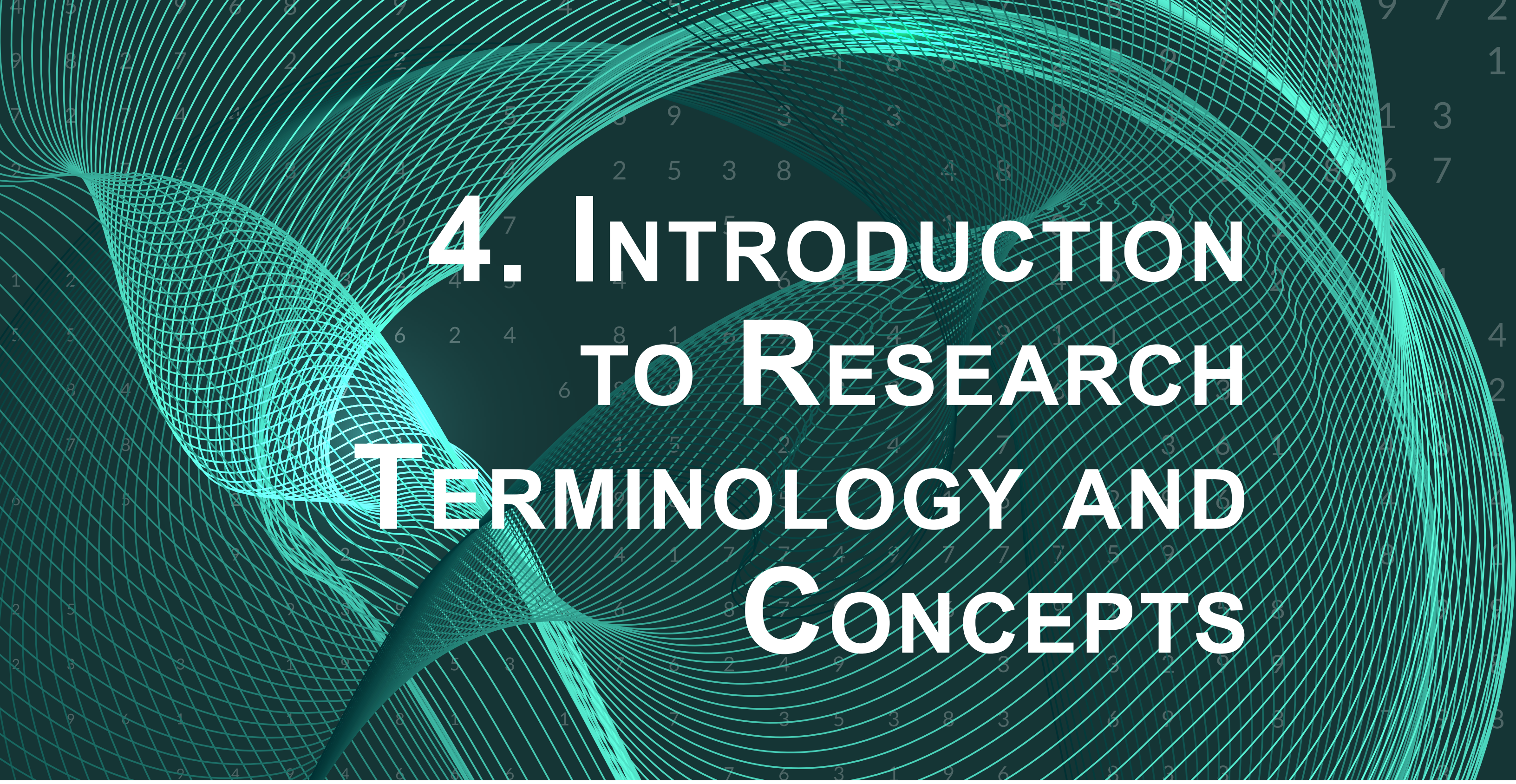

4. Introduction to Research Terminology and Concepts

Sections 4.1 - 4.13

Deborah Chase, J.D., Ph.D

4.1 Introduction

Judges are not required to become statisticians in order to critically read or listen to reports from experts. The world of research and statistics, like law and other disciplines, has its own language. For the most part, the language of statistics is numbers and relationships among numbers. Reports of experts attempt to translate that language into narrative that can be clearly understood by non- statisticians. Familiarity with some of the basic research terminology, concepts and methods can help understand the efforts of experts to translate their work into plain language, and assist the judge as gatekeeper between science and law. It can also help identify the strengths and weaknesses of claims made as the result of the research being presented. Once effectively communicated, most statistical information reveals ideas, inquiries, conclusions and rationales that are common in daily life, although not always intuitive.

4.2 Statistics

Statistics are numbers that represent the data a researcher has collected from a sample, or subset, of a population. A statistic summarizes that data. A **descriptive statistic** allows data to be categorized according to its properties. As its name implies, it simply describes the data and does not make inferences about the larger population. Examples of descriptive statistics include measurements of frequency (counts, percentages); central tendencies (mean, or average; median, or midpoint; mode, or most frequent value); variance in the distance of data points from the mean (standard deviation); and variance in the distance of the data points relative to each other (percentiles, quartiles). An **inferential statistic** generalizes outcomes from a sample to a population. For example, if a researcher wanted information about the ages of individuals filing for divorce in a given court, a sample of divorce cases (the population) would be selected to study, and the resulting outcomes would be statistics. If warranted, the researcher could then make an inference about the average age of the divorcing public. The average age would be a statistic. If instead, the researcher was to try and gather data from the entire population, that process would be called a **census** and the resulting statistic would be a **parameter**.

4.3 Data

Data are the pieces of information that a researcher collects in a study. Most data are expressed numerically, but not all. Identification data, such as eye and hair color, are examples of non-numerical data, while height and weight are numerical data. **Discrete Numerical Data** measure items in which each variable can have only one value. For example, the number of times a five turns up on 50 fair dice rolls will produce one certain number between 1 and 50, a finite number. (However, measuring how many fair dice rolls it takes to produce 50 fives also produces one certain number between 50 and infinity.) **Continuous Numerical Data** may contain any value within some range. For example, the average number of minutes it takes to conduct a hearing on child support is a continuous variable. Any value is possible. This is the opposite of a discrete variable. **Categorical Data** represents values taken from a small group of categories. Values for gender, employment status, or educational level are examples of categorical data. A categorical variable may be assigned a number. For example, female might be assigned the number 1 and male assigned the number 2. These are simply for identification. They have no numerical value and are referred to as **nominal** data. Some categorical data, however, imply a logical order or hierarchy, such as education level or Likert scale data. That subset of categorical data is referred to as **ordinal**. Although there is an order, the distances between data points cannot be quantified, and therefore one should be skeptical of analyses that perform mathematical operations on ordinal data (e.g., averaging scale data). **Labels** simply refer to data sets in which each individual is given a unique name or other identifier.

Data Types -- Examples

Name	No. Of Children	Age	Sex	Employment Status	Education Level	Gross Monthly Income
Susan Wilson	3	55	F(1)	Employed	College	$90,000
Henry Jones	0	75	M(2)	Retired	Ph.D	$150,000
Rose Cohen	2	36	F(1)	Employed	High School	$75,000
Ronald Kruse	2	38	M(2)	Unemployed	Some College	$30,000
Label	*Discrete Numerical*	*Continuous Numerical*	*Categorical*	*Categorical*	*Categorical Ordinal*	*Continuous Numerical*

Table 4.1

4.4 Variables

A **variable** is almost any kind of data that has a quality or quantity that can change from individual to individual. A **dependent variable** is the subject of a research inquiry. Its value is expected to change when exposed to—or be dependent upon—a **treatment**, intervention, or condition of some sort. For example, in a drug test, the treatment is the drug. It is the **independent variable**.

