Chapter 2 Conducting Psychological Research (From unpublished Introductory Psychology Textbook. Feel free to…

Chapter · December 2018

CITATIONS
0

READS
53

1 author:

Shelia Kennison
Oklahoma State University - Stillwater
682 PUBLICATIONS **726** CITATIONS

SEE PROFILE

2 CONDUCTING PSYCHOLOGICAL RESEARCH

Choosing a Research Design

The True Experiment

Quasi Experiments

Correlational Studies

Naturalistic Observations and Case Studies

Collecting Data

Constructing a Sample

Measuring Psychological Variables

Statistical Analysis of Data

Descriptive Statistics

Inferential Statistics

Statistical Significance

Practical Significance

Protecting Human Subjects

History of Research Ethics

Examples of Unethical Research

Research Misconduct

Professional Codes of Ethics

Student Learning Objectives

- Discuss how well the different research methodologies enable researchers to establish a causal relationship between variables.
- Describe how a researcher conducts a true experiment.
- Discuss how psychologists obtain a representative sample from a research population.
- Discuss the importance of validity and reliability in the measurement of variables.
- Describe how researchers evaluate the central tendency of a sample.
- Describe how psychologists measure the variability observed from samples.
- Discuss the extent to which outliers influence large and small samples.
- Discuss the circumstances in which a researcher would make Type I and Type II errors.
- Discuss the significance of the Nuremberg trials and the Belmont Report in the development of contemporary ethical principles for the conduct of research.
- Discuss how the process of informed consent serves to ensure the ethical treatment of human research participants.
- Discuss the code of ethics that psychologists follow. What professional activities are covered in the code?

Psychologists use the scientific method when conducting research. Without a doubt, using the scientific method to investigate psychological phenomena can be challenging. As in other disciplines, researchers in psychology must develop a research question, select an appropriate research design, plan appropriate measurement of the phenomenon of interest, statistically analyze the data, and then interpret the results. Because psychology research routinely involves the testing of human research participants, researcher must ensure that human research participants are treated in an ethical manner at all times. They must also ensure that research data are kept confidential while the study is occurring as well as when the study is over and the data are being analyzed. In this chapter, you will learn more about psychology research is done.

Choosing a Research Design

Every research project begins with an idea or hypothesis about why some human characteristic or behavior might occur. The next important step the research processing is to choose a research design. The researcher must consider the fact that the different types of research designs differ in their **internal validity** or how well the design can establish the cause of the observed outcome. Beginning students of psychology are often surprised to learn that there is only one type of research design that can establish *causation.* If a researcher wants to be 100 percent sure that one variable causes a change in another variable, then there is no doubt about what methodology should be used. Using any other methodology may provide evidence that a relationship exists between the two variables, but one cannot conclude that one variable causes the other.

The True Experiment

What is the one and only methodology that can prove causation? It is the **true experiment**. The true experiment, when conducted in a careful way, achieves the highest level of internal validity. In a true experiment, a researcher manipulates one or more variables in order to determine whether one or more other variables are affected. Any variable that is manipulated in an experiment is called an **independent variable** (IV). Any variable that is considered the "outcome variable" or is expected to be changed by the IV is called a **dependent variable** (DV). Experiments can have one or more IVs and DVs. In order for true experiments to have the highest level of internal validity, the researcher must exert **control** over the research process, so that the only difference among the different versions of the manipulation is the manipulation made by the experimenter. Experiments on animal subjects often achieve high levels of internal validity, because the experimenter can easily control every aspect of the animal's environment, such as the temperature of the room, the layout of the case, the amount of food and water, and the amount of light received each day.

True experiments must also involve **random assignment**. Each member of the research sample should have an equal chance of being assigned to any one of the versions of the IV. Random assignment prevents the possibility that the groups will start out different before the IV manipulation is applied. An easy technique to ensure that participants are randomly assigned to the different groups or conditions is to put all the participants' names in a hat and randomly select participants for the first group, second group, third group, etc. If there are only two experimental conditions, one might determine a particular participant's random assignment with a flip of a coin.

Experiments involving human research participants are more challenging to control or ensure that the other aspect of the situation that is varying is the IV manipulation. Researchers cannot control every aspect of participants' environments as they can in experiments involving animal subjects. Nevertheless, true experiments involving human research participants are routinely conducted. A familiar example is the **randomized clinical trial** (RCT). RCTs are frequently used to test the effectiveness of new procedures, drugs, or non-drug therapies to improve functioning. In a RCT, the effect of a drug or new procedure may be compared to a **placebo**, a pill or procedure that does not contain the experimental substance or procedure. Steps are taken to ensure that the placebo looks similar in all possible ways to the experimental drug or procedure. The use of the placebo is to prevent participants from knowing whether they are receiving the experimental drug, because one may experience physical benefits from merely believing that an experimental drug or procedure has been taken or experienced. This has been called the **placebo effect**. Often, research personnel who monitor the participants' health outcomes also are not permitted to know which participants are receiving the new drug or procedure. This type of design is a **double-blind design.** This type of design prevents the research results from being influenced by **experimenter bias** or the researcher's expectation for a particular type of outcome.

Quasi Experiments

There are circumstances when a researcher cannot test a research hypothesis in a true experiment, such as when the variable of interest cannot be randomly assigned to human research participants. For example, if a researcher is interested in

determining whether men and women perform differently on a particular task, it is not possible for the researcher to flip a coin to determine which participants will be assigned to the *male* condition and which participants will be assigned to the *female* condition. Variables that cannot be randomly assigned to participants are called **subject variables**. Subject variables are frequently used in psychological research. Studies that compare performance for different types of people (e.g., age, smoking status, political affiliation, etc.) are called **quasi experiments**. Quasi experiments cannot achieve the highest levels of internal validity, because it is always possible that the subject variable that was used to create the groups is not the only difference that exists between the groups. The groups might differ in other ways as well. One or more of these other differences between the groups might contribute any difference observed between the groups at the end of the study.

