Chapter

1

Identification of Empirical Setting

Purpose for Identifying the Empirical Setting

The purpose of identifying the empirical setting of research is so that the researcher or research manager can appropriately address the following issues:

1) Understand the type of research to be conducted, and how it will affect the design of the research investigation,

2) Establish clearly the goals and objectives of the research, and how these translate into testable research hypothesis,

3) Identify the population from which inferences will be made, and the limitations of the sample drawn for purposes of the investigation,

4) To understand the type of data obtained in the investigation, and how this will affect the ensuing analyses, and

5) To understand the limits of statistical inference based on the type of analyses and data being conducted, and this will affect final study conclusions.

Inputs/Assumptions Assumed of the Investigator

It is assumed that several important steps in the process of conducting research have been completed before using this part of the manual. If any of these stages has not been completed, then the researcher or research manager should first consult Volume I of this manual.

1) A general understanding of research concepts introduced in Volume I including principles of scientific inquiry and the overall research process,

2) A research problem statement with accompanying research objectives, and

3) Data or a data collection plan.

Outputs/Products of this Chapter

After consulting this chapter, the researcher or research manager should be able to do the following:

1) Understand the purpose of the statistical models or methods to be used, and how they can be used to address specific research hypotheses,

2) Understand the limitations of the sample data as it relates to the population of interest,

3) Understand the type of data and how it will affect the choice of statistical method,

4) Understand which are the dependent variables, independent variables, and the role that the experiment or observational study in collecting them,

5) Issues concerning sample sizes as they relate to conducting statistical hypothesis tests.

Troubleshooting: Empirical Setting Frequently Asked Questions

What is the difference between models that predict and models that explain?
How do I know if the sample data are random?
How do I know if the sample data are representative of the population?
What are the different scales of measurement and why are they important?
How are research objectives converted into research hypotheses?
How large of a sample does the analyst need?
What is a dependent variable?
What is an independent variable?
What are the differences between experiments, quasi-experiments, and observational studies?
What is a statistical error?
What is alpha and beta, and how do they affect research conclusions?

References on Empirical Setting

1) Levy, Paul S. and Lemeshow, Stanley (1999). Sampling of Populations, Methods and Applications. Third Edition. John Wiley & Sons, Inc. New York, New York.

2) Som, Ranjan K. (1996). Practical Sampling Techniques, Second Edition. Marcel Dekker, Inc. New York, New York.

3) Pedhazur, Elasar J. and Schmelkin, Liora Pedhazur. (1991). Measurement, Design, and Analysis: An Integrated Approach. Lawrence Earlbaum Associates, Hillsdale, New Jersey.

Introduction

Transportation research involves studies pertinent to the analysis, design, and optimization of transportation systems including:

- The planning, design, construction, operation, safety, and maintenance of transportation facilities and their components;

- The economics, financing and administration of transportation facilities and services; and,

- The interaction of transportation systems with one another and with the physical, economic, and social environment.

In general, transportation research is applied, i.e., transportation research studies are generally designed and conducted with practical goals and objectives in mind. These goals might include enhanced understanding of a process, prediction of future performance, or process verification and validation.

In addition, research in transportation is based on data that are predominately observational in nature. In other words, much research is based upon data that cannot be controlled by the analyst. Assessing data and building statistical models based on observational data raises a host of additional issues that warrant additional attention and savvy from the researcher.

In order to better understand phenomena, transportation research requires the collection and analysis of data. A variety of issues is associated with data: it's relationship with time, how it is quantified or measured, how much is necessary, how it is handled, and most importantly, it's relationship to the purpose and objectives of the investigation.

Purpose of investigation

The engineer conducts research for a variety of reasons. The intended end-use of a statistical model or hypothesis test helps to guide the researcher through the research process; including the design of the study, the collection of data, the analysis procedures, and the dissemination of results.

Predictive Research

Predictive research attempts to predict the value of a dependent variable using information from a set of "predictor" or independent variables. These studies are concerned primarily with the degree of successful prediction, rather than an improved understanding of the underlying processes that affects the phenomenon. The fact that an understanding of the underlying phenomenon is not of prime importance affects the identification and selection of predictor variables.

Explanatory Research

Explanatory research aims to understand the underlying processes affecting a phenomenon by associating the variability of the phenomenon of interest (dependent variables) based on an assumed set of causal predictor or explanatory variables. Because the prime importance is greater understanding of the phenomenon, predictor variables must be selected with utmost care, and data should be collected under controlled conditions to the extent possible (i.e. an experiment).

Research for Quality Assurance and Control

Research for quality assurance and control focuses on the broad topic of stability of a measurement precision of a response variable typically resulting from a process,

procedure, or system. Often this type of research is conducted to ensure that a process, system, or procedure is being conducted at an acceptable level. An example is research to ensure that a minimum concrete compression strength of mixes is sustained by a concrete manufacturing plant.

Types of Research Investigations

There are generally three types of research investigations that are conducted: experiments, quasi experiments, and observational studies. Descriptive statistics and statistical models estimated on data obtained from experiments are the least susceptible to internal and external threats to validity, followed by quasi-experiments, and then by observational studies. In other words, there is greater assurance that relationships developed on experimental data are in fact 'real', as opposed to illusory.

Experiments

Experiments are research studies conducted under specific and controlled conditions to verify or refute a working set of research hypotheses. Experiments are the cornerstone of statistical theory, and are the soundest method for providing support for or against a theory. Experimental or research hypotheses cannot be proved using statistics, instead, evidence can be mounted in support or against their validity. Essential elements of an experiment require that at least one variable be manipulated, and random assignment be used to assign experimental units to different levels or categories of the manipulated variable.

Bridges example: In an experiment on the longevity of three different protective paints for steel bridges, the analyst randomly selects 24 bridges from her state for inclusion in the study. She then randomly assigns the protective paint to the 24 bridges, such that each paint is applied at random to eight bridges. To control for the effect of application differences, she hires four painting contractors and assigns each contractor to paint two bridges in each treatment group.

Quasi-experiments

For a research study to be classified as a quasi-experiment, the researcher manipulates at least one variable, however, experimental units are not randomly assigned to different levels or categories of the manipulated variable. It should be noted that when random-assignment is not employed, that self-selection bias has the potential to bias results of the study with an omitted or confounded variable problem.

Quasi-experiments suffer from two errors not found in true experiments: specification error, and self-selection error. Specification errors describe cases when relevant variables correlated with the independent variable are not identified nor included in the analysis. Self-selection errors describe where the assignment of the independent variable to specific categories is related to the status of the dependent variable. Many studies conducted in transportation are quasi-experiments.