Suppose a researcher for the forestry service wants to see if fire in a forest with a high density of trees will spread over a wider land area than one with a lower tree density. The extent to which a fire spreads is the dependent variable, and the density of trees is the independent variable. A **confounding variable** is one that intervenes into causal relationships between the dependent and independent variables.[1] If the forestry researcher is studying forests located in a climate with unusually high levels of rainfall annually, the rain might affect the spread of fires. The annual rainfall is a confounding variable that interrupts a causal relationship between forest density and the extent to which fire spreads.

4.5 Hypotheses

A **scientific hypothesis** is an idea about the natural world. This idea may come from physics, economics, psychology, medicine, or any other field that studies the natural or physical world. A scientist wants to find out whether an idea is correct and can actually predict expected outcomes in the natural world. Therefore, a scientist will test the idea (hypothesis) by conducting experiments, making observations and performing statistical analyses that can confirm it to be true, or reject it. A scientific hypothesis is assumed to be true. Therefore, it must be consistent with all possible data in the empirical world which is inexhaustive and open-ended. A scientific hypothesis simply cannot be proved. Statisticians attempt to solve this dilemma by adopting an **alternate hypothesis** – the **null hypothesis.** The null hypothesis is the opposite of the scientific hypothesis. It assumes that the scientific hypothesis is not true. The researcher conducts a statistical analysis of the study data to see if the null hypothesis can be rejected. If the null hypothesis is found to be untrue, the data support the scientific hypothesis as true.

A hypothesis can not be proved; it can only be disproved.

4.6 Populations and Samples

Populations refer to all individuals or objects that are the main focus of a scientific study. For example, if a study looks at one or more characteristics of people living in Chicago, then the population for the study is the entire population of Chicago. Populations tend to be very large and impractical to study. Researchers attempt to address this issue by using a **sample** of the population to study. However, applying results from a sample to an entire population can prove challenging. There are numerous obstacles to measuring a population. Examples include low response rate, omissions, uncollectible data, unintelligible data, and more. The boundaries of a population (**parameters**) must therefore be estimated and adjusted to account for missing data.

A poorly selected sample will lead to unfair, invalid, and inaccurate results when applied to the population being studied.

An interesting example of a sampling challenge is the decades long effort to validate the Minnesota Multiphasic Personality Inventory (MMPI) for use with the general U.S. adult population. In the 1940s, the original MMPI set out clinical scales designed to compare responses of a sample of psychiatric patients to a control group that did not have psychiatric diagnoses. The responses of the psychiatric patients reflected their previously established clinical diagnoses. The control group in this original study was small and consisted of young, white, married individuals from the rural Midwest. This sampling was not representative of and did not support the validity of the MMPI for the general population. Nevertheless, the MMPI was used for the next 50 years. In 1989, the MMPI-2 was published. This version used a broader sample of Americans over the age of 18 years from more diverse and representative backgrounds. In 2008, the MMPI-2-RF was published. It provided extensive reports on external comparative data from numerous subjects in a variety of sample settings.[2] The control group statistics for the MMPI-2-RF include responses from 68,377 individuals, over half of whom were defendants in criminal cases. Research continues to work on expanding the validity of the MMPI-2-RF to more sample settings, and to individuals in countries other than the United States.[3]

4.7 Types of Studies

There are two basic types of statistical studies – **observational** and **experimental**. An observational study simply observes the presence of a factor that occurs in two or more different groups. The researcher cannot manipulate or control some independent variables. Observational studies do not establish causation. Experimental studies are required to establish causation.

4.7.1 Observational Studies

Surveys are the most common types of observational studies. **Polls**, a type of survey, are reported on cable news 24 hours a day. Surveys consist of a set of questions given to subjects in a sample of the population being studied. Questions can come through written questionnaires, either on paper or online, by telephone, in-person interviews, etc. Surveys and polls are often interesting and entertaining, but their results cannot generally be extrapolated to any larger populations. Their validity is restricted to the profile of voluntary participants. Most people who play with the array of online "tests" or magazine surveys do not really expect their results to be meaningful. Scientific surveys, however, are required to be as specific and clear as possible in all facets of the survey.

4.7.1.1 Polls

Suppose a doctoral candidate wants to see what sort of orders for the physical custody of children are made in divorce cases that involve allegations of domestic violence. She is seeking to test her idea (**hypothesis**) that cases involving domestic violence would have more sole physical custody orders than joint physical custody orders. She decides to conduct an **archival study**[4] of case records in one court location filed within a 12-month period. She selects a random sample of cases by identifying the case numbers of all divorces with children filed during one year and using a computerized randomizer to select 500 of these case numbers. After she completes collecting data from the 460 of the case files, she separates the cases into 2 groups, one group involved domestic violence and the other not involved in domestic violence. The types of custody orders in each group are counted and compared. She then uses these totals (**statistics**) to test the **null hypothesis** - that the domestic violence cases will not have more sole than joint custody orders. The data show that there are not more sole than joint physical custody orders. The null hypothesis cannot be rejected, and the hypothesis is not confirmed.

If this is the only information provided in a report of this study, problems arise. There is a **lack of clarity** about what is being studied. It is unclear what constitutes domestic violence such that a case is included in the sample. Were allegations alone enough or was a restraining or other relevant order required? Was the civil or the criminal definition of domestic violence used? What happened to the missing 40 cases? Were they eliminated by the researcher for some reason (e.g., because there were no children involved in the case; were the case files unavailable; did some cases completely lack custody orders)? There is also a lack of clarity about how physical custody is defined. Must a physical custody order state that it is joint or sole, or has the researcher defined physical custody on the basis of the amount of time the children actually spend with each parent? If parental time is used, what is the cutoff point that qualifies as sole or joint custody?

Assuming the researcher is able to clarify these issues, sampling might also be questioned. It is important that a **sample** be **representative** of the **population** being studied. The selection of cases by randomizing case numbers does result in a random sample of cases in that one court. Additionally, the **sample size** is reasonably large. But this sample is taken in only one court location when there are actually three court locations handling the same type of cases. What about the other two court locations? A more representative sample would include cases taken from the three locations. It would also be helpful to take cases that were filed during more than one 12-month period. Both these adjustments to sample selection result in a more representative sample of the population. It would include cases heard by a larger number of different judges. Different judges can have different attitudes and experience with cases involving domestic violence. Selecting participants from more than one 12-month period would also increase the number of cases heard by different judges as assignments rotate and would help account for changes in the law during any of the study periods.

Suppose the researcher selects samples from the three court locations, but wants to keep the aggregate size of the samples to 500. The largest courthouse with the most departments hears about 60% of all cases involving domestic violence, so 60% of the goal number of 500 total sample cases should come from that court location – 300 cases. The first branch court hears about 25% of all cases in the population – 125 cases; the third branch hears about 15% - 75 cases. The resulting sample then has what is referred to as proportional representation of cases from each of the court locations. The researcher now has a more diverse sampling of the population, and the results are more representative of it.