Correlational Studies

Often, a researcher who is interested in understanding more about subject variables will use a **correlational design,** in which no variable is manipulated or compared; rather, the researcher measures two or more variables from a group of individuals in order to learn whether there is any systematic relationship between any of the pairs of variables. Correlational studies generally have low levels of internal validity, because the researcher cannot conclude that the relationship that is observed involves one variable causing a change in the other. For example, one might observe that the number of snow cones sold in a day is related to the number of violent assaults recorded at the police station in the same town. One might first leap to the conclusion that eating a snow cone puts one in the mood to get into a fight. However, one might

then think that it is possible that the experience of fighting leads one to crave a snow cone. Neither of these interpretations may be true. It is also quite plausible that snow cone sales and violent assaults are not directly related at all. A third variable may be related to them both. Researchers refer to this as the **third variable problem**. In this case, temperature could be the third variable. As the temperature rises, some people in the town like to treat themselves to a snow cone, and other people end up having disputes with people and some of those get physical. For studies using a correlational design, one should never conclude that there is a causal relationship between variables.

Despite the low internal validity of correlational studies, media reports of such studies often imply that a causal link between variables exists. In 1996, a study showed that regular coffee drinkers were less likely to commit suicide than those who never drank coffee (Kawachi, Willett, Colditz, Stampfer, & Speizer, 1996). One might assume that the study results provide evidence that there is some substance in coffee that serves to protect one from the state of mind that leads to suicide. However, it is just as plausible that some unidentified variable is related to whether people drink coffee and is also related to people's likelihood of committing suicide. For example, it is possible that the overall health of participants was related to coffee drinking. Many individuals who regularly take medications for various health problems may be advised by their physicians to avoid coffee and other foods containing caffeine. Unfortunately, when listening to media reports of research, one must always be cautious. The journalists who are reporting the results of studies are likely not to have taken a course in introductory psychology and in psychological research methods.

Naturalistic Observations and Case Studies

Although correlational studies cannot provide convincing evidence that one variable causes another, they can be an important first step in a researcher's long-term program of research. If a correlational study provides some evidence that variables are related, then the researcher may conduct a follow-up study, using a different research design. By conducting multiple studies on the same topic, a research may obtain a more detailed picture of the phenomenon. There are two other types of research designs that are routinely used as when researchers are just beginning to study a topic. These research designs are the **naturalistic observation** and **case study**. Both types of designs have very low levels of internal validity. In a **naturalistic observation,** one observes behavior as it occurs naturally in life. The advantage of this type of research is that the researcher can get a picture of behavior as it occurs in real world settings. A disadvantage of this type of research is that there is a lack of control over the setting. As a consequence, the researcher cannot be certain about the actual causes of the observed behaviors. In a **case study**, researchers study a single participant or event in order to gain insight into a phenomenon or process that occurs in others. There is also a lack of control in case studies. A researcher documents that important elements of the person and the situation and attempts to identify possible causes of outcomes; however, no definitive causal conclusion can be drawn.

For some research hypotheses, a researcher may never be able to conduct a study in which a causal relationship between variables can be established. Sometimes, a true experiment can never be done. This is most often true when one is investigating the harmful effects of one or more variables on human research participants. Consider

the numerous studies that have demonstrated that cigarette smoking in humans is correlated with lung cancer rates and rates of breathing-related diseases (Oreskes & Conway, 2010). None of them have *proven* with 100 percent certainty that cigarette smoking causes lung disease in people, technically. This is because, there has never been a true experiment conducted involving human research participants. Why has no true experiment been conducted? Conducting such an experiment would not be ethical. Although, it would be relatively straightforward to recruit a group of research participants for the experiment, offer to pay them well for their time, and then randomly assign half of the sample to smoke a pack a day for a year and randomly assign the other half not to smoke at all for year. After a year, you call them back to the laboratory for a complete physical workup. I hope it is obvious why this type of experiment should never be done. It would be unethical to expose someone to a manipulation that is *expected* to cause them harm.

Constructing a Sample

Regardless of the research design that is used, it is usually the case that the researcher's ultimate goal is to use the study results to make a statement about what is likely to occur in the future other people in other places, particularly for people who are similar to those who were tested in the study. For example, if a researcher finds that taking a new antidepressant causes 45 out of 50 people given the drug to experience an improvement in mood, then it is reasonable to estimate that 90 percent of other people given the drug may also an improvement of mood. The term **external validity** refers to how well a study's results can predict outcomes to other groups. Generally

when conducting a research study, one strives for the study to achieve the highest possible level of external validity.

In order to determine whether the external validity in a study is high or low, one must consider the extent to which those tested in the study are similar to the individuals whose outcomes the researcher wants to predict. The **research population** is the group of individuals who share one or more characteristics of interest to a researcher and whose outcomes the researcher is interested in predicting. The **research sample** is a subset of a research population from which data are collected. Studies that have the highest level of external validity use **representative samples,** which are samples in which the researcher has recruited a small group of people who accurately reflect the characteristics of the entire research population in proper proportion.

Studies that have the lowest level of external validity use **biased samples** or samples that do not accurately reflect the characteristics of the research population. For example, if a researcher is interested in a population of male and female college students and the sample includes only female students, then the sample results might not accurately predict outcomes for all college students. If a researcher is interested in a population that includes adults ranging in age from 18 to 60 and the sample includes only individuals between the ages of 18 and 22, then the results might not generalize to individuals who are over 22. Samples can be biased for many reasons. Whenever a sample is biased, the researcher is likely unable to *fix* it; there is no easy way to transform a biased sample into a representative sample. If researchers report the results of the sample, they must be careful to provide ample detail about the creation

and composition of the sample, the nature of the bias, and how the interpretation of the results is limited by the sample's bias.

The best strategy for obtaining a representative samples by obtaining a **simple random sample**. A simple random sample is achieved when each person in the research population has an equal chance of being selected for the research sample. This can be achieved by placing the names of all the members of the population in a hat, shaking vigorously, and then selecting at random those that will be include in the sample. Unfortunately, constructing a simple random sample is rarely possible in studies involving human research participants, because it is typically impossible to identify all the members of the research population by name. In order to ensure that each member of the research population has an equal chance to be selected for the sample, one would have to throw all their names into the hat. For large populations, not only may one having trouble finding a hat that is large enough, but more importantly, one may have trouble obtaining a list of names of everyone in the research population. If one cannot put all the names in the hat, one cannot ensure that every member of the research population has an equal chance of being selected for the research sample.

