

A Data Analysis Primer¹

Jon L. Pierce
University of Minnesota Duluth

Donald G. Gardner
University of Colorado – Colorado Springs

Students of the organizational and behavioral sciences (e.g., leadership, management, organizational behavior) reading empirical studies will inevitably encounter sections in the manuscript that present the researcher's findings. In the majority of these instances, researchers conducting empirical studies employ quantitative (vs. qualitative or narrative) approaches. As a consequence, statistical tools/techniques are usually employed in order to identify the presence and strength of relationships between the variables embedded in their research questions and hypotheses.

For those of you who may need a brief introduction to (or review of) statistics, we have developed this primer. What is to follow is neither an exhaustive examination of the statistical tools routinely employed in the organizational (or behavioral) sciences, nor should it be seen as a substitute for a regular course in statistics in the behavioral sciences. This primer does, however, provide a brief, non-technical description of several of the statistical techniques that you are likely to encounter in your reading of the leadership, management, and organization literatures. More specifically, we provide a brief overview of: descriptive statistics, reliability, validity, correlations, multiple regression, analysis of variance and multivariate analysis of variance, chi-square, structural equation modeling, and meta-analysis.

Descriptive Statistics

A sample is a group of people that researchers believe have the same characteristics as a much larger population. For example, a group of employees at Coors Brewery might be considered a sample of all brewery workers, or even of all employees. It is not feasible to study all members of a population (e.g., all employees), so researchers look at only a subset of the population (the sample). There are several statistics that are traditionally employed to describe (characterize) a particular sample. We refer to these as 'descriptive statistics.' These descriptive statistics include three measures of central tendency: the **mean**, **median**, and **mode**. Each provides us with insight into the overall distribution of the sample being studied. In addition, researchers and consumers of information are frequently interested in the amount of variability that is present in a sample. Variability is typically revealed by the *range* of scores and the sample's *standard deviation*. Starting with indicators of central tendency, each of these descriptive statistics is described below.

The **mean** is commonly known as the arithmetic average. It is one indicator of central tendency in the distribution of scores. It is computed by adding together all of the scores in a sample and then dividing by the number of scores. For example, if 50 employees responded to the ten-item instrument for the measurement of organization-based self-esteem (OBSE), the mean or average score for that sample would

¹ Reprinted from Leaders and the Leadership Process, edited by J. L. Pierce and J. W. Newstrom. New York: McGraw-Hill/Irwin, 2011.

be determined by adding together each study participant's answers to the 10 OBSE items, adding those 50 scores together, and then dividing by 50.

The **median** is a second measure of central tendency. It is that score which divides a distribution of scores (arrayed from lowest to highest) exactly in half – it is the middle score. If there are an odd number of scores (e.g., 1, 2, and 3) the median is the middle score (i.e., 2). If there is an even number of scores (e.g., 1, 2, 3, 4) the median is the average of the two middle scores (i.e., $2 + 3 = 5 / 2 = 2.5$). In both instances the median is that numeric value that separates the sample (or population) into a lower and upper half – an equal number of people in a sample score above and below the median.

The **mode** is the final measure of central tendency. The word mode means the customary fashion or a popular style. In statistics, the mode is that score that appears with the greatest frequency, or most often. Using our OBSE example, out of the 50 subjects, the most common score might be 4.00 (out of 7.00). It is possible that the mode is the same score as the mean, the same as the median, and it is possible that each reflects a different number.

The three measures of central tendency are related and they are determined by the shape of the distribution of obtained scores. A *symmetrical distribution* is one where the right-hand side of the distribution mirrors the image on the left-hand side. The symmetrical (or “bell shaped”) distribution is one characterized by the mean, median, and mode being exactly the same. If the distribution is said to be multi-modal, the mean and median are the same, while the distribution has two or more modes. Finally, there are *skewed distributions*. These distributions are “lopsided” (skewed). As such the peak (highest frequency or mode) lies to either the right side of the distribution (i.e., a negatively skewed distribution), or the peak lies to the left side of the distribution (i.e., a positively skewed distribution). The mean and median will lie to the left side of the mode in a negatively skewed distribution and to the right in a positively skewed distribution.

Variability simply means that not everything is the same. Researchers and consumers of information are often interested in knowing the amount of variability that is present in a particular set of scores. If we are looking at the strength of organizational commitment among the employees working for New Page Paper Company, we might be interested in knowing how similar or dissimilar these employees are from one another in terms of the strength of their commitment. If everyone has the same commitment score, there is no variability. By the same token, if there are large differences between scores, then the variability is said to be large. Thus, variability provides us with a quantitative measure of the degree to which scores in a distribution are clustered together or spread out from one another.

Range and Standard Deviation are two common indicators of variability. **Range** is calculated as the absolute difference between the largest score and the smallest score in the distribution. Organization-based self-esteem (OBSE), for example, is measured by an employee's answer to ten different questions, each of which is scored on a 7-point scale anchored 1 = Strongly Disagree to 7 = Strongly Agree (this is called a ‘Likert-type’ scale and is very common). Typically each respondent's score is calculated by adding their score to each of the ten items and dividing by ten. This places each individual's score back onto the 1 to 7 response scale that was initially employed in the measurement of the OBSE items. If 50 employees from New Page Paper responded to the OBSE instrument we would have 50 scores. The range of this distribution would be ascertained by identifying the smallest score (e.g., 2.0) and the highest score (e.g., 6.0), and calculating the difference between the two. Thus, our range would be 4 on a scale of 1 to 7 (where the maximum range is 6 [7 – 1]).