4.7.1.2 Interviews and Surveys

Interviews and **surveys** of the subjects of a population are probably more common than conducting an **archival study** as above, or a **meta-analysis** (a statistical study that attempts to identify an effect common to several different existing studies). Interpersonal data collection is more complex since it requires very clear and effective communication between the researcher and the subjects. Suppose a probation director wanted to see what percent of convicted misdemeanor defendants leave the arraignment court with a clear understanding of the terms of their probation. The court agreed to have two experienced criminal department clerks seated at a desk outside the courtroom to interview the individuals as they left. **Ethical issues** were addressed by making it very clear that participation was **voluntary and confidential** and would not affect their court cases. No names would be recorded in the study, subjects were only identified by a unique number assigned by the clerks. The clerks had been provided with a set of questions, an **instrument or tool** to ask the subjects. These questions must be crafted carefully to elicit responses that apply to the study and not exceed its scope. Assessing the quality of questions asked in survey or poll is not unlike making evidentiary decisions at hearings and trials. In both settings, the goal is to access complete, reliable and valid facts. Thus, many of the objections to questions in court mirror problems with questions in a survey. Examples of poor survey questions include those that lack clarity due to vagueness or ambiguity; are leading; compound; call for guesses, conclusions; lay opinions; hearsay; assume facts that are not supported by evidence; and exceed the scope of the study. In studying the degree to which misdemeanor defendants understand their terms of probation, suppose the clerks were provided with a list of questions that included the following:

> *The law and research have the same goal—access to reliable and valid facts.*

1. On a scale of 1-10, please rate how fair the judge was (with 10 being the fairest)

   - *Comment:* This is a leading question. It assumes that the judge was fair, even if only a little.

2. How do you feel about that?

   - *Comment:* This is unclear and vague. The subject doesn’t really know how to answer the question.

3. Have you ever had a close friend or relative convicted of a misdemeanor?

   - *Comment:* Why ask this at all? It calls for hearsay; hardly seems relevant; and, is beyond the scope of this study

4. If, yes, did they understand the terms of their probations?

   - *Comment:* This would be objectionable for the all the reasons mentioned in #3, but also would call for the subject to guess.

5. Was the judge knowledgeable about the facts and law in your case?

   - *Comment:* The judge could be well versed in both the facts and the law without the subject understanding that. It calls for a lay legal opinion. It is also compound.

6. Have you ever been convicted of a felony?

   - *Comment:* This question is totally out of the scope of the study. It is also intrusive and may be unnecessarily off-putting.

7. Did you find the written document you were given setting out the terms of your probation to be clear?

   - *Comment:* The question assumes that the subject can read – a fact not in evidence—and actually did read the document.

Personal interviews of subjects involve considerable subjectivity on the part of the person asking the questions. In this study, the clerks would have to be trained on how to deal with responses from subjects to matters not covered in the list of questions, or inconsistent input from the subject. For example, an individual may not understand why the order was made in the first place. He may answer that he doesn't understand the terms of probation when in fact that is not what he is confused about. Another subject might answer that the terms of her probation were clear, then proceed to ask the clerk a number of questions that demonstrate that she does not understand them. The degree to which the two clerks respond to these situations in like manner determines the **inter-rater reliability** of the clerks. The study should also address the **timing of the interviews**. Scheduling them to occur right after court might result in subjects experiencing heightened levels of emotion—they might be anxious and less able to comprehend what they read until they calm down. They might be quite angry, and their responses would be affected by that. Further, participation was voluntary. Some people agreed to be interviewed, some did not. If many subjects refused to take part, the study would have a low **response rate**. This might suggest that the study was **biased** in favor of those willing to talk to court personnel.

Remember

Observational studies do not establish causation.

4.7.1.3 Surveys (Questionnaires)

Suppose that instead of interviews, the probation department decided to put the questions in a written survey and send it in the mail to subjects. Potentially more subjects could be reached that way. However, not all questionnaires would be returned, and a low response rate would have to be accounted for. Further, subjects might modify the questions in some way and then answer the modified version, or write narrative answers onto the questionnaire. Characteristics of a good poll, survey or interview are fundamentally the same.

4.7.2 Correlational Studies

Correlational studies are observational studies. They seek to demonstrate whether there is a statistical relationship between two or more variables. If a relationship exists, when one of the variables increases or decreases, the other one also increases or decreases. A correlational study does not establish a causal relationship between these variables.

Medical and psychological studies frequently use correlational studies. Correlations are the mainstay of **epidemiological studies**. Although a correlation does not establish causation, it may identify a relationship between two variables that is strong enough to suggest the possibility of future experimentation. A **Pearson's Correlation Coefficient (r)** indicates the direction and strength of the relationship between two variables and ranges from -1 to 1. A coefficient of +0.8 reflects a very strong **positive correlation**. This means that when one of the variables increases, the other variable increases as well. A coefficient of r=-0.5 is a moderately strong **negative correlation.** This means that when one variable goes up, the other goes down. A coefficient of r=0 means there is no relationship at all between the variables.

When plotted onto a graph, the data points underlying a correlation are expected to form a straight line, not a curve. The following graph, which shows the values of two variables plotted along two axes, is called a **scatter plot.** If the data points move in relation to each other, they will cluster around a straight line on the graph. If the relationship is weak or nonexistent, the data points will be scattered around over the area of the graph. See below for examples of scatter plots showing positive (Chart 4.1), negative (Chart 4.2), and no correlation whatsoever (Chart 4.3).[5]