Safety example: In a study of the safety effectiveness of replacement of 4-way stop control with signalized intersections, a researcher notes that 'conversions' from 4-way stop control were conditional upon traffic volumes and crash experience. Thus, sites that received the experimental 'treatment' were self-selected based upon traffic volumes and crash rate, which has a direct impact on crashes, the dependent variable.

Observational Studies

The literature does not differentiate clearly the differences between quasi-experiments and observational studies. An observational study certainly does not contain an important element of an experiment: random assignment of subjects to treatments. To further differentiate it from a quasi-experiment, an observational study might lack the ability to assign levels of the manipulated variable. In other words, the experimenter might not have the ability to control the levels of the manipulated variable, unlike a quasi-experiment. An observational study, furthermore, might have a large number of non-controlled random variables that are thought to affect the dependent variable.

Safety example: A researcher may be interested in determining whether there is a relationship between emergency vehicle response times (in attending a primary incident or crash), and the probability of a secondary incident or accident. In other words, the researcher is interested in knowing whether faster emergency vehicle response times will reduce the number of crashes caused by incident induced congestion. To conduct this type of investigation, the analyst gathers data on incident occurrence, emergency response times to those incidents, and the occurrence of secondary crashes. In this observational study, none of the independent variables thought to affect secondary crashes, such as emergency vehicle response times, incident clearance times, traffic volumes, site characteristics, etc., are under the control of the investigator, and so the investigation is observational in nature.

Statistical models estimated using observational data are the least robust with respect to validity of all research investigations. It is difficult to determine whether relationships identified using observational data are real, or if they are the result of illusory correlation. Because observational studies are typically less well planned than experiments, inferences based on them tend to be more plagued with post-hoc theorizing.

Statistical Models and Modeling

Often the end-result of research, whether for prediction, explanation, or quality control, is a statistical model or set of models. A statistical model is an abstract description of the real world. A model is a simple representation of the complex forms, processes, and functional relationships of a "real world" transportation counterpart. Because many of the processes underlying transportation research are complex and a function of many inputs (independent variables), the statistical models representing them by necessity are much simpler. In fact, a primary objective of statistical modeling is to represent complex phenomenon with as simple a model as possible.

Although it usually possible to identify the primary objective of a research project, models estimated for predictive or explanatory purposes are related. When the analyst can identify functional or cause-effect relationships between objects or events, then superior models result, and often are useful for both explanation and prediction. In many cases, however, the suspected 'causal' independent variables are not directly measurable, are

too expensive or impractical to measure, and so the analyst seeks surrogate variables that are correlated with the causal ones.

Two fundamental features characterize all models: form (relationships) and content (variables.) The choice of form establishes the ease of manipulating and interpreting the content; detecting errors of omission and commission; and refining and improving the model to better serve its purpose.

When describing transportation systems, model can be broken down into four different types (Rubenstein, 1975):

Probabilistic Models

Probabilistic models are used to deduce certain conclusions about a sample drawn from an entire population (deductive inference), or to infer population parameters using measurements from a sample (inductive inference). Estimates of population parameters such as population mean and population variance are known as statistics.

Decision-Making Models

Decision-making models are used to select between alternate courses of action. In essence, any time a decision-maker is faced with a finite set of choices or courses of action, a decision making model can in theory be estimated given relevant attributes about the decision makers and/or the choices.

Decision-making models are classified according to the assignment of probabilities to the states of nature: decision-making under certainty, decision-making under risk, decision-making under uncertainty, and decision-making under conflict (game theory).

Optimization Models

There often exists more than one possible solution to complex transportation problems. Not only does the researcher face the problem of finding one solution, the researcher must select between multiple feasible solutions to determine the “best” solution. The class of models for selecting the “best” is described as “optimization models.”

The procedures used to select between feasible solutions given a known decision criterion are know as “mathematical programming” techniques (to be distinguished from computer programming techniques). Depending on the form of the decision variables or objective function, alternate programming techniques may be used.

Linear programming models attempt to maximize a linear objective function subject to a set of linear constraints. Linear programming problems can be solved graphically for two or three variables. For more variables, the simplex algorithm is used. Nonlinear programming models are similar to linear models, except that at least one of the equations in the model is a nonlinear function of the decision variables.

The general linear programming transportation problem typically has a number of equivalent “best” solutions. Most commercial software picks one to display and ignores all other equivalent “best” solutions. The choice amongst equivalent “best” solutions may have operational or management advantages, and so it is prudent to obtain all equivalent solutions to a specific problem.

Dynamic programming is useful in problems of sequential decisions and is based on the principle of optimality. Dynamic programming is more a fundamental element in the philosophy of problem solving than an algorithm for the solution of a class of problems. Formulation of a dynamic programming model is more an art than a science.

Dynamic Systems Models

A system is a collection of elements aggregated by virtue of the links of form, process, or function, which tie them together and cause them to interact. A system model is constructed to help understand relationships between elements, forms, processes, and function and to enhance our ability to predict system response to inputs from the environment. This understanding may eventually help the engineer to control system behavior by adjusting inputs to achieve a desire output. A dynamic system is defined as a system whose state depends on the history of system inputs.

Types of Variables Used in Models

Statistical models rely on measured variables. A variable represents any attribute or property of an object whose value can be measured. There are numerous types of variables that arise in experimental, quasi-experimental, and observational research.

In terms of statistical modeling, the generic terms left-hand-side (lhs) and right-hand-side (rhs) variables often can be used synonymously with context specific terminology (referring to the position with respect to the equal sign). Some textbooks, software manuals, and other reference materials refer to *rhs* and *lhs* variables to avoid making distinctions between the types of models being estimated by the analyst.

Dependent and Independent Variables

In predictive studies, variables are often referred as “predictors” and “criteria” variables. In developing statistical models for predictions, the right-hand-side variables are called “predictors”, and the left-hand side variables are called “criteria”, or “response” variables.

In explanatory (or causal) studies, variables are often classified as “independent” or “dependent.” The independent variables (right-hand-side variables) are the presumed cause, and a dependent variable is the presumed effect.

Of course, the analyst should always recognize that an independent variable in a statistical model by definition does not cause the dependent variable to vary. Causality is determined by factors other than statistics, and relationships reflected in a statistical model may merely be associative.

Manipulated Variables

Manipulated variables, often called treatments, consist of the intended assignment of experimental units to different levels of an independent variable, with the intent to observe the impact of these variations on the level of the dependent variable(s). This is the most commonly applied approach to manipulating variables. In cases where it is difficult to vary the value of a variable, an analyst may wish to hold the specific variable constant, thus eliminating its effect on the dependent variable—called control through elimination. The difficulty, of course, is that in real-life the eliminated variable may always vary, and so its

effect may be sought. Control by elimination is typically a last-resort to experimental control.