The **Standard Deviation** (SD), within the management and organizational sciences literature, is probably the most commonly used measure of variability. Standard deviation uses the mean of the distribution as a reference point and measures variability by considering the distance between each score and the mean. It determines whether the scores are generally near or far from the mean. It addresses the question –How much spread is there in the scores ‘around the mean?’ That is, are the scores clustered together or scattered? In simple terms, the standard deviation approximates the average distance from the mean (Gravetter & Wallnau, 1996, p. 110). Thus, the larger the SD score, the greater the degree of variability. As an example, a standard deviation of OBSE scores for a sample of New Page employees of 3.76 reflects much more variability amongst the sampled employees than an observed standard deviation of 1.34 from a sample of Minnesota Power employees.

If everyone’s score was the same (e.g., 4 on a 7-point Likert scale) the standard deviation would be 0. If everyone scored between 3 and 5, the standard deviation would be larger than 0, but it would be a small number. If half of the respondents scored 1 and half scored 7, the standard deviation would be a large number, because every score would be a long way from the mean (of 4.00).

Specifically, the standard deviation is equal to the square root of the variance. For example, if a distribution of scores had a *variance* of 25, the standard deviation would be equal to 5. To find a variance from a standard deviation, merely square the standard deviation (i.e., multiply it by itself; e.g., $5 \times 5 = 25$). Variances are computed by: (1) subtracting every score from the mean (called a deviation), (2) squaring that number (making all the numbers positive), and then (3) dividing by the number of people in the sample. Variances are always positive.

Once calculated, we need to interpret a standard deviation, which raises the question --What is big? What is small? These questions are not easy to answer. If we assume that the distribution of scores is a normal distribution (i.e., a bell shaped curve), we can draw the following conclusions: (1) 68% of the respondents have a score that lies between one standard deviation below to one standard deviation above the mean, (2) 95% of the scores will lie between two standard deviations below to two standard deviations above the mean, and (3) 99% are within three standard deviations from the mean (below to above).

The following example is taken from R.M. Jaeger’s *Statistics: A Spectator Sport* (Sage Publication, 1990.) Let’s employ an example assuming a normal distribution: a group of students completed a 100 question test where the mean score was 60 with a standard deviation of 10. Thus, 68 percent of the students will have test scores that are between 50 ($= 60 - (1 \times 10)$) and 70 ($= 60 + (1 \times 10)$). Ninety-five percent will have scores between 40 ($= 60 - (2 \times 10)$) and 80 ($= 60 + (2 \times 10)$). Finally, 99 percent will have scores that range between 30 ($= 60 - (3 \times 10)$) and 90 ($= 60 + (3 \times 10)$). The multipliers (1 x SD; 2 x SD; and 3 x SD) employed above reflect one, two, and three standard deviations from the mean (below to above) respectively.

Reliability and Validity

Reliability and validity will be treated together, as they are related, but distinct and are frequently confused with one another. Contrary to a commonly held belief, a reliable measure is not necessarily a valid one! Yet, a measure must be reliable if it is a valid measure. Reliability simply means that a measurement instrument is consistent or predictable in nature, while a measure that is valid measures what it is supposed to measure (as opposed to something else).

Reliability.

Reliability refers to the *consistency (predictability)* of a measurement, experiment, test, or observation, over two or more series of trials. It is, therefore, an indicator of the amount of confidence that we can place in, for example, the instrument that we are using to provide us with our data. Consider getting on and off of a bathroom scale, several times in a row, and without taking any clothing or other articles 'on or off.' If the reading is 130 each and every time, then we have a reliable (consistent, predictable) reading. If, however, the scale reads 130 the first time, 120 the second, and 150 the third, we would say the scale lacks consistency and/or it lacks reliability. An instrument that is reliable in an absolute sense is one where the 'observed score' does not contain any randomly appearing error (deviations from the real or true score). If the 'observed score' matches the 'true score,' each and every time, the instrument can be said to be absolutely reliable, as well as valid (i.e., it measures what it purports to measure). Reliability is a necessary precondition for validity, though it is not a sufficient condition, as a measure can be reliable and simultaneously not valid. For example, our bathroom scale might be very consistent/reliable, but it is not a valid measure if you weigh 150 pounds. It would be valid to the extent that readings on the scale (say, 130 pounds) reflect a person's true weight (130 pounds).

There are several ways of determining reliability, with test-retest and internal consistency being among the most frequently employed and reported in research conducted in the organizational sciences. **Test-retest Reliability** simply reflects the administration and re-administration of the same instrument to the same sample, coupled with an examination of the strength of the correlation between these two sets of data (correlations are discussed below). For example, on Monday, January 13 a researcher administers the Minnesota Job Satisfaction Questionnaire (MSQ) to a group of employees at General Electric. Then again some time later (e.g., February 14) the same questionnaire is re-administered to the same group of employees. In the event that there have been no major job changes during that one month time period (this is called a "history" effect), and if we find that the correlation between the Time-One data and the Time-Two data is quite high and statistically significant, then we may conclude that we have reasonably strong test-retest reliability (i.e., the two readings reflect a high degree of consistency).

Internal Consistency Reliability is a test of the homogeneity (sameness) of the items employed in a measurement instrument. The short-form of the MSQ employs twenty-items (i.e., 20 questions), with each one inquiring about the respondent's level of satisfaction with twenty different facets (e.g., ability utilization, compensation, security, responsibility, independence) that make up a job. Cronbach's (1951) *Coefficient Alpha* is the most popular test of internal consistency reliability. The alpha coefficient ranges from .00 to 1.00. As the number increases it reflects increasing homogeneity of scale items (i.e., indicating that the set of items are essentially 'tapping into' the same construct). Items which measure a particular variable, for example job satisfaction, should be more strongly related to one another than the strength of their relationship with items intended to measure other variables (e.g., organizational commitment). Thus, an instrument with a high level of internal consistency has items that are strongly related to one another.