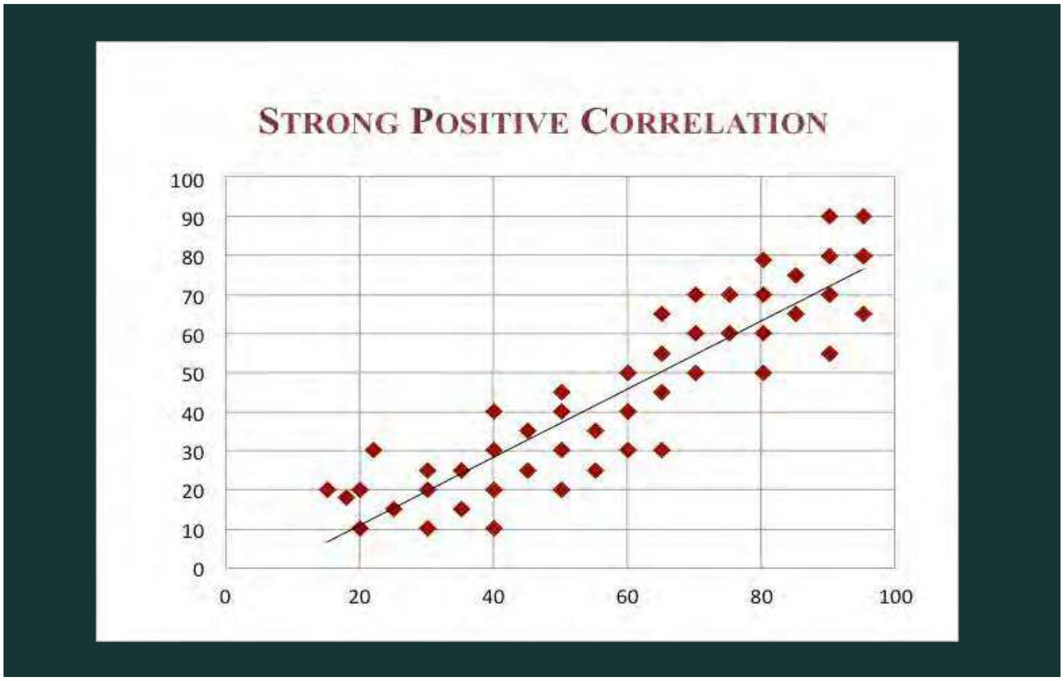

Chart 4.1

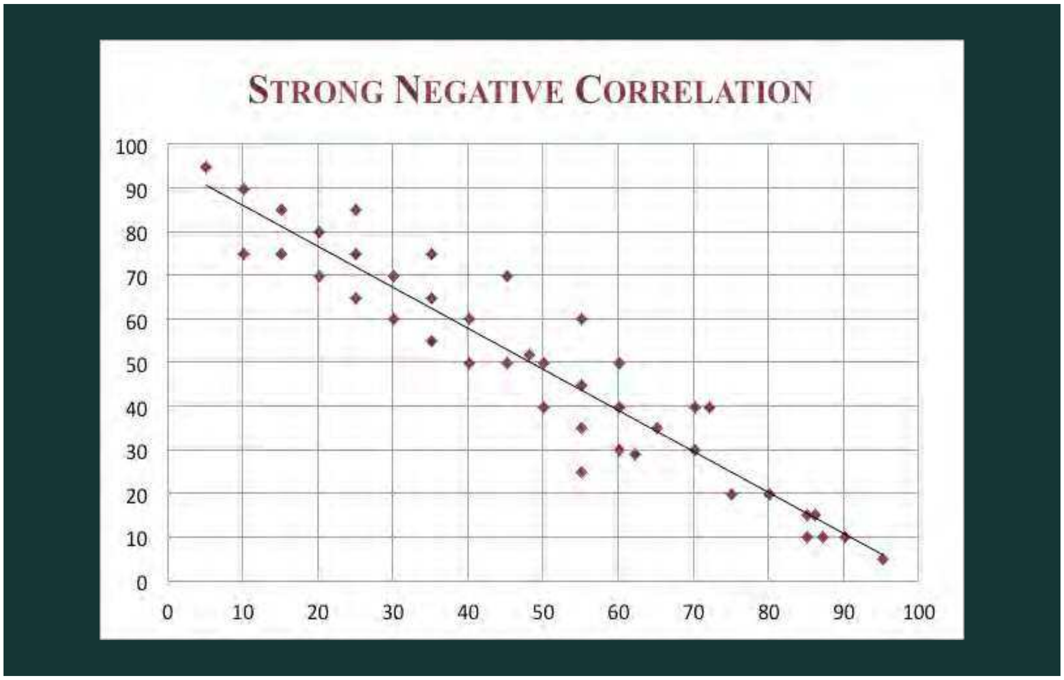

Chart 4.2

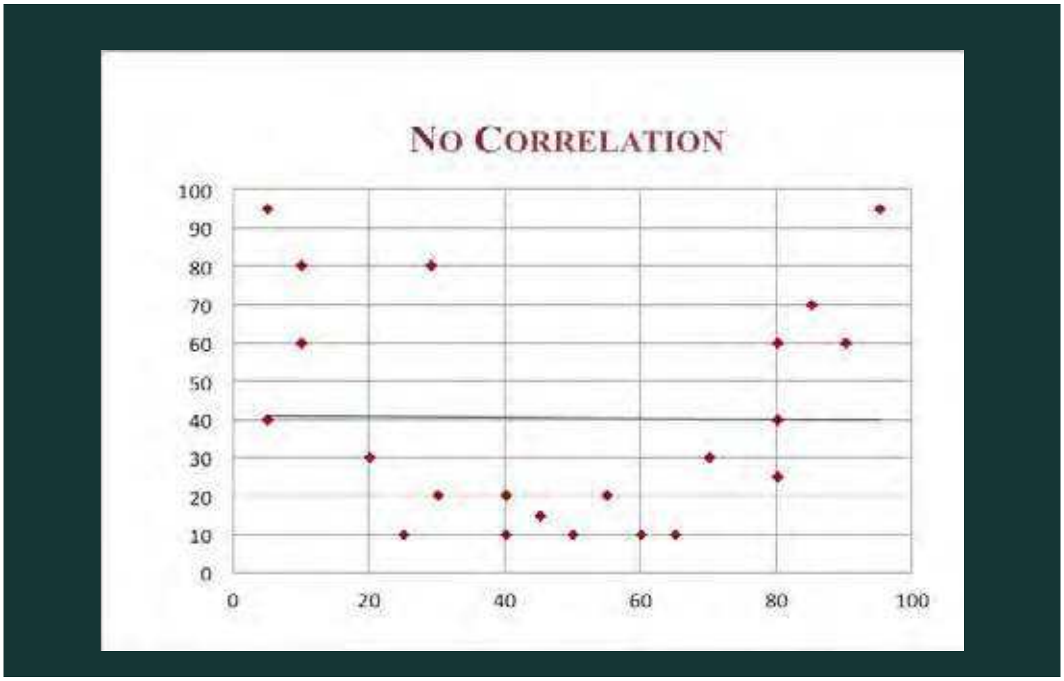

Chart 4.3

Unrelated variables may also produce a linear relationship when measured. These variables are completely independent and simply rise and fall at a similar rate to one another. These are called spurious correlations. Two examples of spurious correlations are: (1) demonstrates a positive statistical relationship between US spending on science, space and technology and suicides by hanging, strangulation and suffocation (Chart 4.4); and (2) another demonstrates a correlation (negative) between the divorce rate in Maine and the per capita consumption of margarine (Chart 4.5).[6]

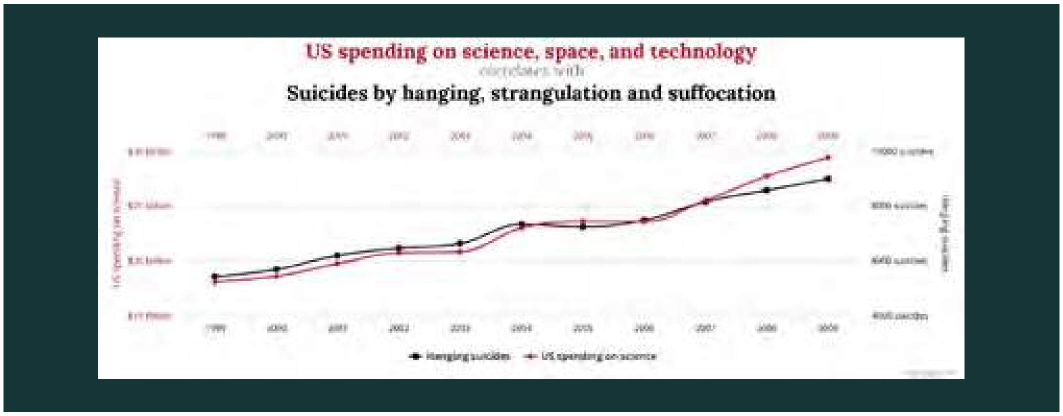

Chart 4.4

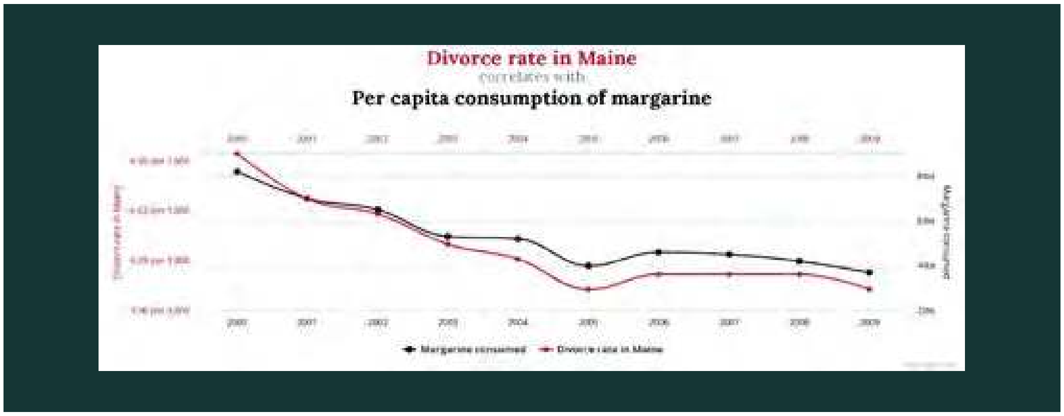

Chart 4.5

Spurious correlations illustrate why these studies cannot establish causation. While it is true that some correlations may suggest possible causation and the need for experimental study, these examples of spurious correlations obviously do not warrant any further study.