In psychology research, the most frequently used type of sample is the **convenience sample,** which involves recruiting and testing participants who meet an eligibility criterion. Participants are recruited from locations convenient to the researcher. Participants may be recruited through flyers posted in public locations, face-to-face appeals, as well as other forms of announcements on television, radio, or the Internet. The participants must possess the characteristics of interest to the researcher, which would be used to define the research population. In convenience

samples, these characteristics serve as the eligibility criteria for recruitment into the study. The convenience sample is the most commonly used sample in research. They are used in even the most rigorous experiments involving human research participants. For example, RCTs that are conducted by the National Institutes of Health utilize convenience samples. Announcements of current trials are listed at www.clinicaltrials.gov (NIH, 2011). Nevertheless, whenever convenience samples are used, there is the possibility of obtaining a biased sample. When researchers use convenience samples, they must be careful to evaluate the representativeness of the sample and discuss the possibility of bias in the interpretation of the results.

There are some research designs whose external validity is always low. For example, in a case study, a researcher may focus on an individual person. Case studies are commonly used to study the functioning of individuals with illnesses, disorders, or injuries. If a researcher can document that a particular therapy or treatment produced a positive outcome for an individual, there is the implication that others with a similar illness, disorder, or injury may also benefit from the therapy or treatment. In order to show that the results of the case study do generalize to other, similar individuals, a researcher may carry out a series of case studies. Many therapies or treatments that start out being tested in case studies ultimately come to be tested in a RCT involving a large sample of participants. In a case study, it is impossible for the researcher to determine whether the results are indicative of what could be observed in others or whether the results reflect idiosyncrasies of the individual.

Measuring Psychological Variables

After a researcher selects the research design and decides how the sample will be constructed, the next step in the research process is to decide the best way to measure the variables of interest. Researchers use the term **operational definition** to refer to the detailed description of how a variable is measured a study. Researchers typically include details of their operational definitions in the methods sections of their research reports. Doing so facilitates the **replication** of research, which is the repeating of research studies for the purposes of determining whether similar results can be obtained. Research results that cannot be replicated may be inaccurate. Only after multiple replications can one be certain that a particular result is accurate and reflects the true state of affairs in the world.. It is often said that *science is self-correcting*, because it involves the multiple replication of important research findings and the continual refining of research explanations. If a result or an interpretation of a result is in error, future replications of the study are likely to improve upon the previous error.

For many psychological variables, there are ways to measure them. For example, when a researcher wants to assess the success of an individual, one might use financial indicators of success, such as one's annual income, or psychological aspects of success, such as one's life satisfaction, or even others. The beginning researcher should consider how the variables of interest have been studied in previous studies. For topics that have been studied many times before, it is likely that researchers have developed some tried and true operational definitions. By using

operational definitions that have been used in prior research, researchers also will find that is it easier to compare their results with those obtained by others.

In most research involving human research participants, at least some of the information collected from participant is reported in either a verbal or written form. Such self-report procedures are referred to as **surveys**. One type of survey is a questionnaire, which is a list of questions to which individuals respond in writing. It is common for questions to be formulated so that participants report a response level using a scale. For example, one might be asked to rate their satisfaction with life on a scale from 1-to-7 with *1* corresponding to *not satisfied at all* and *7* corresponding to *extremely satisfied.* Such questions are called **Likert-type scales** for the psychologist Rensis Likert (Likert, 1932).

In addition to surveys, psychologists use a wide variety of other techniques to measure psychological variables. In some studies, researchers might measure how quickly a participant can respond to a stimulus, such as word or picture displayed on a computer screen. In such studies, a participant may be asked to make a judgment by pressing particular key on a keyboard. The time that the participants take to press the key is recorded. The amount of time that a task takes to perform can be useful information to a researcher who is attempting to understand what steps of processing are occurring when one makes a judgment. Psychology researchers might also measure physiological responses, such as heart rate, skin temperature or moisture as well as electrical activity produced by muscle movements or brain activities.

When choosing how to measure the variables in a study, researchers must be careful to make sure that the measures both valid and reliable. A measure's **validity**

refers to how well the operational definition of a variable accurately captures the concept of interest. For example, imagine that a researcher is interested in measuring the anxiety level of students in a class. The researcher finds a pretty short questionnaire that was developed by researchers at Oxford University in English. The questionnaire asks a series of questions about satisfaction with life, school, family relationships, and various other aspects of daily life. Using the questionnaire as a measure of anxiety would not be valid. Satisfaction and anxiety are different concepts. The **reliability** of a measure refers to the extent to which similar results can be obtained when the measure is carried out multiple times. In some instances, a researcher may rely on human coders to make judgments about events or objects. It is important that that different human coders use the same criteria when providing ratings. **Interrater reliability** refers to how consistent multiple raters are in judging the same events or objects. For example, raters may be asked rate how long an infant looks at an object during a testing session. Each rater would complete their judgments individually. Then, the raters' judgments would be compared. Interrater reliability could be computed by counting the number of times the raters agreed and dividing the number by the number of times that they could have agreed.

Researchers who gather responses from participants or make observations of behaviors must keep in mind that participants sometimes change their responses or behavior because they are aware that they are being studied. The term **social desirability bias** refers to the tendency of human research participants to respond or to behave in a way that they perceive to be consistent with social norms. A person may not want to report that they drink beer every day, starting early in the morning, even if it

is true. They may not want to report that they cheat on their tax returns or dislike their bosses. In other circumstances, participants may guess the purpose of the study and may respond in a way that they perceive would be helpful to the researcher. A person may guess that a taste test is being conducted by a particular company and then provide very high ratings each time they are asked to sample on the company's products. Despite the participants' good intentions, this type of responding prevents the researcher from obtaining an accurate picture of the phenomenon. Researchers try to prevent participants from guessing the purpose of the study. **Demand characteristics** refer to any aspect of the research procedure that might enable the participant to guess the purpose of the study. Researchers may carry out **piloting testing** or practice sessions to run through the research procedures after which they make improvements to the procedures before actually conducting the study.