While there is no absolute rule for what constitutes high internal consistency reliability, generally accepted norms reflect the following: (a) below .50 is unacceptable, (b) in the .60s is generally acceptable for scales in the early stages of development, (c) .70s to mid .90s is acceptable to increasingly strong, and (d) reliability coefficients in the high .90s suggest high reliability, and that some of the instrument's items may be unnecessarily redundant and that the scale could possibly be shortened by eliminating one or more items. (It is unlikely that we would ever reject a measure with a coefficient in the high .90s however.)

Validity

When we talk about validity, in this chapter, we are referring to the 'measurement' (e.g., tests, surveys, questionnaires) of specific constructs or variables in which we are interested (e.g., organizational commitment). Psychological *constructs* are aspects of an individual that we cannot see by just looking at the person (as opposed to something like hair color). In addition, constructs of interest to researchers have properties that vary (e.g., 'commitment' is a construct, in that some employees have no commitment, some have moderately strong degrees, and some have very strong levels). Thus, commitment has properties that vary). You cannot look at the picture of a stranger and say with certainty that she is extraverted or introverted. So, we use measures that produce scores (numbers) that we believe reflect different levels of the construct. We might use a personality test to measure just how extraverted an individual (or group of people) is. **Validity** is the degree to which scores on a measure really do reflect different levels of the construct it is intended to measure. For example, a measure of extraversion is valid to the extent that higher scores on the measure reflect higher levels of extraversion.

Scores on the measurement of most variables in the organizational sciences consist of three distinct components: $X = V + S + E$

Where: X = Score on a Measure
 V = Reliable and Valid Component
 S = Reliable, but Invalid Component
 E = Random Error Component (Unreliable by Definition)

V is the only component of a score that we are interested in. If X is an intelligence score (IQ), then V is that part of the score that really indexes intelligence. As you can see from the equation, V can only be as high as the reliability of the item. That's why it is often said that reliability is the upper limit of validity. That part of the item that is unreliable is E , and cannot be valid or predictable. So what is S ? S is the part of a score that is reliable, but totally unrelated to the construct in which we are interested. An example would be a person who always chooses alternative "D" on a multiple choice test when s/he does not know an answer (this is called a response bias). S/he would be very consistent (reliable), but that component of his/her total score would be unrelated to the construct of interest (e.g., knowledge of leadership theories).

Another good example of this equation is the controversial practice of intelligence testing. Some researchers believe that there is a very large component of IQ tests that is valid (which they abbreviate as g , for general intelligence). Other researchers claim that there is another large, reliable component of IQ scores that is unrelated to intelligence, which they refer to as cultural bias. Because most intelligence tests have been validated using primarily white, middle class samples, they may unintentionally be biased against minority and lower income groups. For example, one intelligence test asks a question about the 1960s pop artist Andy Warhol. Middle and upper class whites are much more likely to know who Andy Warhol was than minorities and lower income groups. This difference is not related to intelligence, but rather is systematic variance associated with the environment in which people grow up. We do not always know what S measures (there are dozens of possibilities). Plus, we are usually not interested in S , and we try to minimize S when we develop measures of constructs.

This discussion has been focused on one type of validity, which is termed **construct validity** (i.e., the extent to which an instrument actually measures that which it purports to measure, nothing more nor nothing less). It is that part of a test that actually measures what we say it measures. There are other types of validity that you may encounter in reading research articles (e.g., face validity, content validity, convergent validity, discriminant validity, criterion validity) which will not be discussed in this context.

Reliability and validity are the two major ways in which we evaluate the *psychometric properties* of a measure. The higher the reliability and validity of a measure, the better are its psychometric properties. Plus, the better the psychometric properties, the better will be the research that employs those measures to examine research questions and to test research hypotheses.

Correlation

In this section, we discuss the popular statistical technique correlation. In addition, we also comment on a couple of its variations, part and partial correlations, and canonical correlations.

Correlation is a statistical technique that is employed to measure and describe a relationship between two variables (this is also referred to as a bivariate correlation, as only two variables are involved). The correlation coefficient is revealed by what is typically called the "r" statistic (e.g., $r = .40$). Correlation coefficients are typically employed to facilitate prediction, and evaluate reliability (see test-retest reliability above), validity, and theory testing. The most common correlation is the Pearson Product-moment Correlation.

If a researcher is interested, for example, in whether or not there is a relationship between job satisfaction and organizational commitment, correlational analysis is a statistical technique that is capable of providing us with an answer to that question. In order to proceed, the researcher would need to collect data from a group of employees by measuring the strength of their job satisfaction and simultaneously (or at a later time period) the level of their organizational commitment. With these two sets of data, correlational analyses would enable the researcher to see if they are related, the strength of that relationship, whether or not the relationship is positive or negative in nature, and the level of statistical significance.

The strength of the relationship is revealed by a correlation coefficient that ranges from .00 (i.e., reflective of no relationship) to 1.00 (known as a perfect correlation). The larger the coefficient (e.g., $r = .80$ vs. $r = .20$) the stronger the relationship between the two variables in question. Taking a correlation and squaring it (i.e., r squared or r^2) converts the correlation to an indicator of percentage (which is sometimes called an "effect size"). This percentage reflects the amount of variance in one variable that can be explained or accounted for by knowledge of another variable. For example, if a researcher were to observe a correlation coefficient of .70 between job satisfaction and organizational commitment, that would mean that 49 percent (i.e., $.70 \times .70 = .49$) of the variance in organizational commitment (job satisfaction) can be explained by job satisfaction (organizational commitment).