A correlation establishes the existence (or not) and strength of a relationship between two variables. There is no dependent or independent variable. Once a correlation of at least moderate strength has been established, (r= at least +0.5 or -0.5), a researcher might want to measure how much a change in the value of one variable is affected by the change in value the other. For example, if you increase the value of one variable by 3 units, and the second variable increases by 1 unit, can the researcher make an accurate prediction that every time the first variable is increased by 3, the second variable will increase by 1? Now there is a dependent and an independent variable. The variable the researcher increases by 3 is the independent variable, the responding variable that increases by 1 is the dependent variable. The statistical test to determine whether this pattern is predictable is called a regression. A **regression** measures the relation between the mean value of one variable (the dependent variable) and the corresponding values of another variable (independent variable). The strength of relationship between the independent and dependent variables in a regression analysis is denoted as **R square**, or **R,^{2}** and the values are expressed numerically between 0 and 1. If $R^2 = .75$, then 75% of the variation in the dependent variable is explained by the independent variable(s).[7] The higher the R^2, together with a low p-value ($p \leq .05$) the more successful is your statistical model (see Table 4.2).

Regression (R^2)	Probability (p)	Meaning	Quality of Model
High (0.8)	Low ($p \leq .05$)	Model explains a lot of the variance in the data and is statically significant	**Best** **Strong Inferential Value (probative)**
Low (0.2)	Low ($p \leq .05$)	Model doesn't explain much of the variance, but it is significant	**Better than no model** **Weak inferential Value (might be probative, but not very informative)**
High (0.8)	High ($p \geq .05$)	Model explains a lot of the variance, but is not significant	**Mostly Worthless** **Mostly no Inferential Value (probably prejudicial)**
Low (0.2)	High ($p \geq .05$)	Model doesn't explain much of the variance and is not significance	**Worst** **No Inferential Value (prejudicial)**

Table 4.2

4.7.3 Experimental Studies

Experimental studies are different in that they can establish a cause and effect relationship among variables. The goal of an experiment, or observational study, is to achieve results that are statistically significant; that is, not occurring by chance. In an experiment, a researcher will select a random sample from the population being studied. Subjects can be randomized twice – once in random selection from the population and again in random assignment to a **test** group or a **control** group. The test group is then exposed to a **treatment** (or other intervention) that is expected to cause an effect. The control group is not exposed to the treatment. The conditions of subjects in the sample are the dependent variables; the treatment is the independent variable. The control group is used to establish a baseline without which a comparison is impossible. Assignment to either the test or control groups should be made by the researcher or other third party, not self-selected by the subjects. Ideally assignment to test or control groups should be double-blinded to eliminate potential bias. A **double-blind study** is one in which neither the researcher nor the subjects know which group is being given the treatment being studied. In a **blind study**, the researcher knows which group received the treatment, but the subjects do not. Random assignment to a sample group is not always possible. A **quasi-experimental design** can use some other criteria for sample selection. For example, one sample group has subjects with a last name starting with the letters A through N, another group includes those whose last names start with M through Z.

Experimental researchers must pay careful attention to **ethical issues** frequently encountered. Medical research regularly uses experiments, particularly in testing the efficacy and dangers of new drugs. Suppose the research is testing a lifesaving drug that would extend the life span of patients waiting for heart transplants. The population consists of heart patients on the list for transplants. The researchers select a random sample of these heart patients; then randomly assign them into two groups. One group receives the drug, the other does not. However, it is unethical to deny this treatment to any of the sample patients – two groups could not be selected.

The ethical issues for a court wanting to conduct an experiment are substantially the same. For example, even though there are no life-threatening outcomes confronted by a study of subjects' understanding of their terms of probation, serious ethical concerns might arise. Suppose the court wanted to expand the study to include testing to see if providing subjects with help to understand their orders and connect with community based services might affect their rates of probation violations. Two samples were selected from the population of misdemeanor defendants. One group included those whose hearings fell on Mondays, the other group had their hearings on Wednesdays. The Monday group was provided with help, the Wednesday group was not. The cases were reviewed after passage of a specified time (e.g., six months). When compared, the Monday group had fewer probation violations than the Wednesday group. The court's decision to deny help to one group while providing it to the other group creates an ethical problem. This study design might not only violate the ethical standards of a committee designated to protect the welfare of human subjects, it would also infringe on the court's duty to be, and appear to be, neutral. Experiments involving human subjects, as well as other biological life-forms is usually quite challenging. Human subjects must volunteer to participate, as in clinical trials. Conducting experiments on involuntary human subjects (prisoners, civil commitment patients, etc.) is not permitted.

Sample size must be sufficiently large to achieve powerful results. A powerful result includes the level of statistical significance, the effect size and the available data. The larger the sample, the more powerful will be the results. It is crucial, however, to accurately define the sample. For example, suppose a legal services agency wants to test the hypothesis that individuals involved in a court case get better outcomes if they are represented by counsel than if they proceed without an attorney. An attorney is assigned to represent a sample of pro se individuals who are involved in eviction proceedings. The outcomes in these cases are compared to an equal number of eviction cases in which an individual appeared without an attorney. A sample of 100 cases is selected from the population of eviction cases during a one-year period. Over the course of that year, the legal services attorney represented 50 cases. Therefore, as a control, there were fifty cases selected from the pool of eviction cases in which there was no attorney representation. Legal aid assigned the same attorney to represent all 50 of the legal aid test cases. Was the sample size 100, or was it possibly only 51 (50 in the control group and only 1 in the test group where the same attorney was representing all 50)? Was the study about how well litigants did with representation generally, or was it a study of the effectiveness of this one legal aid attorney? The sample was not representative of the population of eviction cases. It was seriously biased toward those who were represented by this one lawyer. Additionally, it is unclear as to the definitions of good and bad outcomes.

4.8 Validity and Reliability

Validity refers to the degree to which a concept is accurately measured in a study.[8] **Content validity** measures the degree to which measurement instruments and methods used in a study, actually measure what they are supposed to measure and cover the **domain** of the subject being studied. Suppose a study wanted to find the average weight of 25-year-old men. A sample of 100 25-year-old men is selected and weighed. The scale used to weigh the men produced accurate weights, but it could only weigh up to 150 lbs. It was thus unable to measure the many men who weighed over 150 lbs.

Construct validity is the degree to which a study is measuring the construct it claims to measure, and determines whether a researcher can draw inferences from a study's results. Taking the MMPI as an example, a subject who scores high on the paranoia scale should be expected to demonstrate paranoid behavior in his daily life. Simpler studies that measure only one thing are more likely to have construct validity. The instruments used in a study should produce values that are related to other studies measuring the same variable. For example, if the MMPI-2-RF paranoia scale values are positively related to the values on the Suspicious/Paranoid scale of the Millon Clinical Multiaxial Inventory (MCMI-III), that indicates that the MMPI has construct validity.

Reliability refers to the degree to which measurements in a study are consistent. In other words, the study will produce repeatable results in subsequent measures. One way to assess reliability is to split the measurements made into two halves, then calculate correlations of each half to see if they are highly correlated. The result will be a number between 0-1. If the correlation is weak (less than .7), the data is less reliable. This test can only be used for questions have two responses (yes or no). With more than two responses, statisticians average all the correlations for every combination of results. Again, the result is between 0-1 and, as a general rule of thumb, should be .7 or higher. A test for the internal consistency of a psychometric measure is called **Cronbach's Alpha(α).** This test uses the number of items in a test, the average joint variability between item pairs, and the variance in the total score. The MMPI includes several validity scales within the inventory. The two most common ones are the L Scale (intended to detect when a subject's answers are untruthful); the K scale (intended to spot when a subject is overreporting or underreporting psychopathology).