Randomization Variables

Control through random assignment attempts to account for the effects of the countless extraneous variables that may confound the relationships between independent and dependent variables. When controlling through random assignment, the researcher randomizes the effect of extraneous variables across treatment groups, such that any omitted variable effects are incorporated into the random error term.

Exogenous and Endogenous Variables

An exogenous variable refers to a variable whose variability is determined by influences (variables) outside of the statistical model. An assumption of statistical modeling is that explanatory variables are exogenous.

In contrast, an endogenous variable is a variable whose variability is influenced by variables within the statistical model. In other words, an independent variable whose value is influenced by the dependent variable is endogenous. In many cases endogenous variables are problematic, because it results in violations of one or more standard modeling assumptions.

Planning example. In a study of the relationship between land-use density and other independent variables on transit ridership, the analyst notes that transit ridership might be endogenous. That is, the presence of a transit station is largely responsible for high-density zoning and development, thus transit ridership has a direct or indirect effect on land-use density. This is problematic, because the 'causal' relation is assumed to 'flow' from independent variables to the dependent variables.

Latent, Proxy, and Surrogate Variables

Latent variables are variables of interest that cannot be measured directly, and so instead a proxy or surrogate variable is employed. A proxy variable may be measured precisely and accurately, but measures the latent variable with some degree of error. That is, a proxy variable is equal to the latent variable and an error term. As a result, when proxy variables are used in statistical models, an additional error term is being implicitly introduced into the model. If this additional error term is not modeled explicitly or removed from the model, then potential bias is brought into the model. For instance, the variable "Intelligence" cannot be measured directly, but instead is measured with proxy variables such as IQ score, SAT score, GRE score, and GPA.

Indicator or Dummy Variables

Indicator or dummy variables are used to enable the modeling of marginal effects of discrete variables. In other words, an indicator variable is a transformation of a discrete variable such that the effect of a change in state of a discrete variable can be represented and estimated in a statistical model.

Scales of Variable Measurement

Variables are measurement using an instrument, device, or computer. The scale of the variable measured drastically affects the type of analytical techniques that can be used on the data, and what conclusions can be drawn from the data. There are four scales of measurement, nominal, ordinal, interval, and ratio. The least amount of information is contained in nominal scale data, while the most amount of information can be obtained from ratio scale data.

Nominal

Nominal scales assign numbers as labels to identify objects or classes of objects. The assigned numbers carry no additional meaning except as identifiers. For example, the use of ID codes A, N and P to represent aggressive, normal, and passive drivers is a nominal scale variable. Note that the order has no meaning here, and the difference between identifiers is meaningless. In practice it is often useful to assign numbers instead of letters to represent nominal scale variables, but the numbers should not be treated as ordinal, interval, or ratio scale variables.

Ordinal

Ordinal scales build upon nominal scales by assigning numbers to objects to reflect a rank ordering on an attribute in question. For example, assigning ID codes 1, 2 and 3 to represent a persons response to a question regarding use rate: 1 = use often; 2 = use sometimes; 3 = never use. Although order does matter in these variables (unlike nominal scale variables), the difference between responses is not consistent across the scale or across individuals who respond to the question.

Interval

Interval scales build upon ordinal scale variables. In an interval scale, numbers are assigned to objects such that the differences (but not ratios) between the numbers can be meaningfully interpreted. Temperature (in Celsius or Fahrenheit) represents an interval scale variable, since the difference between measurements is the same anywhere along the scale, and is consistent across measurements. Ratios of interval scale variables have limited meaning because there is not an absolute zero for interval scale variables. The temperature scale in Kelvin, in contrast, is a ratio scale variable because its zero value is absolute zero, i.e. nothing can be measured at a lower temperature than 0 degrees Kelvin. Time is an example of variable measured on the interval scale.

Ratio

Ratio scales have all the attributes of interval scale variables and one additional attribute: ratio scales include an absolute “zero” point. For example, traffic density (measured in vehicles per kilometer) represents a ratio scale. The density of a link is defined as zero when there are no vehicles in a link. Other ratio scale variables include number of vehicles in a queue, height of a person, distance traveled, accident rate, etc.

Hypothesis Testing and Its Role in Statistical Model Development

Hypothesis tests are statistical constructs (tools) used to ask and answer questions about engineering phenomena. The usefulness of hypothesis tests for engineers depends upon sound and informed application of them. Hypothesis tests are easily misused and misinterpreted by engineers and scientists. The potential consequences of improperly applied and/or interpreted hypothesis tests often can be quite serious. For these important reasons, the two concepts are elaborated carefully.

Engineering studies should begin with a set of specific, testable research hypotheses that can be tested using the data. Some examples include:

1) Whether crash occurrence at a particular site tends to support claims that it is a 'high risk' location.

2) Whether an adjustment to the yellow-time at an intersection increases the number of run-through-red violations at the intersection.

3) Whether traffic-calming measures installed at a particular site reduce traffic speeds.

4) Whether alternative route guidance information provided via variable message signs successfully diverted motorists (or not).

Hypothesis tests are typically used to assess the evidence on whether a difference in a measurable quantity (e.g. crashes, speeds, travel times, etc.) on two or more groups (e.g. motorists with and without travel information, locations with and without enhanced police enforcement, sites with and without countermeasures, etc.) is likely to have arisen by random chance alone. Statistical distributions are employed by the engineer in the assessment process to estimate the probability that the data occurred, given a prior assumption about what 'should have' occurred. When the observed results (data) are extremely unlikely to have occurred by chance, the engineer concludes that the difference between the two groups (e.g. roadway re-surfacing, enhanced police enforcement, increased speed limits, etc.) provides a better explanation for the observed differences.

Traffic example. Suppose an engineer is interested in quantifying the impact of roadway re-surfacing on free-flow speeds on a particular section of rural road. She collects representative speed data in periods both before and after the re-surfacing project. As a simple test, the engineer decides to compare the mean speeds between the before and after periods. In statistical terminology, her charge is to assess whether the observed difference in mean speeds between the periods is likely (or not) to be explained purely by the natural sampling variability in the mean speeds for the two groups. The engineering question can be phrased as: does the natural variability in sampling alone explain the observed difference in mean speeds, or is the re-surfacing project responsible for the difference? In this example, the engineer will compute the probability that observed speeds increased (or decreased) from the before to after period under the assumption that no difference should have been observed. If the observed data are extremely under this assumption, then the engineer concludes that re-surfacing was instrumental in bringing about the change. Conversely, if the observed difference in speeds is not all that unusual, and consistent with a difference that could have arisen by random chance, then it is difficult to attribute any observed differences to the effect of re-surfacing.