The direction of the relationship is either positive or negative (or neither). A positive relationship means that the value of the two variables tends to move in the same direction. As we move across the sample from respondent to respondent, we witness that an increase in variable X (e.g., job satisfaction) is accompanied by an increase in variable Y (e.g., organizational commitment). A negative correlation between two variables implies that the two variables are moving in the opposite direction. For example, with an increase in variable X (e.g., job satisfaction) we witness a decrease in variable Y (e.g., organizational turnover). These two directions reflect the following scenarios: (1) as job satisfaction increases across our study participants we witness a simultaneous increase in their organizational commitment (i.e., a positive relationship), and (2) with an increase in job satisfaction we witness a decrease in the level of organizational turnover (i.e., a negative relationship). (You should note from this

example that a positive relationship is not necessarily “good,” just as a negative relationship is not necessarily “bad.”)

Typically when a correlation coefficient is presented it will be accompanied by its level of **statistical significance**. A statistic (like r) is said to be significant if we believe that it is not a ‘chance’ or ‘random event.’ In the organizational sciences (and the entire range of behavioral sciences) a ‘rule of thumb’ has been adopted which reflects a cut-off point where a particular relationship is treated as significant or not significant. This rule of thumb is $p \leq .05$, where p is the probability, $\leq$ reads less than or equal to, and .05 is a proportion out of 1.00. This simply means that the reported correlation between the two variables in question, given this particular sample size, could be this strong 5 or fewer times out of 100 **due to chance alone**. A significance level of $p \leq .01$ would mean that the observed correlation, given the sample size involved, would appear 1 time or less out of 100 by chance alone. Statistical significance and practical significance are not necessarily the same, as a very small correlation coefficient accompanied by a very large sample size could reach acceptable levels ($p \leq .05$) of statistical significance, and at the same time the relationship may not be very practical in terms of its total explanatory or predictive power.

The presence of a correlation should not automatically be seen as proof of a causal relationship between the two variables in question. Two variables may have a strong relationship simply because they co-vary (move together) as illustrated by the crowing of the rooster and the rising of the sun (along with a sleeping rooster and a setting sun). By the same token, under conditions where variable X is the cause of variable Y , we would expect to see a strong and significant relationship between them. If the relationship were zero, then it is probably true that X (the independent variable) does not cause Y (the dependent variable).

Part and Partial Correlations reflect a small variation upon correlations as just described. In each case there is, for example, a third variable whose effect upon the dependent variable, or both the independent and dependent variables respectively over which the researcher wants to exercise control over. In the case of *part correlations*, in Step One the dependent variable is correlated with the control variable and the unexplained variance (i.e., the residual, or ‘left-over’ variance) in the dependent variable is saved for employment in Step Two. In Step Two the independent and the ‘residual’ in the dependent variable are correlated with one another. The procedure for *partial correlations* is essentially the same, except that in Step One the control variable is correlated with both the independent and the dependent variable. The unexplained variance in both variables is used in Step Two, where these two residuals are correlated with one another in order to ascertain the relationship between the two variables independent of the effect in their relationship caused by the third variable—the one being treated as a control variable. The use of partial (part) correlations allows researcher to test research questions like “What is the correlation between job satisfaction and job performance, after controlling for any effects caused by organizational commitment?”

Finally, **Canonical Correlation** analysis is the multivariate analog of bivariate correlations (cf. Cooley & Lohnes, 1971). That is, canonical analysis allows for the simultaneous comparison of several independent *and* several dependent variables. Instead of examining a single predictor with a single criterion as in bivariate correlations, or multiple predictors against a single criterion as in multiple regression (see below), canonical correlation simultaneously uses multiple predictors as a single ‘predictor set’ and multiple criteria as a single ‘criterion set.’ For example, assume that a researcher is interested in the job design effects that are associated with the increasing bureaucratization of an organization’s structure. Job design has typically been portrayed multidimensionally (i.e., skill variety, task identity, task significance, autonomy and feedback) as has been organizational structure (e.g., centralization,

formalization, stratification, complexity). Treating organization structure as the independent variable and job design as the dependent variable canonical correlations will enable the researcher to ascertain whether or not the predictor and criterion sets of variables are related (correlated) to one another.

A significant canonical correlation (e.g., .50) between organization structure and job design indicates that the set of organization structure measures has a substantial relationship with the set of job design measures. Further analysis (called a 'variate analysis') would enable us to see which of the structure measures (e.g., formalization and complexity) most strongly correlates with the set of job design measures. Variate analysis would also enable us to identify the job design measures (e.g., variety and autonomy) most affected by organization structure.

In summary, canonical correlation analysis provides the researcher with several insights. The canonical correlation expresses the strength of association between the predictor and criterion variates (i.e., combinations of independent and dependent variables), and its squared value reflects the amount of variance common to that relationship. In addition, canonical correlation analysis informs the researcher as to the number of significant predictor-criterion variates that derive from the original variable sets, of those significant criterion variates which dependent variables (e.g., skill variety and autonomy) were involved and the amount of variance in each which can be explained. Those variables in the predictor set (e.g., formalization and complexity) that played the largest role in producing the relationship are identified. Finally, canonical correlation analysis reveals the overall all amount of variance in the criterion set that can be accounted for by the variables in the predictor set.