Reliable measurements should be stable. That is commonly tested by giving the same subject the same test two times to see if the responses are highly correlated. The questions can be worded differently, but test the same concept. A strong correlation is .7 or higher. Another test for reliability is **inter-rater reliability.** In studies with two or more observers, their observations are compared to see the level of agreement among them. The greater the degree of difference, the less reliable the measurements in the study. Measures of inter-rater reliability can also be used in the pilot phase of a study to assess the quality of an instrument.

Measures should be both **accurate and precise**. For example, one clock will measure time in hours and minutes. A second clock measures the hours, minutes, seconds, and tenths of a second. The second clock is much more precise than the first. However, the second clock is accurate only within a range of ±5 minutes from the true time. The clock is precise, but not very accurate. The extent of inaccuracy is called the **standard error.** The greater the standard error, the less reliable any estimate based on the measurement will be.

The **confidence interval** is also a way to gauge the reliability of an estimate. The confidence interval predicts the parameters within which a sample value will fall. It looks at the distance from the mean a value will fall, and is measured by using standard deviations. For example, if all values fall within 2 standard deviations from the mean, about 95% of the values will be within that range. The larger the confidence interval, the wider the array of values. Given the accuracy issues of the second clock above (±5 minutes from the true time), a researcher might expect a greater variance in the time measurements of that clock. Suppose the times measured would fall within 2 standard deviations from the mean producing a confidence level of 95%. Combined with the standard error of the clock, estimates made using this measure would seem highly unreliable.

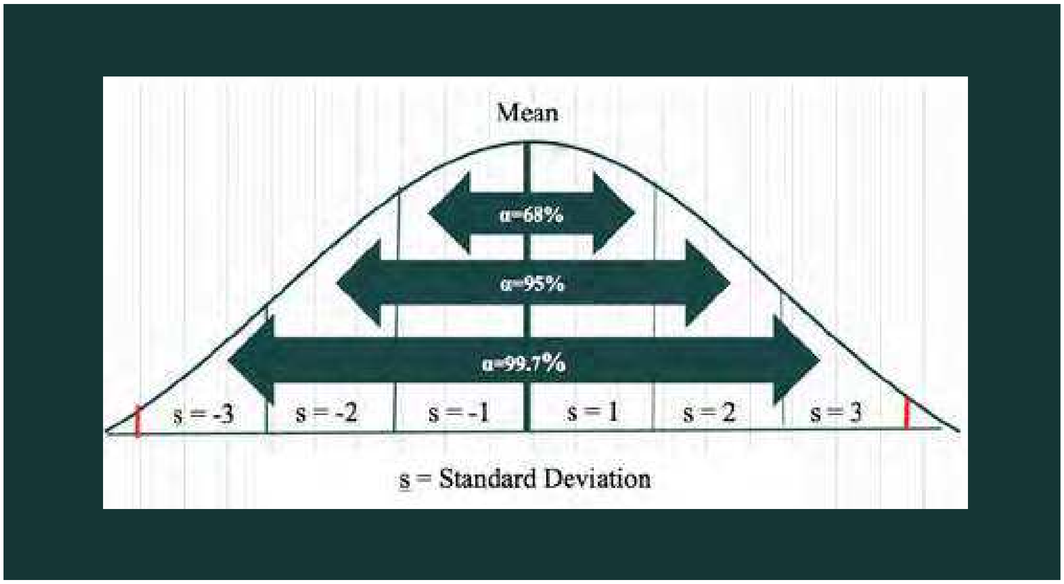

Chart 4.6

Bias refers to a systematic error in a study that overestimates or underestimates a true value in the population. Bias can be found in two basic areas – sample selection and data collection. Lack of random sampling is the main cause of sampling bias. As discussed, in the study of understanding probation terms, the sample was selected from anyone who left the misdemeanor courtroom on a given day. Participants were self-selected, so only those willing to talk to court staff were represented and the study would be biased in their favor.

Observer bias can also disrupt results. If one of the observers in the probation terms study inadvertently spends more time talking with female subjects than male subjects, the results would be biased in favor of the women subjects. A study can also be biased because of **incomplete or missing data.** Suppose a family court judge wants to know whether ordering a parent with alcohol use problems to demonstrate their sobriety by testing daily with a hand-held breathalyzer is helpful. The hypothesis is that if the breathalyzer helps the person stay sober, the number of contested custody hearings might be reduced. The frequency of hearings for both groups is then measured over a specified time period. The outcome shows that fewer contested hearings occur in the group using the breathalyzer; however, a question of bias arises. The researcher has ignored a background factor. There are some individuals in the breathalyzer group that have participated in substance use disorder treatment programs during the study period. The reduction in the number of contested hearings may also be due to a parent's work in recovery. The omission of this background factor would influence the outcome of the study – it is a confounding variable. Many other potential confounding variables probably exist for this study. Suffice it to say that a report claiming that use of the breathalyzer in custody cases involving alcohol reduces the need for contested hearings is inaccurate and misleading at best. The outcome would overestimate the relationship between court-ordered use of the breathalyzer and number of contested custody hearings. Confounding variables are probably the most frequent examples of bias in statistics.

Confounding Variables are probably the most frequent examples of bias in research.

Publication bias occurs when research reports only include results that are statistically significant, and omit results that failed to produce significance. An investigation of publication bias covering more than 4600 publications from different disciplines found strong evidence that publication bias is increasing in frequency, particularly in areas where research funding is scarce.[9] Publication bias can result in intentional research fraud, some from respectable instructions or noted peer reviewed journals. For example, Duke University recently paid $112 million to settle a federal law suit after prosecutors identified falsified and fabricated research data used to win multiple governmental grants from the National Institute of Health, the Environmental Protection Agency, and other federal agencies.[10] Likewise, in 1998, Andrew Wakefield published a study in the esteemed medical journal *The Lancet*, claiming that there was a correlation between measles vaccine and autism. A subsequent investigation in 2010 of his research revealed that data related to the research participants had been misrepresented and was based on only 12 subjects. Further, Wakefield was partially funded by an attorney seeking to sue the manufacturer of the measles vaccine. In 2011 the editors of the British Medical Journal labeled the research as fraudulent.[11]

Confirmation bias is the tendency to search for, interpret, favor and recall information that confirms one's own pre-existing beliefs or ideas. There are many types of bias possible in a study, but all of them have the same result—the overestimating or underestimating of the true values of study measurements.

Cognitive bias refers to errors in reasoning often occurring as a result of clinging to pre-existing beliefs regardless of evidence to the contrary. The persistent use of the Rorschach ink-blot test by clinical psychological experts in court cases is an example of cognitive bias. This test has been in clinical use for decades and has been generally accepted in the clinical community, but less so in the research community. There is scarce scientific evidence that the Rorschach measures anything in the real world. In fact, scientific studies of the Rorschach conclude that it has no psychometric value, except for identifying severe thought disorders. Yet, the clinical world clings on to it in the face of criticism from research psychology community.[12] Nevertheless, trial courts continue to admit it as scientific evidence across many case types.[13] Another example of cognitive bias in psychology was demonstrated in many cases involving recovered memory.[14]

4.9 Hypothesis Testing

Central Tendency refers to measuring the central or typical value in a sample distribution. There are three common measures of central tendency—the **mean, median** and **mode**. The **mean** is the average of the data set values. It is derived by adding up the sum of the all the numbers and dividing by the number of data points. The **median** is the data point at which there are an equal number of data points above and below the mean. If you arrange all of the sample values from smallest to largest, the median is the midpoint of that array. The **mode** is the value most frequently appearing point in the data set. Suppose a family court judge is interested in how many pre-trial hearings are normally required per divorce case. There are 10 people on the morning court calendar with final judgments. To start the study, the judge asks the clerk to review the case records for those 10 cases to see how many hearings were held prior to judgment. The case records revealed the following numbers of pre-trial hearings: 3, 8, 10, 14, 8, 9,15, 8, 7, 18. The **mean** (average number) of pre-trial hearings per case was 10. More of these cases (6) had less than the mean of 10 while only 4 had 10 or more. The point at which there are an equal number of data points above and below the mean is 8 cases (5 cases above and 5 cases below) is the **median** point. This data shows that although the average number of hearings is 10, most of the cases have 8 or fewer hearings. Although less frequently reported by experts than the mean, the median is also necessary to understand the distribution of data points, particularly if there are outliers or data points on the extreme ends of the range. Researchers should always provide the mean and the median. The mean alone is not sufficient. The data point that appears most frequently (for 3 cases) is 8 hearings. It is referred to as the **mode**. All the other data points appear only once. In this data set, the median and the mode are the same.