Hypothesis tests provide the engineer a way in which to feel confident with a decision resulting from the analysis of data. Hypothesis tests are decision-making tools.

The formal uses and development of hypothesis tests is now provided. The application of statistical distributions in the conduct of hypothesis testing consists of five basic steps.

Step 1: Generate the Null and Alternative Statistical Hypotheses from the Engineering Project under Investigation

To ask and answer questions about engineering phenomena the engineer must pose two competing statistical hypotheses, a working hypothesis (the hypothesis to be nullified, sometimes referred to as the null hypothesis), and an alternative hypothesis. Generally, the engineer intends to demonstrate that the data are supported by the alternative hypothesis.

Traffic Example. Suppose an engineer has off-ramp use data for a one month before and after installation of a variable message sign with route-diversion information. By comparing downstream off-ramp use data during the periods before and after the provision of alternative route guidance information, the engineer hopes to quantify the effect the variable message sign had on route diversion. The engineer therefore sets up two hypotheses:

Working Hypothesis (H_0): There has not been an increase in downstream mean off-ramp traffic volumes during provision of alternative route information on the variable message sign (an increase is likely purely by probabilistic chance).

Alternative Hypothesis (H_A): There has been an increase in downstream mean off-ramp traffic volumes during provision of alternative route information on the variable message sign (an increase is likely purely by probabilistic chance).

Of course, the engineer must ensure that other changes in the after period have not taken place, such as seasonal changes in volumes, special events, etc. Finally, the engineer should carefully consider whether the effectiveness of the sign would remain constant over time.

The choice of the null and alternative hypotheses is important, and should be conducted with careful thought and consideration of the study objectives. Recall from Volume I, Chapter 2 that the null hypothesis can be the hypothesis of no effect, often called the nil hypothesis. Since the test statistic will be constructed to test the feasibility of the working hypothesis, the choice of the null hypothesis is crucial. The analyst must consider the practical consequences of making type I and type II errors (see step 3) prior to selecting the null and alternative hypotheses. Since level of confidence and power associated with type I and type II errors respectively may not be equivalent, the choice of the working and alternative hypothesis should be chosen as to maximize the usefulness and practicality of the statistical test results.

The testing of statistical hypotheses is probabilistic and not deterministic. Thus, using knowledge of statistical distributions, the engineer will make judgements as to how likely an observed event would have been given that random chance alone were acting (the working hypothesis). If random chance alone would not likely produce such data, then the engineer prefers the alternative hypothesis explanation of the data.

Step 2: Select Test Statistic and Identify Assumptions of Statistical Distributions

The engineer must decide whether the hypothesis test should be employed to compare the means, variances, or shapes of statistical distributions. Hypothesis tests can be constructed to test almost any aspect of the data in which an engineer is interested.

The type of statistical test selected depends on the engineering questions being asked. For instance, if an engineering intervention is expected to change the mean of a distribution (e.g. advanced warning information expected to increase on-ramp use as in previous example), then a test of means is appropriate. If an intervention is expected to change the spread or variability of data (e.g. an intervention that causes peak-period spreading of traffic volumes), but not necessarily the mean, then the engineer should conduct a test to compare variances. Finally, if an intervention (or program, project, etc.) is expected to change the entire shape of a distribution (e.g. changing both the mean and variance), then a test comparing distributions is most appropriate. Regardless of the type of hypothesis test, the engineer will assess the empirical evidence supporting (or refuting) that an observed difference is likely to have occurred merely by chance fluctuations.

Step 3: Formulate the Decision Rule and Set Alpha and Beta.

The decision to be made by the engineer is which of the two hypotheses, the working or an alternative, provides a better description or explanation of the observed data. Empirical evidence against a chance explanation (the working hypothesis) supports an alternative explanation (not necessarily the one posed by the engineer). Lack of evidence to reject a chance explanation generally raises some doubt as to the validity of an engineering claim, and often leads to further data collection and scrutinizing. The decision rule and selection of alpha and beta should be based on a subjective evaluation of an acceptable level of risk associated with making type I and type II errors.

The significance level, alpha (σ) represents the probability that the analyst would reject the working hypothesis explanation of the data when the working hypothesis is in fact true, and is commonly known as a type I error. A type II error represents the probability that the analyst would not reject the working hypothesis, even though it is false (because the observation resulted under an alternative hypothesis). Table 1 shows the outcomes of statistical hypothesis test results.

Table 1: Outcomes of Statistical Hypothesis Testing and Associated Probabilities

	Reality	
Test Result	*Working Hypothesis is correct*	*Working Hypothesis is Incorrect*
Reject H_0	Error: Type I Probability = α	Correct Decision Probability = 1-β
Do not reject H_0	Correct Decision Probability = 1-α	Error: Type II Probability = β

For several reasons, the probability of making a type II error is often ignored. First, many introductory statistics courses—the type most engineers have taken—do not include detailed discussions about type II errors. Second, there has become an over-emphasis on reporting levels of significance (p-values) in the research literature. Finally, until recently the majority of statistical packages did not provide information on power of statistical tests, so the information is not readily available to practicing engineers. Fortunately, current and newly proposed versions of many statistical packages are offering at least some limited power analysis capabilities. Engineers should always take into account both alpha and beta in the application of the decision rule.

Several important lessons are learned by inspecting the relationship between type I and type II errors. First, the smaller is alpha, the larger is beta for a given test. Thus, if alpha is chosen to be relatively small, the 'cost' is higher probability of a type II error. Second, the larger the effect size, or difference between the working and alternative hypothesis means, the lesser the probability of making a type II error (alpha fixed). In practical terms, bigger effects (differences between the means of the distributions under the working and alternative hypothesis) will be easier to discern than smaller ones. Finally, the variance of the distribution will affect decision-making errors. The larger is the variance, the greater is the probability of making a type II error (alpha fixed).

Traffic example continued: To illustrate the importance of statistical errors associated with hypothesis testing, consider again the advance warning sign example. If a type I error is made, an engineer would conclude that the advance warning sign did avert a significant number of travelers off of the freeway, and perhaps conclude that the sign was indeed effective when in actuality it was not. If implemented, the county/city/state would spend funds on a system that was ineffective. If a type II error were made, the engineer would conclude that the advance warning sign did not significantly impact travelers when in fact it did. In this case, the county/city/state would fail to invest funds into something that was effective.