Next, we turn our attention to multiple regression. This statistical tool, somewhat similar to correlational analysis described above permits us to look at the association between multiple predictor variables and a single dependent variable. Thus, we will have examined tools for an examination of the relationship between a single independent and a single dependent variable, multiple independent and multiple dependent variables, and multiple independent and a single dependent variable.

Multiple Regression

In this section we describe the statistical technique called multiple regression. In addition, there are two special forms of the application of regression analyses for the testing of moderated and mediated relationships among a set of variables. Thus, we will also comment on moderated and mediated regression.

While the Pearson correlation is a technique for the assessment of a linear relationship between two variables, **Multiple Regression** is a technique used in the examination of a linear ("straight line") relationship between a single dependent variable (Y), and multiple independent variables (e.g., X_1 , X_2 , and X_3). It can also be used to evaluate curvilinear relationships ("curved line"). Multiple regression is one of the most frequently used statistics for testing research questions.

Presented below is an example of the results from a multiple regression where the dependent variable—organization-based self-esteem—was regressed on three individual difference (personality) variables, each of which is hypothesized to be a cause of the level of organization-based self-esteem that an individual develops within the work and organizational context (see Pierce & Gardner, 2009). This particular example uses what is called *stepwise* multiple regression

In Step One, the program looks to see which of the independent variables has the strongest relationship with the dependent variable and enters it into the equation first. In this instance Generalized Self-efficacy had the strongest relationship and was entered at Step 1. This step gives us an R statistic (strength of association much like a correlation), R Square (R^2 converts the R into a percent, which tells us the percent of the variance in the dependent variable that can be explained by the first variable to enter the equation), and the F statistic and its significance level (which assesses the overall significance of all of the variables in the equation at this juncture). At Step 2, Internal Locus of Control entered the equation. We can see a corresponding value for R (this R value is cumulative, as we now have two variables in the equation) and R^2 (the R^2 value is also the cumulative amount of variance in organization-based self-esteem that we can explain with the two variables in the equation), the percent of the observed changed in R^2 , and a new F statistic and its level of significance. Finally, at Step 3 the third variable, Global Self-esteem entered and we can see a new R, R^2 , the change in R^2 , and the new F statistic and its corresponding level of significance. In this case, all three personality variables were significantly related to OBSE (i.e., accounted for a significant portion of the variance in OBSE), and in the final analyses (i.e., at Step 3) the F statistic was significant ($< .01$) and the three personality variables accounted for 32 percent of the variance in OBSE.

Finally, the beta coefficients and their respective level of statistical significance are presented. At Step 3 all three independent variables are in the model. The *Beta Coefficient*, like a correlation coefficient, depicts the final strength of the relationship between the independent and dependent variables. Thus in this full model, Global Self-esteem correlates .20 with organization-based self-esteem and this relationship is significant at a level of less than .01. The beta coefficients for Generalized Self-efficacy and Internal Locus of Control respectively are .31 and .22, each significant at the $< .01$ level.

Dependent Variable: Organization-based Self-esteem

	R	R^2	Change in R^2	F	Significance of F
Step 1	.46	.21	.21	60.90	$< .01$
Step 2	.54	.29	.07	22.40	$< .01$
Step 3	.56	.32	.03	9.27	$< .01$

Entering at Step 1: Generalized Self-efficacy

Entering at Step 2: Internal Locus of Control

Entering at Step 3: Global Self-esteem

Predictor Variables	Beta Coefficient	Significance
Global Self-esteem	.20	$< .01$
Generalized Self-efficacy	.31	$< .01$
Internal Locus of Control	.22	$< .01$

There are two additional relationships that commonly appear in the organizational sciences for which a variation on multiple regression is employed. Some relationships are moderated, while others are mediated. A *moderated relationship* is one where a third variable (called a moderating variable) works on and produces a change in the relationship between the independent and the dependent variable. Thus, there is an interaction effect that is produced by the simultaneous appearance of the independent and the moderating variable. As the moderating variable changes, its interaction with the independent variable changes, thereby producing a different relationship between the independent and dependent variables.

A statistical technique called **moderated hierarchical regression** is commonly employed to test for this moderating effect. The following example illustrates what is happening in moderated hierarchical regression. It has been proposed in Hackman and Oldham's (1975) theory of job/work design proposed that complex job design and the individual difference variable growth need strength interact with one another as they influence the level of an employee's internal motivation. That is, growth need strength is presented as a moderator of the job design–internal motivation relationship, such that this relationship is hypothesized to be more positive for those with strong growth need and less positive for their low growth need counterparts.

Moderated hierarchical regression could be employed to test this relationship. The process would unfold in the following manner. First (i.e., Step 1 as described above in our discussion of multiple regression), the dependent variable (i.e., internal motivation) would be regressed on job design. This step would provide us with the same statistics as previously discussed, enabling us to see if there is a significant job design — internal motivation relationship, the amount of variance in the dependent variable accounted for, and the level of significance depicted by this relationship. In the next step (i.e., Step 2), the dependent variable would be regressed on the moderating variable (i.e., growth need strength). Once again, we will be able to ascertain whether or not there is a significant relationship between the moderating variable and internal motivation. Step 3 provides a look at the interaction effect. More specifically, each respondent's job design score is multiplied by their growth need strength score (i.e., Job Design score X Growth Need Strength score), and the dependent variable is regressed on this interaction term. If the interaction accounts for a significant portion of the unexplained variance in the dependent variable (i.e., variance in the dependent variable that has not been accounted for by job design in Step 1 and growth need strength in Step 2) there will be a significant change in R revealing support for the hypothesized moderated relationship. (That is, support for an interaction effect between job design and growth need strength and its effect on internal motivation.) Once again, Step 3 provides us with the same statistics as described in Step 3 above (i.e., in our general discussion of Multiple Regression).