Statistical Analysis of Data

One may think that most of the hard work in the research project is over after a researcher finds a research hypothesis, selects the research design, decides how the sample will be constructed, operationally defines the variables. Actually, only then is the important work beginning. The next step in the research process is to analyze and to interpret the results. Researchers use **statistics** to organize and to summarize numerical data. Researchers have two types of statistics at their disposal. **Descriptive statistics** involve using only the data from the sample to make statements about the sample itself. The data are not used to make predictions. When researchers do want

to use the results to make predictions, as is the case with most psychological research, they use **inferential statistics.**

Descriptive Statistics

When describing samples, one is typically most interested in the **central tendency** or the score or observation that is most typical of the entire sample. For example, imagine that you have a summer job at the local YMCA. Your boss would like you to collect some data about how many different activities the children and parents in the community would like the YMCA to offer. You devise a short survey in which you ask people to list make a recommendation. You gather your survey responses to show your boss. Your boss asks you "What did you find?" As you begin to describe all the different responses that you received, the boss stops you and says, "What activity do most people want?" The boss is asking you to report the central tendency. There are three common measures of central tendency. The **mode** is the most frequently observed response. The mode is useful for both data measured on a numerical scale as well as data that are not numeric, such as responses to the question "What is your favorite activity at the YMCA?" In fact, the mode is the only measure of central tendency that can be used with non-numerical data. For numerical data, one can also use the **median,** which is the middlemost score or the **mean**, which is the arithmetic average.

When determining the central tendency of a sample of numerical data, it is useful to make a graphical representation of the **distribution** or all of the observations in the dataset. On the horizontal or x-axis, one displays the different types of responses. On

the vertical or y-axis, one displays the frequency or how many responses were observed. Sometimes, the graphs are displayed with bars representing the frequency of each response category. Sometimes, only a smooth curve is used to indicate the frequency of each category. For the bar graph, the highest bar corresponds to the mode. For the smoothed curve graph, the response category that is below the highest part of the curve is the mode. When the mode, median, and mean are the same, then determining the central tendency of the dataset is easy. This occurs when the data follow a perfect **normal distribution or bell curve.** A normal distribution is symmetrical and has an equal number of scores below and above the median, mean, and mode. In contrast, a **skewed distribution** is not symmetrical. There are either more scores above the median, which occurs in a **negatively skewed** distribution, or more scores below the median, which occurs in a **positively skewed** distribution. A distribution is skewed because there are a relatively small number of scores that are unusual in comparison with the rest of the dataset. Such scores are called **outliers.** For skewed distributions, determining the central tendency is less straightforward, as it is usually the case that the mode, median, and mean will differ. The median is recommended as the best measure of central tendency for any skewed distribution, because is affected less by the presence of outliers than is the mean.

When describing a dataset, one also pays attention to the sample's **variability**, which refers to how much the observations in a sample are different from one another. A useful measure of variability is the **range** of a distribution, which refers to the size of the difference between the highest score in the distribution and the lowest score in the distribution. When a distribution has a larger range than another, the distribution has

higher variability. Most often, researchers assess the variability of a distribution using the **standard deviation,** which reflects how spread out the sample is relative to the sample mean. When computing the standard deviation, one measures how far each score is from the mean and then finds that average. Samples with low standard deviations have scores that are more similar to the mean that samples with higher standard deviations.

Researchers who want to describe the relationships among variables in a sample can use descriptive statistics to quantify the relationship that exists between variables. The most commonly used statistic is **Pearson's *r*** (Gravettner & Wallnau, 2009)**.** The statistic captures the degree to two variables share a linear relationship. If two variables are not related in a linear way, the value of *r* will be 0. When the two variables are perfectly related in a linear way, the value of r is either +1.00 or -1.00. When two variables are perfectly correlated and their values increase and decrease together, they have an r of +1.00. When there is a perfect correlation and the values of one variable increases as the values of the other variable decreases, the value of r is -1.00. Correlations with r values close to ±1.00 are stronger than correlations with values close to 0.

When two variables are correlated, it is possible to think of the relationship as indicating how much of the variability observed in one of the two variables is explained by the other variable. Consider the relationship that has been found to exist between happiness and income. Research has shown that the value for *r* = .18 (Hagerty, 2000). One might wonder what percentage of the variability in happiness is explained by income. One can compute this using the *r* value, squaring it, and multiplying by 100.

According to this formula, only 3.2% of happiness is explained by income, leaving 96.8% of happiness level unexplained. The take home point here is that the stronger the correlation, the more one variable can account for variance observed in the other variable. For variables that are strongly correlated, such as those with an $r = +.80$, there is 64% in one variable that can be explained by the other and 35% of the variance unexplained. Variables that have an $r = +.95$, 90% of the variance can be explained, and 10% remains unexplained. Still, one must always be mindful that the relationship cannot be assumed to be causal.

Inferential Statistics

When researchers use data from a research sample to make a prediction about a research population, they use inferential statistics. Work conducted by mathematicians over 300 years ago laid the foundation for the inferential statistics that researchers use today. The mathematicians from the distant and not-so-distant past worked hard to uncover some remarkable regularities that occur when samples are drawn from a population. At the heart of this work is the concept of **probability,** which refers to the likelihood that an event will occur. Probability is most easily explained in terms of random events, such as flipping a coin, rolling dice, or drawing a card from a well-shuffled deck. One can estimate the likelihood of an event by divided the number of possible times that event could be observed by the total number of possible outcomes. So, when a coin is flipped, the probability of a heads is one heads in every two flips or 1/2. The probability of observing a tails is also ½. If a single six-sided die is rolled, there are six possible outcomes (i.e., 1, 2, 3, 4, 5, or 6). So the chance of observing any one of these outcomes on a roll is 1/6. The chance of rolling an even number is 3/6,

because there is there are 3 ways in which that event could be observed. The chance of rolling a 4 or a 5 is 2/6.

This simple rule of probability can also be used to predict what selections are likely when sampling. Imagine that a teacher has a class of 40 students, 15 of them male and 25 of them female. If she places all their names in a hat to draw out one a random, she has a 15/20 chance of drawing out the name of a male student and a 25/20 chance of drawing out the name of a female student. Now imagine that you know a little more about the class, specifically that it contains 10 freshmen, 10 sophomores, 10 juniors and 10 seniors. Of the freshmen, 1 is a male and 9 are female. Of the sophomores, 4 are male and 6 are female. Of the juniors, all are male, and of the seniors, all are female. What is the probability of drawing out a name that is a junior? That would be 10/40 or ¼. What is the probability of drawing out a name that is male? That would still be 15/40? What is the probability of drawing out one name at random that is a freshman male? That would be 1/40.