The determination of which statistical error is least desirable depends on the research question and consequences of the errors. Both error types are undesirable, and thus attention to proper experimental design prior to collection of data will help the engineer minimize the probability of making errors.

Often the probability of making a type I error, alpha, is set at 0.05 or 0.10 (5% and 10% error rates respectively). The selection of alpha often dictates the level of beta for a given hypothesis test. The computation of beta under different hypothesis testing scenarios is beyond the scope of this chapter, and the reader is referred to more complete references on the subject such as Glenberg, (1996).

There is nothing magical about a 5% alpha level. As stated previously, it is the probability of incorrectly rejecting the working hypothesis. The selection of an appropriate alpha level should be based on the consequences of making a type I error. For instance, if human lives are at stake when an error is made (which can occur in accident investigations, medical studies, etc.), then an alpha of 0.01 or 0.005 might by appropriate. On the other hand, if the result merely determines where monies are spent for improvements (e.g. congestion relief, transit level of service improvements, etc.), then perhaps a less stringent alpha is most appropriate. Finally, the consequences of type I and II errors need to considered together, as they are not independent.

Step 4: Randomly Sample from the Population, Check the Statistical Assumptions, and Compute the Test Statistic.

Perhaps surprising to those whom conduct engineering field studies, field data should be collected after steps 1 through 3 have been completed. Thus, before collecting field data, the engineer should have a detailed understanding of the nature and intent of the engineering investigation, and have some specific ideas as to the magnitude and direction of expected effects. In some cases a pilot study may need to be conducted in order to estimate the size of anticipated effects, and to estimate the variability in the data.

Because the engineer will be inferring things about the larger population based on information contained in the sample—the nature of the sample is of utmost importance. Collecting a random sample of data in which to apply statistical techniques is arguably the hardest part of an engineering investigation. To the extent possible, humans should not be relied upon to select objects from a population at random, as humans are not generally good at performing this function due to inherent biases. Instead, a random process must be devised that stands up to outside scrutiny concerning potential sampling biases. For instance, an engineering study designed to estimate mean off-peak travel speeds on rural two-lane highways (in a specific region) might require a random number generator to select both rural highway segments (where all segments have equal chance of being selected) and off-peak travel periods (e.g. one-hour off-peak sampling periods). An extensive discussion on potential sampling biases is beyond the scope of this chapter. There are many textbooks that cover sampling procedures to various extents (including discussions on potential biases), and these should be consulted for detailed discussions. Once a random sampling scheme is devised, it should be adhered to throughout the course of the study. Observations should be carefully recorded for ensuing analysis.

All test statistics have underlying assumptions, which should be thought of as requirements. These requirements, if not met, will impose various consequences on the results of the hypothesis test or the construction of a confidence interval. In the most favorable of circumstances, unmet requirements will yield analysis results with a reasonable approximation to reality. In the worst cases, they will yield analysis results that are misleading. Unfortunately, if distribution requirements go unchecked there is nothing to indicate whether the results are 'best' or 'worst' case. Thus, a remedy is to ensure that the underlying assumptions hold to a convincing level of comfort.

The following list of questions can be asked to test most assumptions and requirements of test statistics:

Are the experimental units continuous or discrete? Often the distinction between discrete and continuous data is blurred. In practice, there are some occasions when discrete data can be assumed continuous without serious quantitative consequences. The treatment of discrete data as continuous is more acceptable as the number of potential discrete outcomes in a defined interval increases. For instance, the number of speeding violations that can occur on a highway segment over a period of time may range from 0 to 1000 or more. Even though these data are technically discrete, the consequence of applying a continuous method or model may not introduce significant error. In most cases, however, there is an equivalent method designed for discrete data that should be used to model and analyze discrete data, and continuous methods should in general be restricted to continuous data.

Are there any sampling requirements? Some distributions require independently sampled data, while others do not. If the observation of one datum affects the probability of observing another, then the data are dependent. Often, sampling from a small population of entities without replacement back into the data 'pool' violates the assumption of independence. Even without replacement, samples drawn from very large populations often can be assumed independent.

Are there restrictions on the assumed distribution's parameters or values? Will the distribution allow negative values? Are there restrictions on the mean and variance or the relation between them? As examples, the Poisson distribution requires that the mean be approximately equal to the variance and the normal distribution requires that the mean is approximately equal to the median.

There are various tools available to help inspect the assumptions of statistical distributions. Graphical plots such as histograms and boxplots are common among many spreadsheet and statistical analysis packages. Graphical plots can be used to inspect the shape of distributions, such as distribution symmetry and presence of outliers. The chi-square, Kolmogorov-Smirnov, and Lilliefors tests may be used to test whether the data or errors terms of a model comply with the assumed distribution. If the chi square test is used it is important to recognize that it formally applies only to discrete data, the bin intervals chosen influence the outcome, and exact methods (Mehta) provide more reliable results, particularly for small sample size.

Test statistics are calculated using the data and are used to draw inferences about the larger population of interest. These should be computed after all statistical assumptions have been met, or remedial measures have been taken to satisfy or circumvent unmet assumptions. In the advance warning sign example, for instance, random measurements on off-ramp usage are used to infer whether the off-ramps are being used more (or less or the same) at all times and conditions when advance warning information is provided to motorists. The test statistic in this example is the difference in the mean traffic volume measured before and after installation of the variable message sign.

Step 5: Apply the Decision Rule and Draw Conclusions

In previous steps the engineer formulated the decision rule (the working and alternative hypotheses) and selected levels of alpha and beta level for making type I and II errors respectively. The engineer collected data, tested statistical assumptions, and computed test statistics in which to learn things about the larger population from the sample data.

What then, does the engineer learn about the population by rejecting (or failing to reject) the working hypothesis? The answer to this question is different from what is often reported, and the interpretation of statistical results is a likely occasion for the analyst to misapply statistics. Correct interpretation of hypothesis tests results is aided by applying the following five rules.

Misconception #1: Alpha is the Most Important Error Rate

For several reasons, the probability of making a type II error is often ignored. First, many introductory statistics courses—the type most engineers have taken—do not include detailed discussions about type II errors. Second, there has become an over-emphasis on reporting levels of significance (p-values) in the research literature. Finally, the majority of statistical packages to date do not provide information on power of statistical tests, so the

information is not readily available to practicing engineers. Fortunately, current and newly proposed versions of many statistical packages are being offered with at least some limited power analysis capabilities. Engineers should take into account both alpha and beta in the application of the decision rule.

Safety example. In many safety studies the impact of countermeasures on human safety are often being assessed. Suppose an engineer is conducting a statistical test of the effect of freeway concrete median barriers on fatal crashes—whether or not they reduce fatalities. A type I error would result in spending money on ineffective median barriers, whereas a type II error would result in failing to spend money on effective median barriers, resulting in a missed opportunity to save lives.

