables, depending on the sample size. The function we need would tell us exactly "how likely it is to obtain a relation of a given magnitude (or larger) from a sample of a given size, assuming that there is no such relation between those variables in the population." In other words, that function would give us the significance (p) level, and it would tell us the probability of error involved in rejecting the idea that the relation in question does not exist in the population. This "alternative" hypothesis (that there is no relation in the population) is usually called the null hypothesis. It would be ideal if the probability function was linear, and for example, only had different slopes for different sample sizes. Unfortunately, the function is more complex, and is not always exactly the same; however, in most cases we know its shape and can use it to determine the significance levels for our findings in samples of a particular size. Most of those functions are related to a general type of function which is called normal.

Why the "Normal distribution" is important.

The "Normal distribution" is important because in most cases, it well approximates the function that was introduced in the previous paragraph (for a detailed illustration, see Are all test statistics normally distributed?). The distribution of many test statistics is normal or follows some form that can be derived from the normal distribution. In this sense, philosophically speaking, the Normal distribution represents one of the empirically verified elementary "truths about the general nature of reality," and its status can be compared to the one of fundamental laws of natural sciences. The exact shape of the normal distribution (the characteristic "bell curve") is defined by a function which has only two parameters: mean and standard deviation.

A characteristic property of the Normal distribution is that 68% of all of its observations fall within a range of ±1 standard deviation from the mean, and a range of ±2 standard deviations includes 95% of the scores. In other words, in a Normal distribution, observations that have a standardized value of less than -2 or more than +2 have a relative frequency of 5% or less. (Standardized value means

that a value is expressed in terms of its difference from the mean, divided by the standard deviation.) If you have access to STATISTICA, you can explore the exact values of probability associated with different values in the normal distribution using the interactive Probability Calculator tool; for example, if you enter the Z value (i.e., standardized value) of 4, the associated probability computed by STATISTICA will be less than .0001, because in the normal distribution almost all observations (i.e., more than 99.99%) fall within the range of ±4 standard deviations. The animation below shows the tail area associated with other Z values.

Illustration of how the normal distribution is used in statistical reasoning (induction). Recall the example discussed above, where pairs of samples of males and females were drawn from a population in which the average value of WCC in males and females was exactly the same. Although the most likely outcome of such experiments (one pair of samples per experiment) was that the difference between the average WCC in males and females in each pair is close to zero, from time to time, a pair of samples will be drawn where the difference between males and females is quite different from 0. How often does it happen? If the sample size is large enough, the results of such replications are "normally distributed" (this important principle is explained and illustrated in the next paragraph), and thus knowing the shape of the normal curve, we can precisely calculate the probability of obtaining "by chance" outcomes representing various levels of deviation from the hypothetical population mean of 0. If such a calculated probability is so low that it meets the previously accepted criterion of statistical significance, then we have only one choice: conclude that our result gives a better approximation of what is going on in the population than the "null hypothesis" (remember that the null hypothesis was considered only for "technical reasons" as a benchmark against which our empirical result was evaluated). Note that this entire reasoning is based on the assumption that the shape of the distribution of those "replications" (technically, the "sampling distribution") is normal. This assumption is discussed in the next paragraph.

Are all test statistics normally distributed? Not all, but most of them are either based on the normal distribution directly or on distributions that are related to, and can be derived from normal, such as t, F, or Chi-square. Typically, those tests require that the variables analyzed are themselves normally distributed in the population, that is, they meet the so-called "normality assumption." Many observed variables actually are normally distributed, which is another reason why the normal distribution represents a "general feature" of empirical reality. The problem may occur when one tries to use a normal distribution-based test to analyze data from variables that are themselves not normally distributed (see tests of normality in Nonparametrics or ANOVA/MANOVA). In such cases we have two general choices. First, we can use some alternative "nonparametric" test (or so-called "distribution-free test" see, Nonparametrics); but this is often inconvenient because such tests are typically less powerful and less flexible in terms of types of conclusions that they can provide. Alternatively, in many cases we can still use the normal distribution-based test if we only make sure that the size of our samples is large enough. The latter option is based on an extremely important principle which is largely responsible for the popularity of tests that are based on the normal function. Namely, as the sample size increases, the shape of the sampling distribution (i.e., distribution of a statistic from the sample; this term was first used by Fisher, 1928a) approaches normal shape, even if the distribution of the variable in question is not normal. This principle is illustrated in the following animation showing a series of sampling distributions (created with gradually increasing sample sizes of: 2, 5, 10, 15, and 30) using a variable that is clearly non-normal in the population, that is, the distribution of its values is clearly skewed.

However, as the sample size (of samples used to create the sampling distribution of the mean) increases, the shape of the sampling distribution becomes normal. Note that for n=30, the shape of that distribution is "almost" perfectly normal (see the close match of the fit). This principle is called the central limit theorem (this term was first used by Pylya, 1920; German, "Zentraler Grenzwertsatz").

How do we know the consequences of violating the normality assumption? Although many of the statements made in the preceding paragraphs can be proven mathematically, some of them do not have theoretical proofs and can be demonstrated only empirically, via so-called Monte-Carlo experiments. In these experiments, large numbers of samples are generated by a computer following predesigned specifications and the results from such samples are analyzed using a variety of tests. This way we can empirically evaluate the type and magnitude of errors or biases to which we are exposed when certain theoretical assumptions of the tests we are using are not met by our data. Specifically, Monte-Carlo studies were used extensively with normal distribution-based tests to determine how sensitive they are to violations of the assumption of normal distribution of the analyzed variables in the population. The general conclusion from these studies is that the consequences of such violations are less severe than previously thought. Although these conclusions should not entirely discourage anyone from being concerned about the normality assumption, they have increased the overall popularity of the distribution-dependent statistical tests in all areas of research.