Finally, researchers are frequently interested in *mediated relationships*. These are relationships where the effect of the independent variable upon the dependent variable is seen as 'working through' a third variable that we call the intervening variable. An example would be that job design affects job performance through the intervening variable of worker motivation. While there are several statistical techniques that can be employed to test such a relationship, mediated hierarchical regression (Baron & Kenny, 1986) has frequently been employed. Mediated regression is simply another structured use of multiple regression.

To illustrate the use of the Baron and Kenny (1986) technique, we will employ a hypothetical example. Envision three variables: X (the independent variable), Y (the dependent variable), and Z (the intervening or mediator variable). It is hypothesized that X affects Y by working through Z — thus: $X \rightarrow Z \rightarrow Y$. In order to test for the mediating effect of Z in the relationship between X and Y, the following steps are prescribed. At the outset there are three preconditions that need to be met for a mediated relationship to exist. More specifically, it is necessary for there to be a significant relationship between X and Z, X and Y, and Z and Y (that is, a significant relationship between the independent and intervening variables, between the independent and dependent variables, and between the intervening and dependent variables). These three relationships are necessary preconditions for the presence of mediation. Four statistical steps are employed:

Step One -- Regress the intervening variable on the independent variable.

Step Two -- Regress the dependent variable on the independent variable.

Step Three -- Regress the dependent variable on the intervening variable.

Step Four -- Regress the dependent variable on both the independent and the intervening variables.

According to Baron and Kenny (1986), a mediated relationship is supported when the following conditions are met:

1. The independent variable significantly influences the intervening variable in the first equation.
2. The independent variable significantly influences the dependent variable in the second equation.
3. The intervening variable significantly influences the dependent variable in the third equation.
4. If these conditions hold, then the influence of the independent variable on the dependent variable must be less in the fourth equation than in the second equation.

Finally, full mediation occurs when the independent variable has no significant effect in Step 4 after the intervening variable is controlled. (Thus, we have gone from a significant relationship between the independent and dependent variable without controlling for the effect of the intervening variable on the dependent variable, to a situation where there is no longer a significant relationship between the independent and dependent variables when the effect in the dependent variable due to the intervening variables has been taken into consideration.) Partial mediation occurs when the independent variable continues to influence the dependent variable (in Step Four) after controlling for the intervening variable, though by an appreciably smaller amount (i.e., R while still significant it is much smaller than it was previously). Thus, under conditions of partial mediation the independent variables works on the dependent variable in part by working through the intervening variable (this is called an 'indirect effect'), and in part by it having a 'direct effect' on the dependent variable.

As with all of the approaches involving multiple regression the statistics employed (e.g., R , R^2 , F , and its level of significance) are virtually the same.

Analysis of Variance and Multivariate Analysis of Variance

It is common for researchers working in the behavioral sciences to employ a factorial design, where they are interested in ascertaining whether or not there are significant differences between groups (e.g., those employees working standard hours vs. those working flexible hours), or within groups across time and/or situations (e.g., comparing a group of employees initially working a standard 5 day, 40 hour week vis-a-vis their working under a new compressed work week schedule where they now work a 4 day, 40 hour week).

Analysis of Variance (ANOVA) and **Multivariate Analysis of Variance (MANOVA)** are commonly employed statistical techniques in the analysis of groups. The primary difference between **ANOVA** and **MANOVA** is that the former is designed to make a comparison across two or more groups on a single dependent variable (e.g., job satisfaction), where we are essentially asking –Is there a significant difference, for example, in terms of the level of job satisfaction between those who are working standard hours (e.g., 8:00 to 5:00) and those who are working flexible working hours (where each employee can exercise an element of discretion in defining the time of day that they start and end their eight hour day). **MANOVA** facilitates the simultaneous examination of two or more dependent variables. For example, we might be interested in whether there is a lower level of stress, and a higher level of job satisfaction and work attendance (i.e., three dependent variables) among a group of employees who are moved from a 5/40 work week to a 4/40 compressed work week.

The following example derives from Pierce and Newstrom's (1982) examination of flexible working hours. There were four groups of employees, one working a 'fixed hour schedule' and three groups each working 'flexible working hours' were studied. These three flex systems were differentiated from one another in terms of increasing levels of flexibility as we transition from flex-schedule #1 to flex-schedule #3. Based upon a literature review, it was hypothesized that with increasing flexibility there would be an increase in worker job attitudes: job satisfaction, hours of work satisfaction, and job involvement. The table below presents the study results and reveals the statistics employed in both ANOVA and MANOVA.