Engineers might set an alpha level associated with a t-statistic to 0.05, only to find that one or more variables in their models, which were thought to be important on theoretical grounds, had p-values associated with t-statistics greater than their accepted "critical p-value"—and so they subsequently remove them from their model. The researcher may want to have these variables included in the model, especially if there is strong theoretical support for such variables, if past research has shown these variables to be important, and if the observed p-values are near the critical p-value. A mechanistic process of removing variables simply on the basis of its p-value should be avoided, and instead the observed p-value, its theoretical importance, and its role in past research should be carefully considered."

The determination of which statistical error is less desirable depends on the research question and consequences of the errors. Because these errors are related—smaller alpha equals larger beta, all else being equal—careful decisions need to be made concerning selection of alpha and beta, and attempts need to be made to quantify beta when appropriate. There are various software packages available for calculating type II error rates, and many textbooks also provide the necessary tools for computing power of statistical tests.

Misconception #2: Hypothesis Test Results are Unconditional

The correct interpretation of a hypothesis test is as follows: If repeated many times (i.e. many samples drawn from the population), the outcome (data) observed by the analyst/engineer and reflected in a computed test statistic (e.g. t-statistic, F-ratio, chi-square, etc.) would occur x percent of the time if the working hypothesis were true. Stated another way, the probability of occurrence is conditional upon the working hypothesis being true. If x is less then alpha, then the working hypothesis is rejected. When the working hypothesis is rejected, the statistical evidence suggests that the working hypothesis is not true, and that some alternative hypothesis provides a better account of the data. What is important to understand (and which is commonly misinterpreted), is that the result does not provide the probability of the working hypothesis being true, nor does it provide evidence that the particular alternative hypothesis is true. In contrast, however, it provides the probability of observing the data if the working hypothesis were true.

Safety example. Suppose a researcher is interested in testing the following question: What is the probability that a motor vehicle crash results in a fatality, given that a motorist traveling on a freeway (at high speeds) is involved in a head-on collision with an immovable object? Upon analyzing the available data, the analyst finds of all fatal crashes, 15% involved motorists traveling on freeways that resulted in a head-on collision with immovable objects. So, is 15% the correct answer to the researcher's original question? From the given information the correct answer is not known, but is likely to be as high as 80%. The difference in interpretation of the two related probability statements is rather fuzzy, as illustrated by the two apparently similar but very different probabilistic statements:
The probability that a crash is fatal, given a head-on collision with an immovable object on a freeway » 80%.
The probability that a crash is a head-on collision with an immovable object on a freeway, given than that the crash is fatal is approximately 15%.

However unrealistic (and morbid), this example, it illustrates the common interpretive error made with hypothesis tests. The probability (alpha) associated with an hypothesis test does not provide the probability of greatest interest—the probability that the null hypothesis is true given the data, but instead provides an interesting and related probability—the probability of obtaining the data given that the null hypothesis is true. These apparently similar probabilistic statements may reflect markedly different probabilities in some cases, as illustrated in the example.

Misconception #3: Hypothesis Test Conclusions are Correct

Conducting hypothesis tests comes with errors, pre-determined by the engineer. Type I errors (rejecting the working hypothesis when it is true) will be committed for approximately alpha percent of tests where the working hypothesis is rejected in the long run. Similarly, type II errors (failing to reject the working hypothesis when it is false) will be committed for approximately beta percent of the tests when the working hypothesis is not rejected in the long run. It is not knowable whether the analyst made an error on any particular test, only the long-term likelihood of the error over multiple tests.

Consider the following. If, over the course of a study, series of studies, year, etc., a researcher were to conduct 300 independent hypothesis tests using $\alpha = 0.05$ and $\beta = 0.10$, and there were 100 rejections of the working hypothesis, then there should be (100)(0.05) = 5 type I errors and (200)(0.10) = 20 type II errors—a total of 25 errors in the course of testing hypotheses. Thus, over this given course of study the researcher would have made five false claims of 'success', and would have incorrectly concluded 'failure' on 20 other occasions. Unfortunately, there is insufficient information for the engineer to know which of the conclusions were correct, and which were incorrect. Interestingly, this is one of the arguments against data snooping and data mining. If a researcher searches long and hard enough through statistical testing (i.e. modeling, testing means, etc.), then he is bound to make type I errors and claim "success" eventually.

What becomes evident is that by conducting increasing numbers of hypothesis tests the chance of obtaining an incorrect rejection of the working hypothesis becomes increasingly likely. If the careless engineer discards (or neglects to report) the failures—the inability to reject the working hypotheses—then the results will be misleading. To combat this problem, the engineer can reduce the error rates by dividing by the number of tests to be conducted in a particular study. For instance, in a study where 200 tests are to be conducted the engineer could use alpha = .05/200 = 0.00025 on each individual test, which would result in approximately 200(0.00025) = 0.05 type I errors. Of course, one must balance out the effect this would have on statistical power, and would only be able to

implement such corrections when sample sizes were significantly large, effect-sizes were large, or variances were small.

Traffic example continued. Assume off-ramp traffic volumes were collected during 100 different periods. Instead of pooling the data, the engineer decides to conduct 100 different hypothesis tests for difference in means, each with a 5% alpha and 35% beta. He finds that of the 100 tests, 80 showed rejection of the working hypothesis (he rejects the "no difference in mean volumes" hypothesis). Since there were 80 rejections of H_0, there were approximately (80)(.05) = 4 incorrectly rejected hypotheses (on average). However, there were also 20 working hypothesis that could not be rejected, and so there should be about (20)(0.35) = 7 type II errors. Thus, in all likelihood, there were 80 – 4 + 7 = 83 occasions when the variable message sign successfully diverted travelers. It is impossible to determine specifically which observation periods were successful and which were not from the data.

Misconception #4: Statistical Significance (Not Effect Size) Determines Practical Significance

Unfortunately, it is all too easy to focus too heavily on p-values, the probability of committing a type I error provided in most statistical package outputs, instead of effect size. The difference between alpha and *p* values is subtle. Alpha is the probability of committing a type I error, and is a constant selected by the engineer before conducting a hypothesis test. It is a subjective value that is intended to delineate chance from non-chance outcomes. P-values are provided in most statistical package output. They are specific alpha levels associated with particular test results, where the alpha associated with an individual test result is calculated exactly. Often, a large number of highly significant statistical tests can mislead an engineer into claiming early success. In actuality, statistical significance is a necessary but not sufficient condition for ensuring practical significance.