Relying on the regularity of the laws of probability, statisticians have demonstrated that large samples are always more representative of the population from which they are drawn than are small samples. This fact is referred to as the **law of large numbers.** Small samples have a greater chance of being influenced by outliers than large samples. On each draw from the population, there is a greater chance of typical examples of the population than unusual examples in the population. When one draws from the population many times as happens when the sample is large, there are more opportunities for typical scores to be observed than when one draws just a few

times from the population. Small samples are more vulnerable to outliers have a big impact on the sample characteristics, such as the mean and standard deviation.

Mathematicians have been intrigued by probability and sampling since the 1600s. In early 1700s, the mathematician Abraham de Moivre (1667-1754) discovered that there is a remarkable regularity that occurs when one samples many, many times from a population. When one keeps track of the samples, one observes something quite amazing – the distribution of sample characteristics (i.e., means, median, standard deviations) *always* approximates a normal distribution. This is true even when the population that is being sampled from is not itself a normal distribution (Gravettner & Wallnau, 2009). This discovery is called the **Central Limit Theorem**. Without the central limit theorem, scientists would not have the ability that they have to use the results from their research samples to make predictions about research populations. Research into the regularity of samples further found that the sample size mattered. For very large samples, such as samples that are infinitely large, then the distribution would be a perfect normal distribution. For smaller sample sizes, the distribution was found to deviate systematically from a perfect normal distribution as the sample size decreased. These demonstrations allowed statisticians in the 20th century to develop the inferential statistics that we have today, such as the t-test and analysis of variance (ANOVA).

Prior to the discovery of the central limit theory, there was another discovery whose existence makes possible modern day inferential statistics. This discovery is that normal distributions are special. Most measurements taken of the physical word, including human characteristics and measurements of human behavior, tend to follow a

normal distribution. Furthermore, it happens to be the case that all normal distributions can be described by the **68-95-99.7 rule.** In all normal distributions, one will observe that 68% of the entire dataset will be within one standard deviation of the mean of the distribution, which means that 34% of the dataset will lie between the mean and one standard deviation above the mean and 34% of the dataset will lie between the mean and one standard deviation below the mean. Further, 95% of the entire dataset will be within two standard deviations of the mean of the distribution, which means that 47.5% of the dataset will lie between the mean and two standard deviations above the mean and 47.5% of the dataset will lie between the mean and two standard deviations below the mean. Go out three standard deviations away from the mean and you will always find 99.7% of the entire distribution.

Consider men and women's heights in the United States. Research has shown that measurements of height follow a normal distribution. The average height for U.S. men is 69.5 inches and the standard deviation is 3 inches, and the average height for U.S. women is 64 inches with a standard deviation of 2.5 inches (National Health Statistics Reports, 2008). Having information about the mean and standard deviation allows us to extrapolate that 68% of men in the U.S. are between 65.5 and 72.5 inches tall, 95% are between 63.5 and 75.5 inches tall, and 99.7% are between 60.5 and 78.5 inches tall. Sixty-eight percent of women in the U.S. are between 61.5 and 66.5 inches tall, 95% are between 59 and 69 inches tall, and 99.7% are between 56.5 and 71.5 inches tall. With the foundational knowledge of the central limit theory and the regularity of the normal distribution, modern day researchers can collect a reasonably large sample from a population whose characteristics are unknown and be about 95% sure

that their sample mean is within two standard deviations of the true mean of the population. This range of values that contains the true population mean is a **confidence interval.** For a confidence level of 68%, the range would be the sample mean plus or minus one standard deviation. For a confidence level of 99.7%, the range would be the sample mean plus or minus three standard deviations.

Statistical Significance

A phrase that might already be familiar is the phrase "Is it significant?" This is what researchers really want to know. For example, in an experiment with an experimental condition and a placebo control condition, the researcher wants to know whether any observed difference between the two groups is significant. The mathematics and statistical knowledge that enables researchers to estimate unknown population values from values observed in a sample also enables researchers to carry out a procedure known as **significance testing** and to determine whether an observed result is *significant* or is likely to reflect a true difference that could be observed again and again in future experiments.

The significant testing procedure usually strikes the beginning student as rather unintuitive. The researcher always begins by formulating a **null hypothesis,** which states that the there is no difference of the type that the researcher believes might occur. In our example, there is no difference for the IV measurement for those who took the experimental drug and those who took the placebo. Then the researcher states the **alternative hypothesis,** which states that there is a difference of the type that the researcher believes might occur. There is a difference between the experimental and

placebo groups. Researchers must approach significance testing in this way because the statistical tools available do not enable researchers to determine whether the alternative hypothesis is true; rather, researchers can only determine whether it is likely or unlikely for the null hypothesis to be true. Only when the probability is very low that the null hypothesis is true can one decide to reject the null hypothesis in favor of the alternative hypothesis.

Researchers can choose the level of certainty that they use when evaluating the null hypothesis. The term **alpha level** is used to refer this criterion. The most commonly used alpha level is .05, which means that in order for the null hypothesis to be rejected, the observed sample must be among the most extreme 5% of samples that are possible in the population when no manipulation has been applied or no effect would be expected. Alpha levels of .01 and .001 are also routinely used. For these alpha levels, in order for the null hypothesis to be rejected, the observed sample must be so unusual that they are among the most extreme 1% or .1% samples in the natural variation of that population when no manipulation has been applied or no effect would be expected. For an observed sample, a researcher can estimate the likelihood of that such a sample occurs in the population in which no manipulation has been applied or no effect is expected. This estimate likelihood is called the probability value or **p-value**. Consequently, when p-values are lower than .05, .01, or .001, the null hypothesis can be rejected.

Whether the researcher rejects the null hypothesis or accepts it, the researcher may be incorrect. Two types of errors are routinely possible when one engages in hypothesis testing – Type I errors and Type II errors. When a **Type I error** is made, the

researcher concludes that a real difference was observed in data, but the difference is does not exist. The result would be unlikely to be observed in a replication. When a **Type II error** is made, the researcher concludes that no difference was observed in the data, but the difference does exist. The likelihood of making a Type I error is directly related the alpha level. Type I errors are more likely to be discovered in future research than Type II errors, because research journals are more likely to publish significant results. When a Type II error occurs, a non significant result is observed. The likelihood of a Type II error is directly related to the statistical **power** of a study, which refers the study's ability to detect a difference or effect, assuming that it exists. Studies that have been carried out with careful methods and with adequate numbers of participants have higher statistical power than studies carried out carelessly with small samples. Studies with adequate statistical power are less likely to result in Type II errors than those with inadequate statistical power. Researchers rely on power calculations to determine the number of participants that are needed in a sample to observe an effect.