	Means				Univariate		Multivariate	
	Fixed	Flex-1	Flex-2	Flex-3	F	p	F	p
<u>Job Attitudes:</u>							5.28	.00*
Job Satisfaction	59.18	70.05	71.64	71.40	5.25	.00*		
Hours of Wk Satisfaction	4.41	5.37	6.08	5.90	10.10	.00*		
Job Involvement	11.52	11.58	---	14.90	5.59	.01*		

Column 1 lists the three job attitudes that are being treated as the study's dependent variables. Columns 2 – 5 present the mean score on each of the dependent variables that were derived from the employees working under each of the four working hour arrangements. Columns 6 and 7 present the ANOVA results (i.e., the F statistic [Wilks Lambda F statistic] and its level of significance for each of the dependent variables, as ANOVA only examines the effects for a single dependent variable at a time). Finally, Columns 8 and 9 present the results from the MANOVA, where the set of three dependent variables are treated simultaneously. As we can see there is a significant difference across the four work schedules in terms of employee job-related attitudes (i.e., $F = 5.28$, $p \leq .05$). In addition, from the univariate analyses (ANOVA) we also see that there are significant differences across the groups for each of the three dependent variables. We can look at the means in order to see what is happening for each dependent variable as we move from those working under the Fixed Hour Schedule to each of the successive Flexible Working Hour arrangements. Note, for example, in each instance those working under the fixed hour schedule had the lowest job attitudes.

What we don't know at this juncture is 'where' it is that these significant differences lie as there are more than two groups. If our study only involved two groups, we could simply look at the mean scores to see which group has the highest and which group has the lowest mean score, thereby revealing the nature of the significant difference detected by the ANOVA F statistics. Here, however, we have four different groups and it is not yet known exactly where the significant difference that was detected stems from.

To move beyond that summary and to ferret out a more specific examination of the location of the significant differences that have been identified by either ANOVA or MANOVA, we need to employ additional (post hoc) analyses. By employing any one of several analytical techniques (e.g., Scheffe test), the researcher can perform further 'planned comparisons.' In the study presented above, Pierce and Newstrom (1982) employed the Scheffe test to examine differences between, for example the fixed and flexible schedules. They reported that the fixed hour group was significantly less job involved, and less job- and hours of work satisfied, than employees working under some form of flexible working hours. In other words, the flex groups had more positive attitudes than the fixed-hours group.

Chi-square

Sometimes researchers collect data that is categorical in nature. Examples of categorical data are gender (male vs. female), type of organization (e.g., profit vs. non-profit), and race (e.g., white vs. black vs. Asian). It does not make any sense to look at the mean in such a data set. What is the average of gender? Of race? Instead, researchers report the numbers and types of data points within each category, which we call frequencies. The **Chi-square** statistic allows researchers to determine if there is a significant difference in frequencies between the different categories. Because the chi-square statistic is not dependent on the mean (or standard deviation) in the data, it is sometimes called a *non-parametric* statistic (there are many others, as well).

An example will help clarify what the chi-square statistic does. Assume a researcher wants to know if there are gender differences in perceptions of effective leaders. He collects data from a group of 100 employees, 50 men and 50 women, asking each: Do you think men are more effective leaders than women (Yes or No)? If there are no differences between men and women in response to this question (our “expected frequency”), the frequencies of yes answers should be the same for both groups (and implicitly the no answers as well). If 40 out of 50 of the men actually answered “yes” to the question, and 30 out of 50 of the women answered “yes,” is that difference ($40 - 30 = 10$) significantly different? Those are the types of questions that the chi-square statistic allows us to examine. It is a very flexible statistic, and is used in a wide variety of analyses. It addresses basic questions that compare the actual data collected to an expected result, usually finding that there are no differences between the multiple categories.

Structural Equation Modeling

Theory forms the basis for structural equation modeling. Starting with what is known from prior research on a set of psychological constructs (e.g., organization-based self-esteem, psychological ownership, perceived organizational support), new hypotheses about causal relationships between these constructs are proposed. **Structural equation modeling** (SEM) may then be used to test those causal relations among those psychological constructs of interest. Structural equation analysis allows researchers to examine causal relationships among hypothetical constructs (termed latent variables), as opposed to only the product moment correlations among the measured (or manifest) variables. Foremost among the well-documented advantages of structural equation analysis is that random measurement error is removed from the estimated relationships among the latent variables. As a result, the magnitudes of the statistical associations among them are not attenuated by a lack of measurement reliability. Therefore, structural equation analyses provide estimates of relationships that are less biased than they would be if one examined only the measured variables of a conventional correlational technique. SEM allows us to approximate the ‘V’ (see the preceding discussion of validity) in the true score equation without the ‘S’ and ‘E.’

In a structural equation model the latent variables are hypothesized to cause variation in the measured variables. The degree to which the causal relationships specified by the theory actually account for observed variances and covariances are assessed using several measures of fit (termed goodness-of-fit indices), because of problems that can occur when only a single measure is used (e.g. parsimony, type of estimation method, sample size, non-normality, and type I errors). There are a number of goodness-of-fit indexes that may be used to assess the fit of a SEM model to the data, though absolute cutoff values for these indexes are a matter of debate. Currently, the recommended indexes and cutoff values are .95 (or higher) for the comparative fit index (CFI), .08 (or lower) for the standardized root mean square residual (SRMR), and .06 (or lower) for the root mean squared error of approximation (RMSEA), before concluding that a model fits the raw data reasonably well.

SEM is not perfect. Inferences about causality using SEM are no stronger than inferences made from a correlation matrix, even when the original data are derived from well-controlled experiments. Hypothesized causal arrows may actually go in the reverse direction, or be reciprocal in nature. SEM does not control for weak theorizing. SEM does allow researchers to examine hypothesized relationships between variables after controlling for error variance and other sources of "noise" in the measured variables.

Meta-Analysis

Finally, **Meta-analysis** is a statistical technique commonly employed in the conduct of a literature review, where the reviewer is interested in summarizing several empirical studies quantitatively as opposed to qualitatively (i.e., judgmentally). The studies of interest are those that addressed a similar hypothesized relationship (e.g., There is a positive relationship between organization-based self-esteem and organizational citizenship behaviors). In a generalized sense, meta-analysis provides a statistic which reflects the averaging of effect size from the several studies included in the literature review.