Effect size is the magnitude of a difference being tested (e.g. the difference in traffic volume means), the slope coefficient in a linear model (e.g. the rate of change of trips with respect to household income), and the difference between observed and expected cell frequencies in a contingency table analysis, to name a few.

Traffic example continued. Suppose that in testing the effectiveness of the variable message sign the engineer claims success by showing a 2% increase in off-ramp usage during peak travel times to a 99.99% level of significance—a p-value of 0.01. Assuming that the three-lane facility carries 2000 vehicles per lane per hour (vplph) during peak times, and the one-lane off-ramp carries 1000 vplph, this amounts to in increase of (1000)(0.02)=200 vehicles per hour on the off-ramp, or a diversion of about 66 vplph on the mainline--a diversion of about 3% of all mainline vehicles. Whether a 3% diversion of mainline vehicles has practical significance is dependent upon the specific location. However, at a location where a known high proportion of motorists could use the alternative route information (which is presumably a good candidate location for sign placement), a 3% diversion could be considered a failure and not a success. Thus, successfully demonstrating statistical significance by no means ensures practical significance.

The need to assess effect-size is perhaps the biggest argument in favor of using confidence intervals instead of hypothesis tests. Confidence intervals are used to quantify the range of effect size, so analysts can assess the practical significance of the statistical

findings. Confidence intervals are a natural extension of hypothesis testing and are extremely useful to aid in the interpretation and presentation of research findings.

Misconception #5: Non-Significant Hypothesis Test Results are Unimportant

Many engineering researchers and analysts have the impression that the inability to reject the working hypothesis is an undesirable result. Although understandable, this perception is ill conceived. The hypothesis test gives us an objective account of the likelihood of observing data given certain assumptions or pre-conceptions about the process under study. No matter the result of a particular hypothesis test, the engineer departs with more knowledge than she had before the test. Failure to reject the working hypothesis merely suggests that one of several possible explanations provides an account of the results:

1) the expected effect, relationship, or difference did not manifest itself in these data;

2) a type II error occurred;

3) the sample size was too small to discern an effect;

4) the effect-size was too small to detect;

5) the inherent variability in the data is too large to discern an effect; or

6) there is no effect.

A failure to reject a working hypothesis leaves the engineer with a number of possible responses. First is to believe the result as truth and report the findings—this might be the appropriate response when findings are consistent with expectations, past theory and research, and no fatal flaws in the research occurred—and probably is the most appropriate response in most cases. A second possible response is to deny the results as truth and accept one of the latter four explanations for the results. A type II may have occurred. However, disaggregation of the data to single time periods might provide evidence against this explanation, as one would expect to be able to reject at least some hypotheses if a type II error occurred on the aggregate data. If hypothesis tests were conducted for each of the 100 days of off-ramp data, an engineer might rightly conclude that a large number of type II errors are likely. In this case, explanations 3, 4, and 5 are likely candidates to explain the consistent failings to reject the working hypothesis.

Fortunately, the engineer can minimize the impact of problems related to sample size, variance, and effect size by careful experimental design and data collection. Pretests can be used to sample from the population and obtain initial estimates of both effect size and sample variance. These results can then be used to design a larger engineering study to minimize the likelihood of rejecting the working hypothesis when it is indeed false. Under these well-planned engineering studies, the inability to reject working hypotheses is more likely to reflect truth rather than deficiencies in experimental or observational design.

Once an outcome of a well-planned engineering study is determined, either favoring or rejecting working hypotheses, valuable information is obtained from the study. In either case, the statistical results obtained from an engineering investigation provide valuable insight and information for engineers to forge ahead and make engineering improvements. It is of critical importance for engineers to report the results of research, despite the statistical outcome. Failure to report ‘failures’ may result in other researchers spending

money to discover the same result—an inefficient method for conducting research. Only through the collective accumulation of research findings can objective assessments of phenomena under investigation be made.

Sample Size Determination

The determination of adequate sample size is a process that is unique for each study. Thus, no single formula can be provided to compute the correct sample size for a generic study. Sample size requirements for any individual field investigation are affected by the following factors.

First, the entire process of statistical modeling and hypothesis testing is data driven. Thus, statistics estimated from the data, such as the means and variances, will be the primary determinants of sample size for a particular study. Note that data varies across fields of discipline and across studies, as do measurement procedures, thus sample sizes must be estimated for each study that is unique.

The type of statistical procedures to be employed will affect the sample size requirement. Thus, analysis of variance, regression, cross-classification, etc., all have varying requirements underlying their use, and are affected by decisions of the analyst along the way.

The level of confidence, $1-\alpha$ that the engineer would like to employ in hypothesis testing will also affect sample size. Recall that this amounts to setting the alpha level in analyses—or the probability of rejecting the working hypothesis when it is true (a type I error). Similarly, the power, $1-\beta$, required of an analysis will affect sample size requirements. Recall that beta is the probability of failing to reject the working hypothesis when it is false (a type II error).

The method of sampling employed in a study—random, stratified random, cluster, or some other method not mentioned here, will also affect sample size requirements.

Given the above factors affecting sample size requirements, how then, does an engineer or analyst determine appropriate sample sizes needed to conduct a particular study? There is much written on sample size analysis in specific analytic contexts, and for details one should consult textbooks on the subject provided in the references for this chapter. In addition, many software packages are becoming more sophisticated with sample size estimation, and so some packages can be used for this purpose. A general procedure is described here to illustrate process:

1) Identify the formulas relating hypothesis test probabilities with the means, variances, and sample sizes of the variables of interest. This may be the regression equation, and the associated sampling variability of beta's estimated in the statistical model. Conversely, it could be a simple t-test or F-ratio test statistic.

2) Select an appropriate level of significance α and power β for use in the field study.

3) Estimate from preliminary data or past research the variance and means of the relevant variables. This step often requires a pre-study to estimate the variability inherent in the process.

4) Solve the equation identified in step 1 for sample sizes required for conducting hypothesis tests with α and β error rates, observed or estimated variability σ^2, and effect size or sample statistic. The analyst can either solve or iterate back and forth to determine sample sizes sufficient for the study of interest.

Model Validation

A model is a simplified representation of a real world physical process or phenomenon. Thus, a good model is a reasonable approximation of reality. Errors in prediction or explanation arise from differences between the real world and the model. These errors, depending on their source and size, can greatly impact the usefulness and aptness of a model.

Validation describes the process of assessing the appropriateness of the model assumptions, variables, etc. (internal validity), and comparing the model to independent measurements or observations in the real world.