Practical Significance

Researchers are concerned not only with the statistical significance of their data, but also with the practical significance of their data. The researcher should consider the extent to which the result that was observed in a sample is likely to translate into a meaning difference in everyday life. When evaluating the practical significance of any research result, one wants to know what does the observed difference mean to me in my life or in my work. Unfortunately, reports of research studies may not do a very

good job of explaining the practical implications of the results. We all must keep in mind that a statistically significant difference in a research study does not always translate into an important difference that one can experience in everyday life. The term **ecological validity** is used to describe the extent to which a research finding reflects a real life situation.

A numerical way for researchers to discuss the practical implications of their results is to discuss the extent to which the statistical difference that was observed involves a large, medium or small **effect size**. Consider an example of a study that tested the effectiveness of a drug to reduce anxiety. It is possible for a researcher to observe a statistically significant reduction in anxiety due to the drug when the reduction is relatively small. Consumers may want to know if the reduction in anxiety that is expected from the drug worth the possibly high cost. Researchers calculate the size of their effects using a statistic known as Cohen's *d* (Cohen, 1992). Cohen's *d* is equal to the observed difference divided by the standard deviation. Large effect sizes have values for *d* of .80 and greater. Medium effect sizes have values for *d* around .50. Small effect sizes have values for *d* of .20 and smaller. In some cases, researchers may investigate the effect sizes observed across similar studies in order to investigate the range of circumstances in which the effect is observed and the typical effect size. Such an investigation is called a **meta-analysis**.

Frequently, studies in which different groups of people are compared find statistically significant differences in the performances of the groups. One fact that seems to get lost in the reporting of the results is that when a difference is observed between the average performances of two groups of people, it is *not* the case that every

member of one group will always out score every member of the other group. Consider the differences observed between men and women's heights. When one examines that tails of the distributions, one finds that among the shortest people, there are more women than men and among the tallest people, there are more men than women. However, there are many women who are as tall as or taller than the average man. Conversely, there are many men who are as short as or shorter than the average woman. When one knows only there is a difference between two groups, it is not the whole story.

Consider the differences that have been found between the math performance for men and women (Halpern, 2000). On standardized tests, the average performance has been found to be higher for men than for men. In the case of math performance and the difference between men and women, the distributions of performance are extremely overlapping. There are many women who perform equal to or above the average performance of men. Conversely, there are many men who perform equal to or below the average performance of women. Further, when one examines that tails of the distributions, one finds the lower tails identical, suggesting that there are comparable percentages of men and women who perform very poorly in math. The upper tails differ, indicating that the extremely small percentage of the population of men who are math geniuses is somewhat larger than the percentage of the population of women who are math geniuses.

When evaluating the practical significance of group differences, one must always consider the entire distribution of scores, rather than just the group means. When one does this, one finds that the variability that occurring within each group is far greater

than the variability that exists between the two groups. Stated in another way, in math performance, women are far more different from each other and men are far more different from each other than the two groups are. Because within group differences are *always* bigger than between group differences, one should be extremely cautious about predicting an individual's likely performance, based on the existence of a performance difference involving groups. If one is parenting a son or daughter, one cannot ever be sure whether the child will be similar to most other boys or girls on the measured dimension or whether the child will be an outlier for his or her group.

Our discussion of the practical implications of research reporting group differences shines a light on a fact that easily goes unnoticed in most discussions of psychological research. It is important to remember that there is no methodology in existence that can predict with accuracy any outcome for any single individual. Statistical procedures provide a useful framework for researchers to make predictions about outcomes for large groups of individuals, particularly when using data obtained from large samples. For example, imagine that researchers have investigated how well new drug reduces improves memory. It was found that a sample of 100 people who received the drug remembered 20 percent more on a memory test than a sample of 100 people who received a placebo. The researcher predicts that on average, those taking the drug in the future will remember 20 percent more than they do normally when not taking the drug. However, it will be the case that some may experience smaller improvements in memory and some may experience larger improvements in memory. The average improvement is expected to be 20%. Someone who takes the drug and

does not experience any improvement is possible. The event would be unexpected, but an example of an outlier.

Protecting Human Research Participants

An essential part of the research process is ensuring that the research is carried out in an ethical manner. In the United States, research funded by the government is regulated by Part 46 of Title 45 Code of Federal Regulations or 45 CFR 46. This regulation is referred to as the **Common Rule.** One of the requirements specified by the common rule is that institutions receiving federal research grants involving human research participants establish **Institutional Review Boards** (IRBs) that review all research on ethical grounds prior to the research's implementation. IRBs are composed of scientists and non-scientists as well as members of the community in which the institution is located. The members of the IRB are charged with the task of ensuring that research participants are not harmed through their participation in research and that they are ethically treated at every stage of the research process.

History of Research Ethics

The modern procedures involved in the ethical conduct of research have evolved over the last half century. Following World War II, the public became aware of the need for governments to ensure that scientific research is ethically conducted. In 1945 and 1946, the Nuremburg trials documented the Nazi's experimentation on prisoners (Tusa & Tusa, 2010). After the trials, the tribunal produced the **Nuremburg Code,** which provides guidelines for the ethical treatment of research participants. These guidelines are displayed in Table 2.1. Among the most infamous of the Nazi doctors was Josef Mengele (1911-1979), also known as the angel of death. He performed numerous

Table 2.1 Nuremberg Code (1947)

1. The voluntary consent of the human subject is absolutely essential.

2. The experiment should be such as to yield fruitful results for the good of society, unprocurable by other methods or means of study, and not random and unnecessary in nature.

3. The experiment should be so designed and based on the results of animal experimentation and a knowledge of the natural history of the disease or other problem under study that the anticipated results will justify the performance of the experiment.

4. The experiment should be so conducted as to avoid all unnecessary physical and mental suffering and injury.

5. No experiment should be conducted where there is an a priori reason to believe that death or disabling injury will occur; except, perhaps, in those experiments where the experimental physicians also serve as subjects.