> *The report of an expert should always provide the mean and the median. The mean alone is not sufficient.*
>
> *Likewise, the standard deviation should always be included in the report.*

The **Standard Deviation** is the number that describes the extent of variance in the values of a variable. The larger the standard deviation, the more the data points vary from the mean. For the example above looking at pre-trial hearings, the standard deviation from the mean is 4 cases. It is derived as follows:

1. take the mean of the data set;
2. subtract it from each data point;
3. square the differences (to eliminate negative numbers);
4. find the mean of the squared differences;
5. find the square root of the mean of the sum of squares - that will be the standard deviation.

Standard Deviation--Pre-trial Hearings per case

Data Points Hrgs/ Case	Subtract the mean -10	Square each result	Sum the squares $\sum$	Find the mean of the sum of squares $\bar{x}$	Find the square root of the mean of the sum of squares $\sqrt{}$	Standard Deviation S
3	-7	49	176	17.6	√17.6	4.2
8	-2	4				
10	0	0				
14	4	16				
8	-2	4				
9	-1	1				
15	5	25				
8	-2	4				
7	-3	9				
18	8	64				

The **Standard Deviation** is 4 (rounded). Therefore, cases with ≤14 hearings are within +1 standard deviation from the mean; also, cases with ≥6 hearings are within +1 standard deviation. Cases with +2 standard deviations would be those with ≥2 and ≤18 hearings.

In this data set, 7 cases, (70%) are within +1 standard deviation, and 100% are within +2 standard deviations. If plotted on a graph, this distribution would form a normal distribution.

Table 4.3

Percentiles look at how the values of the data points differ from each other rather than how they differ from the mean. The percentile of a distribution is measured by arraying the data points from largest to smallest and locating the value below which a given percentage of the data points fall. In the example above, the data point with 9 pre-trial hearings would be in the 20th percentile. Two cases out of the10 data points (cases with 7 and 3 hearings) are below 9 hearings – 20% of the

10 cases. **Interquartile** ranges are another way of looking at central tendency. It is particularly useful if there are outliers because it does not consider them. To calculate the interquartile ranges, the data points are grouped into percentile sections with each section containing a fourth of the data (25th, 50th, 75th and 100th percentiles). The interquartile range takes the middle of section (25th – 75th percentiles) to analyze, thereby eliminating any outliers (see Chart 4.7).

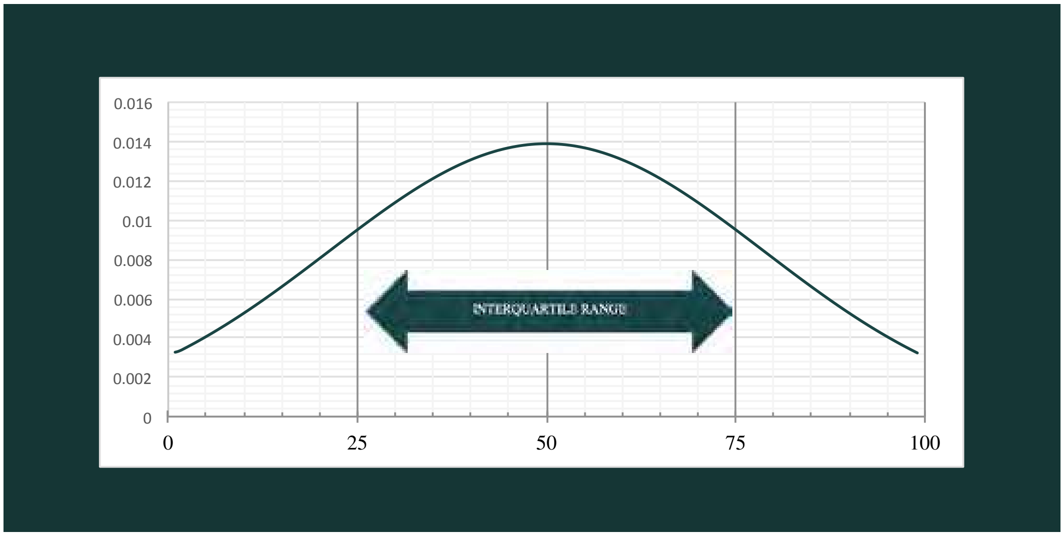

Chart 4.7

4.10 Statistical Significance

The research field agrees that study outcomes must demonstrate they are not the result of random chance. Leaving room for an error of .05, the study must achieve a 95% level of confidence that the results were the product of the study. This is denoted as **p≤.05**. (or .01 or .1)

Type I and Type II Errors

Hypothesis testing must account for making errors in conclusions. For example, in court X, self-represented litigants are referred from the courtroom to the court self-help center to have written orders-after- hearing drafted. The court administrator knows that it takes staff an average of 1 hour to prepare these orders. She believes that if drafting staff were in the courtroom, the amount of time required to prepare the orders would be reduced.

1. The null hypothesis (H_0) is that the outcome of the staff change ≥ 1 hour;
2. The alternative hypothesis (H_a) is that the outcome of the staff changes <1 hour.

After running time testing, it is found that the staff change did not result in a change in the time required to prepare an order. It still took 1 hour or more to prepare an order. The null hypothesis can not be rejected. Nevertheless, a certain percentage of cases that took 1 or more hours to prepare were inaccurately measured as requiring less time. This percentage of cases inaccurately measured constitutes a **Type I error**.

Suppose, however, that time testing revealed that having staff in the courtrooms to prepare orders reduced the amount of time required to 20 minutes. In this case, the null hypothesis is rejected; the amount of time required for order production is <1 hour. Nevertheless, a certain percentage of cases that took less than1 hour to prepare were inaccurately measured as requiring 1 or more hours to complete. This percentage of cases inaccurately measured constitutes a **Type II error**. The null hypothesis can be rejected, and support is found for the alternative hypothesis.

	In Reality	
H_0 = Null Hypothesis	H_0 is True	H_0 is False
Reject H_0	Type I Error	Correct Conclusion
Do Not Reject H_0	Correct Conclusion	Type II Error

Table 4.4

4.11 Probability

4.11.1 Frequentist Probability

Suppose you toss a fair coin; that is, a coin that is not weighted towards either heads or tails. What are the chances you get heads? Heads has a 50% chance of turning up; same for tails. Probability is always denoted by a number between 0-1. Thus, heads will turn up one half the time (50%) – denoted as a **probability of p=0.5**. The same goes for tails. Next, you toss the coin five more times (for a total of 6 tosses). Would you expect to see heads a total of 3 times and tails a total of 3 times (HTHTHT)? You may very well see an unequal distribution of heads and tails. The chances are random. The probability of turning up heads is still p=0.5 and same for tails. This is because probability is intended to predict events over many tests, not just a few. So, if a you toss a coin 600 times, heads will appear about half the time, and tails about half. But it does not say in what order they would appear. The individual tosses are independent of each other – the outcome of one does not affect the outcome of the others. With **expected (a priori) probability** of p=0.5, if a coin is tossed 600 times, and heads comes up 450 times, then you might be inclined to conclude that the **coin was biased** - weighted in favor of heads.