Internal Validity

Internal validity refers to the appropriateness of the proposed relationships between independent and dependent variables and their counterparts in the physical world. Often, mistakes can be made regarding the functional relationships between variables included in the statistical model, and mistakes can be made when selecting which variables to include and exclude from the proposed model.

Inclusion of irrelevant variables occurs when independent variables, in actuality unimportant in the modeled process, have been included in a model. Including irrelevant variables usually inflates model standard errors, with the amount of inefficiency proportional to the correlation between the included variables.

Problems also arise when independent variables thought to be important in the modeled process have been omitted from the model. Often referred to as an omitted variables problem, the result is typically biased parameters of estimates in the statistical model.

Internal validity can be checked by reviewing the reasonableness of model coefficient signs and magnitudes, by using a hold-out data set to determine if the model has been over-fit to the estimation data, by comparing the results to past research results, and by checking the entire model for logical consistency.

External Validity

External validity refers to the ability of the statistical model to be applicable to other populations or over time. In many cases, external validity is the most important attribute of a statistical model, and often the least performed. External validity may be performed by using data collected at another point in time or space to validate the appropriateness of the relationships captured in the statistical model. In model validation, the analyst attempts to determine whether the model parameters are the same for both the estimation and validation data. There are numerous measures for comparing the predictions from an estimated statistical model to external data, such as the PRESS statistic, mean squared error, mean absolute deviation, correlation coefficient, and average prediction bias.

Troubleshooting: Empirical Setting Frequently Asked Questions

What is the difference between models that predict and models that explain?

Statistical models used for prediction can contain many more proxy variables than models used for explanation. In other words, it is not as important to measure the underlying causal variables in a predictive model, since the intent is not primarily to explain the relationships, but instead predict a response. There are usually practical constraints that motivate the selection of proxy variables, such as financial, personnel, and equipment costs associated with capturing the underlying true causal variable.

Because predictive models can be estimated using proxy variables and often satisfy the purpose of the model, more care must be taken to interpret the model results, and statements inferring causality should be avoided. In addition, the external validity of predictive models is more at risk when proxy variables are employed, since proxy variables may work well in one region but not in another, and proxy variables may not be consistently applied over time.

How do I know if the sample data are random?

Unfortunately, the analyst cannot determine whether data have been obtained randomly by inspecting the data after they have been collected. There are ways to ensure that recruitment of experimental units is conducted in a random fashion, however. The first step is to identify accurately and specifically the population in which the analyst is interested. Then a sampling frame, the "window" through which the population is obtained, is identified. A random selection process is then adopted that provides each member of the sampling frame an equal probability of being sampled. Threats to the ability to obtain random samples include a sampling frame that is not representative of the population of interest, a non-random sampling process, and a host of problems with keeping recruited subjects in the sample once identified (refusing to participate, inability to contact, etc.).

How do I know if the sample data are representative of the population?

In many cases, the final sample of data can be compared to a known statistic of the parent population, such as some classification of the population. If the sample differs considerably from the population statistic, then some evidence is provided that suggests the sample is not representative. For instance, census data might be used to verify whether households from a particular region of the country were sampled representatively. Unfortunately, this method can only be used to provide evidence against a random sample, and cannot ensure that a random sample has been obtained, as there may be unknown biases in the sample that cannot be tested.

What are the different scales of measurement and why are they important?

The different scales of variable measurement are nominal, ordinal, interval, and ratio. They are important because they affect the amount of information that can be obtained from statistical models estimated from them.

How are research objectives converted into research hypotheses?

In order to make use of statistical models, research objectives must be converted into testable research hypotheses. Essentially, a research hypothesis is testable if data obtained in the study can be used to test the 'truth' of a particular hypothesis. The most common mistake made is that research hypotheses are not made to be specific enough. For instance, the question, "are sport utility vehicles dangerous?" is not specifically testable, whereas the statement "do sport utility vehicles have a significantly higher crash rate than do passenger cars?" is testable.

How large a sample does the analyst need?

The determination of sample size is a function of many factors, including the variance in the data, the size of the effect that is being investigated, the assigned probabilities of making statistical type I and II errors, and sampling variability. There is not a standard or 'cook book' answer for determining sample size, especially for complex studies with many objectives found commonly in transportation research. Generally, sample size is back-calculated using information obtained from past studies, a pilot study, or other methods for obtaining estimates of effect-size and sample variance. There are many good references on sample size determination that can aid the analyst through the design of a particular study.

What is a dependent variable?

Variables on the left-hand-side of the equal sign are often classified as the dependent variable, in that its' variability depends on values of the independent variables in the model. It is the variable whose outcome the analyst is most interested in, and is usually denoted as Y.

What is an independent variable?

The independent variables, typically denoted as X's, are right-hand-side variables, and are presumed to be associated with or cause a change in the dependent variable, typically denoted Y.

What are the differences between experiments, quasi-experiments, and observational studies?

Experiments are research conducted under specific and controlled conditions to discover, verify, or illustrate a theory, hypothesis, or relationship. Essential elements of an experiment require that at least one variable is manipulated, and random assignment is used to assign experimental units to different levels or categories of the manipulated variable. In a quasi-experiment, the researcher manipulates at least one variable; however, experimental units are not randomly assigned to different levels or categories of the manipulated variable. In an observational study, the analyst might lack the ability to assign levels of the manipulated variable. In other words, the experimenter might not have the ability to control the levels of the manipulated variable, unlike a quasi-experiment. An observational study, furthermore, might have a large number of non-controlled random variables that are thought to affect the dependent variable.

What is a statistical error?

If, as the result of a test statistic computed on sample data, a statistical hypothesis is rejected when it should be accepted, i.e. when it is true, then a type I error has been made. Alpha, or level of significance, is pre selected by the analyst to determine the type I error rate. The level of confidence of a particular test is given by 1 – alpha.

If, in contrast, as the result of a test statistic computed on sample data, a statistical hypothesis is accepted when it is false, i.e. when it should have been rejected, then a type II error has been made. Beta is pre selected by the analyst to determine the type II error rate. The Power of a particular test is given by 1 – beta.

What is alpha and beta, and how do they affect research conclusions?

Alpha is the probability assigned by the analyst that reflects the degree of acceptable risk for rejecting the working hypothesis when in fact the working hypothesis is true. The degree of risk is not interpretable for an individual event or outcome, instead it represents the long-run probability of making a type I error.

Beta is a probability assigned by the analyst that reflects the degree of acceptable risk for accepting the working hypothesis when in fact the working hypothesis is false. Similar to the type I error, the degree of risk is not interpretable for an individual event or outcome; instead it represents the long-run probability of making a type II error.