6. The degree of risk to be taken should never exceed that determined by the humanitarian importance of the problem to be solved by the experiment.

7. Proper preparations should be made and adequate facilities provided to protect the experimental subject against even remote possibilities of injury, disability, or death.

8. The experiment should be conducted only by scientifically qualified persons. The highest degree of skill and care should be required through all stages of the experiment of those who conduct or engage in the experiment.

9. During the course of the experiment the human subject should be at liberty to bring the experiment to an end if he has reached the physical or mental state where continuation of the experiment seems to him to be impossible.

10. During the course of the experiment the scientist in charge must be prepared to terminate the experiment at any stage, if he has probably cause to believe, in the exercise of the good faith, superior skill and careful judgment required of him that a continuation of the experiment is likely to result in injury, disability, or death to the experimental subject.

horrific surgeries and procedures on prisoners at Auschwitz. He focused particularly on twins, pregnant women, and individuals with physical deformities (Ware & Posner, 1986). The dissection of living people without anesthesia was common. Mengele was not among the 23 Nazi doctors who were placed on trial at Nuremberg. He died in Paraguay in 1979. His identity was not confirmed by DNA analysis until 1992.

In 1964, the World Medical Association published the Declaration of Helsinki, which was an ethics code for medical research. In 1974, the Congress of the United States created a National Commission for the Protection of Human Subjects in Biomedical and Behavioral Research. The committee met in Eskridge, Maryland at the Belmont Conference Center, and their report was called the **Belmont Report.** The report set forth three core principles governing the ethical treatment of human research participants: a) **respect for persons**; b) **beneficence**; and c) **justice**. Each of these principles is today reflected in the procedures that researchers are required to follow when conducting research.

Respect for persons refers to "protecting the autonomy of all people and treating them with courtesy and respect." Participants should be volunteers. Their participation should not be forced or coerced in any way. In order for participants to exercise their right to volunteer, there must be a process of **informed consent**, during which the participant is made aware of the nature of the research and the procedures that will be performed in the research study. In research involving research participants who are younger than 18 years of age, parental consent must be obtained. When parental consent is obtained, the researcher must then formally invite the child to participate and provide information about the nature of the study in a process that is referred to as

assent. All research participants, whether they are adults or children, should be allowed to withdraw from the research study at anytime without penalty. In studies involving deception, researchers must provide **debriefing**, during which participants are informed that deception was used.

Beneficence refers to "the philosophy of *do no harm* while maximizing benefits for the research project and minimizing risks to the research subjects." Harms may be physical as well as non-physical, such as psychological stress. Researchers should strive to maximum benefits of the research while at the same time, minimizing any physical or psychological harm. In circumstances in which there is no clear benefit expected from a research study, the research study would typically not be approved by an IRB. Participants' time and energy should not be wasted for research that does not stand to produce some actual benefit to society.

Justice refers to "ensuring reasonable, non-exploitative, and well-considered procedures are administered fairly — the fair distribution of costs and benefits to potential research participants — and equally." Participants who experience the burden of the research procedures should also benefit from the products of the research. For example, imagine that a community of individuals living near marshes is recruited to participate in a new study to test a vaccine for malaria. Malaria is a disease transmitted through the bite of a mosquito (Packard, 2007). It is a particularly deadly disease, killing 100 million people per year and making twice as many sick. Those who survive it, suffer lifelong problems. Marshes are one of the favorite breeding grounds for mosquitoes. The study is carried out and found to be effective in preventing malaria. The researchers conducting the study leave the community and work toward bring the

drug to the market. If it is the case that those who participated in the study will not themselves have an opportunity to benefit from the fruits of the research, then one could say that they endured the cost of the research as a participant but did not benefit. This would violate the principle of justice. If those who participate in research are ensured that they can also one day benefit from the future results of the study, should any benefit occur, then the principle of justice would be satisfied.

Examples of Unethical Research

Modern day research ethics has been greatly informed by past examples of unethical research. One of the most shocking examples of unethical research was carried out by the United States government from 1940-1970. It has come to be known as the **Tuskegee Syphilis Study** (Jones, 1981)**.** The Department of Health and Human Services directed the study in Tuskegee, Alabama from 1932 to 1972. Syphilis is a contagious, sexually-transmitted disease caused by bacteria. If left untreated, syphilis can cause blindness, brain damage, and death. Syphilis can cause miscarriages in women and premature births of infants. Children born with syphilis may have deformities, developmental delays, and seizures. In 1932, the United States Public Health Service enrolled approximately 600 African-American men living in Macon County, Alabama in the study. Of the 600, 399 had been diagnosed with syphilis and were monitored for the purposes of determine the progression of the disease; 201 did not have the disease. The 399 men who had syphilis were not told that they had it. They were never told that they were enrolled in a research study. By 1947, penicillin was known to be the cure for syphilis. The men were not given the treatment. Furthermore, steps were taken to prevent the men from seeking medical treatment

elsewhere, where they might have been informed about their disease and also cured of it. When the study was halted in 1972, only 74 of the 399 men with syphilis were still alive. Twenty-28 men had died of syphilis. One hundred others died of complications directly related to syphilis. Forty of the men had wives who had also contracted the disease. Nineteen children are known to have been born with syphilis. In 1997, the United States government formally apologized to the survivors and the families of those harmed in the study. The film and play *Miss Evers Boys* (1997) tells the story of Tuskegee by focusing on Miss Evers a nurse who was hired to recruit men for the study.

Unfortunately, the Tuskegee Syphilis study is not the only example of unethical research conducted with human research participants in the 20th century. In the decades following the invention of the atomic bomb, countless men, women, and children living in the United States and Canada were exposed to radiation for research purposes without their knowledge or consent. These studies have been referred to as the **human radiation experiments** (Welsome, 1999; Jones, 2005). In 1994, then President Bill Clinton called for an investigation into these studies. In 1995, the final report was presented to Congress and is now available on the World Wide Web (Department of Defense, 2011). The report was approximately 1000 pages long and detailed thousands of experiments carried out between 1944 and 1994, including the injection of radioactive substances into infants and pregnant women, the placement of radioactive subjects in the daily meals of children who were housed in schools for those with intellectual disabilities, and the exposure of members of the military as well as incarcerated prisoners to high levels of radiation.