Imagine a situation where a researcher collects intelligence and job performance data from engineers at a division of a high tech firm (e.g., Hewlett Packard in Palo Alto, CA). She correlates the two measures and obtains an $r = .45$ ($p < .05$), meaning that the higher the intelligence, the higher the job performance. Now, imagine that same researcher, with the same measures, collects data from a similar group of engineers at the same company, but at a different location (e.g., Colorado Springs). She correlates the two measures and obtains an $r = .15$ ($p > .05$, non-significant). Why would the correlation be strong and significant in one location, yet small and non-significant (ns) in a different location?

In many if not most cases of discrepancies like this, the cause may be attributed to what is called "statistical artifacts," or the presence of an extraneous variable which is causing variability in the observed relationships. Statistical artifacts all act to lower the strength of the observed relationships (such as correlations). Some research studies are plagued by these statistical artifacts more than others, causing those researchers to obtain smaller observed relationships. If some research studies are affected by these artifacts, and others are not (or less so), then we would expect to see this inconsistency in obtained results across different research studies.

This is a problem that researchers have encountered since they first started collecting data. Karl Pearson, one of the "fathers" of modern statistics, is credited with performing the first meta-analysis in 1904. Meta-analysis is a procedure that averages effect sizes (such as correlations) across a number of different studies of the same relationship (e.g., job satisfaction and job performance), after correcting for the effect sizes for the statistical artifacts. The resulting *average* effect size is believed to come closest to the "true" effect size in the population from which the research samples were drawn (e.g., employees in general). One important example from organizational behavior was Judge et al.'s (2001) meta-analysis of the relationship between job satisfaction and job performance, in which he concluded that the population estimate is an $r = .26$ (small, but significant).

Meta-analysis cannot correct for all types of errors and biases (such as errors in calculations and data-entry errors). Which statistical artifacts are corrected in meta-analysis? The most common corrected artifacts in management research are (Hunter, Schmidt, & Jackson, 1982):

1. Sampling error: Some studies use samples that are more like the population than others.

2. Restriction of range: Smaller samples usually have a smaller range of scores on measures than do larger samples.
3. Unreliability in the predictor measures.
4. Unreliability in the criterion measures.

If these corrections reduce the variability of reported results by 80% or more, usually the meta-analyst is done. The average effect size is our expected value for the population. If there is still some variability left across different research studies, researchers might look for moderator or mediator variables (as discussed above). For example, Bowling (2007) conducted a meta-analysis in which he found that the relationship between job satisfaction and job performance is reduced to zero after controlling for OBSE.

The major advantages of meta-analysis are that they quantify the degree to which one variable is related to another variable as revealed across several different investigations, while avoiding biases which might occur in qualitative reviews of the same body of research. It allows us to get a better estimate of the “true” strength of relationship between two (or more) variables – a very important feature in organizational research.

In Closing

We hope that this short tutorial, without an emphasis on formulae and an extensive use of numbers will be of assistance to you as you read the empirical organizational sciences literature. There are a number of very good statistical books, written for introductory students in the behavioral sciences (e.g., F. J. Gravetter & L. B. Wallnau, *Statistics for the Behavioral Sciences*, Minneapolis/St. Paul, MN: West Publishing, and previously mentioned R. M. Jager) that we strongly recommend finding their place in your permanent library. Your future practice of **evidence-based management** is dependent upon your ability to be a sophisticated consumer of the research literature as it unfolds in your discipline. An understanding of basic data analytic techniques is and will be of extreme importance, making that job easier, more enjoyable, and ultimately more informative.

References

- Baron, R. M., & Kenney, D. A. (1986). The moderator-mediated variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, 51, 1173-1182.
- Bateman, T. S., & Ferris, G. R. (1984). *Method and analysis in organizational research*. Reston, VA: Reston Publishing.
- Bowling, N. A. (2007). Is the job satisfaction-job performance relationship spurious? A meta-analytic examination. *Journal of Vocational Behavior*, 71, 167-185.
- Coolley, W. W., & Lohnes, P. R. (1971). *Multivariate Data Analysis*. New York: Wiley.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Gravetter, F. J., & Wallnau, L. B. (1996). *Statistics for the behavioral sciences*. Minneapolis/St. Paul, MN: West Publishing.

- ackman, J. R., & Oldham, G. R. (1975). Development of the job diagnostic survey. Journal of Applied Psychology, 60, 159-170.
- Hunter, J. E., Schmidt, F. L. & Jackson, G. B. (1982). Meta-analysis: Cumulating research findings across studies. Beverly Hills, CA: Sage.
- Judge, T. A., Thoresen, C. J., Bono, J. E., & Patton, G. K. (2001). The job satisfaction-job performance relationship: A qualitative and quantitative review. Psychological Bulletin, 127, 376-407.
- Pierce, J. L., & Gardner, D. G. (2009). Relationships of personality and job characteristics with organization-based self-esteem. Journal of Managerial Psychology, 24:5, 392-409.
- Pierce, J. L., & Newstrom, J. W. (1982). Employee responses to flexible work schedules: An inter-organization, inter-system comparison. Journal of Management, 8:1, 9-25.
- Warwick, P. V. (1975). Canonical correlation analysis: Subprogram cancorr. Chapter 25, In N. H. Hie, C. H. Hull, J. G. Jenkins, K. Stein-Brenner, & D. H. Bent (Eds.), Statistical Package for the Social Sciences, New York: McGraw-Hill.

January 2010