Research Misconduct

The maltreatment of research participants is just one way in which researchers may breach codes of research ethics. Anytime researchers violate the ethical code of conduct, they are described as engaging in **research misconduct.** A familiar example of research misconduct is **plagiarism,** which occurs when an author uses the verbatim words of another with acknowledgement and proper citation. Another example of research misconduct is **data fabrication,** which occurs when a researcher states that data were collected when they were not; rather, the researcher *fakes* the data. A recent example of data fabrication is the research reported by Dr. Andrew Wakefield (born 1957), a surgeon and medical researcher, who proposed a link between childhood vaccines and autism (Godlee, Smith, & Marcovitch, 2011). After many years of controversy surrounding the vaccines caused autism, it was found that the data supporting the claim had been fabricated. A second example of data fabrication comes from the early days of psychology. Cyril Burt (1883-1971) published many studies on the inheritance of intelligence, from studies involving identical twins (Fletcher, 1991). The fraud was only discovered after his death.

There are penalties for those who engage in research misconduct. Individuals may lose their positions, depending on their employment contracts. Those who plagiarize and profit from the work may be sued for damages by the individual whose work was stolen. Those who commit research misconduct while working on a federally funded research project can be barred from receiving future research grants for either a number of years or for life (Altman & Hernon, 1997). Typically, one would be barred

from receiving future research grants only after being found guilty of research misconduct more than once.

Professional Codes of Ethics

The American Psychological Association has a detailed code of ethics. The APA code covers both the ethical issues involving research and professional practice. APA formed its first committee on the ethical standards for psychologists in 1947 (Hobbs, 1948). The first ethics code was published in 1953 and has been revised nine times, most recently in 2009 (APA, 2011). The code provides guidelines for every aspect of a psychologist's professional life. For psychologists involve in delivering treatments in a clinical setting, the ethics code cautious against having dual relationships with clients. Psychologists should refrain from having any type of personal relationship with a client away from the clinic. The current code also provides recommendations regarding the treatment of human research participants as well as animal subjects. The code emphasizes the responsibility that researchers have in ensuring that the results of the research are not used by a third party to bring harm to others. The code also states that researchers have an obligation to share their research data with others.

The APA Ethics Committee is receives and reviews allegations of unethical conduct by psychologists (APA, 2011b). In the event that the committee finds evidence of unethical conduct, there are a number of different sanctions that can be issued. The psychologist in question may be reprimanded if it was found that there was a breach of ethics, but that no one was harmed and the professional was also not negatively affected. A censure can be issued if there was some level of harm to an individual and to the profession. Typically, the level of harm involved in a censure is not extreme. In

cases of extreme harm to an individual or to the profession, an expulsion is issued. In cases of expulsion, one loses membership to the APA. It may be the case that when an expulsion from APA has occurred, the offending member has already lost their license to practice psychology at the state level and may have lost their membership privileges at the state level. The most frequently cited reason for expulsion involves dual relationships (Phelan, 2007).

Key Terms

1. 68-95-99.7 Rule
2. Alpha Level
3. Alternative Hypothesis
4. Bell Curve
5. Belmont Report
6. Beneficence
7. Biased Sample
8. Case Study
9. Central Limit Theorem
10. Central Tendency
11. Common Rule
12. Confidence Interval
13. Control
14. Convenience Sampling
15. Correlational Design
16. Data Fabrication
17. Debriefing
18. Demand Characteristics
19. Dependent Variable
20. Descriptive Statistics
21. Distribution
22. Double-Blind Design
23. Ecological Validity
24. Effect Size
25. Experimenter Bias
26. External Validity
27. Human Radiation Experiments
28. Independent Variable
29. Inferential Statistics
30. Informed Consent
31. Institutional Review Board
32. Internal Validity
33. Interrater Reliability
34. Justice

35. Law Of Large Numbers
36. Likert-Type Scale
37. Mean
38. Median
39. Meta-Analysis
40. Mode
41. Naturalistic Observation
42. Negatively Skewed
43. Null Hypothesis
44. Nuremburg Code
45. Normal Distribution
46. Operational Definition
47. Outlier
48. Pilot Testing
49. Probability
50. P-Value
51. Placebo Effect
52. Plagiarism
53. Positively Skewed
54. Power
55. Questionnaire
56. Randomized Clinical Trial (RCT)
57. Random Assignment
58. Range
59. Reliability
60. Replication
61. Representative Sample
62. Research Misconduct
63. Research Population
64. Research Sample
65. Respect For Persons
66. Significance Testing
67. Simple Random Sample
68. Skewed Distribution
69. Social Desirability Bias
70. Standard Deviation
71. Statistics
72. Subject Variable
73. Survey
74. Quasi Experiment
75. Third Variable Problem
76. True Experiment
77. Tuskegee Syphilis Study
78. Type I Error
79. Type Ii Error
80. Validity
81. Variability

Review Questions

1. What is internal validity? What type of methodology has the highest level internal validity?
2. Compare and contrast a representative sample and a biased sample? Which type of sample is the best to use for empirical research?
3. What is the difference between descriptive statistics and inferential statistics?
4. Discuss the internal validity and external validity of the methodology referred to as the naturalistic observation?
5. Discuss the internal validity and external validity of the case study methodology?
6. What are the three most commonly used measures of central tendency?
7. Explain the differences between a skewed distribution and a normal distribution.
8. What is the central limit theorem? How do researchers use information about the central limit theorem to evaluate samples?
9. What is the 68-96-99.7 rule? Explain how one can use it to estimate what percentage of cases in a population will have a characteristic of interest?
10. Contrast the circumstances in which a researcher would make Type I and Type II errors.
11. What is the relationship between a study's statistical power and sample size?
12. What is the relationship between a study's statistical power and the chance of a Type II error occurring?
13. What is meant by the statement "science is self-correcting"?
14. Discuss the extent to which researchers can predict the outcomes of large populations and the outcomes of individuals.

15. What is the Nuremberg Code? Identify three out of the 10 principles included in the code.
16. What are the three principles set forth in the Belmont Report?
17. What is informed consent and why is it used in research involving human research participants?
18. What is *debriefing* and when is it used in research involving human research participants?
19. What was the Tuskegee syphilis study? What is its significance in the history of research involving human research participants?
20. What were the *human radiation experiments?* What modern principle of research ethics did they often violate?