4.11.2 Conditional Probability

What if you toss a fair coin 3 times? What is the probability you would turn up 3 heads? There are 8 possible outcomes. In this case, the expected (a priori) probability changes with each toss. The table below (Table 4.5) shows that 1 of the 8 possible outcomes would result in 3 consecutive heads (HHH). The probability of getting 3 heads in a row is then 1/8 or p=0.125.

No. Tosses	Outcomes								Probability of 3 heads in 3 tosses
1st toss	H				T				
2nd toss	H		T		H		T		
3rd toss	H	T	H	T	H	T	H	T	**1/8 or 0.125**

Table 4.5

What then would the probability be of tossing the coin 3 times and getting 2 heads? The table below (Table 4.6) shows that 3 of the 8 possible outcomes results in 2 heads. Therefore, the probability of getting 2 heads is 3/8 or p=0.375.

No. Tosses	Outcomes								Probability of 2 heads in 3 tosses
1st toss	H				T				
2nd toss	H		T		H		T		
3rd toss	H	T	H	T	H	T	H	T	**3/8 or p = 0.375**

Table 4.6

The table below (Table 4.7) shows that 1 of the 8 possible outcomes results in 0 heads (that is 3 tails). The probability of showing 0 heads is 1/8 or p=0.125. The chances of getting 4 heads on 4 tosses, or 4 tails on 4 tosses, is reduced again by 0.5 to 0.125/2 = 0.0625. This pattern of the reduction in the probability that heads (or tails) will appear in repeated consecutive tosses continues to decline by .5 into infinity. Suffice it to say, that after several tosses turning up nothing but heads would tip of the observer that the coin was biased in favor of heads – it would not take 450 heads out of 600 tosses to signal this bias.

No. Tosses	Outcomes								Probability of 0 heads in 3 tosses
1st toss	H				T				
2nd toss	H		T		H		T		
3rd toss	H	T	H	T	H	T	H	T	1/8 or p = 0.125

Table 4.7

4.11.3 Bayes' Probability

There are other situations in which the outcomes of a test are not as random as the flipping of a coin. In the following example, we have a starting point that is informed by some known population parameters. Suppose we know that about 2% of those diagnosed with alcohol use disorder will develop Korsakoff's Syndrome,[15] a type of brain damage, at some future point. Our population includes those diagnosed with alcohol use disorder. This means that we do not have to start out with the assumption that individuals with alcohol use disorder have a random chance (p=0.5) of developing or not developing Korsakoff's.

Also suppose that scientists have developed a new test that can predict the likelihood that an individual with alcohol use disorder will develop Korsakoff's within the next 5 years. We also know the following:

1. The test accurately predicted those individuals who did not develop the disease within five years 95% of the time;
2. However, 5% of those without the disease at five years were inaccurately predicted as likely to develop it.
3. The test accurately predicted those individuals who did develop Korsakoff's within five years 97% of the time;
4. However, 3% of those who had developed the disease were incorrectly predicted as unlikely to develop it.

Individual Alpha, who was diagnosed with alcohol use disorder, was tested. The test predicted that Alpha will develop Korsakoff's within the next five years. What is Alpha's actual chance of developing the disease? Sometimes it helps to think geometrically (i.e., visually). Let the square (100 sq. units) below (Table 4.8) be the population of all those diagnosed with alcohol use disorder.

Those Without Korsakoff's at five years – 98%

Those With Korsakoff's at five years – 2%

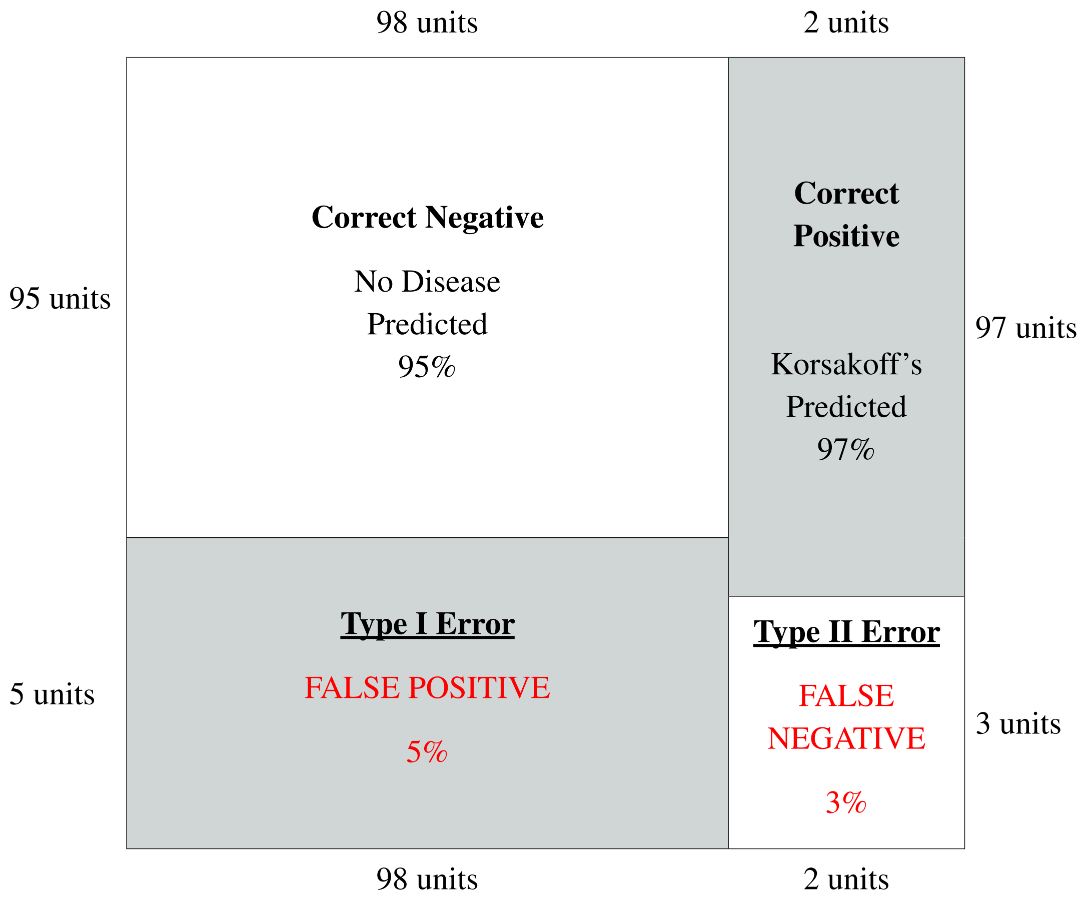

Table 4.8

Alpha must be in one of the two shaded areas of the square above – he will be either in the area of the square correctly predicting he will develop Korsakoff's, or alternatively he could be in the false positive area of the square. The area of the square correctly predicting that Alpha will develop the disease is 97 units x 2 units = 194 units. The area of the square of false positives is 5 units x 98 units = 490 units. The two possible areas of the square in which Alpha must be found, combine to equal an area of equaling 684 units.

What percentage of the combined areas of the two shaded sections (684 units) is accounted for by the portion that correctly predicts Alpha will develop Korsakoff's? To find the answer, we divide the area of that section (194 units) by the combined area of both shaded sections (684 units). The result is .2836, or, a probability of p=0.2836 (28%) that Alpha will develop the disease. Alpha is less likely to develop Korsakoff's than he is to develop it.

A common misreading of this data interprets Alpha's chances of developing the disease as 97%. This is simply the test's accuracy rate for correctly assessing those who have developed Korsakoff's. The test has an excellent accuracy rate, but that rate does not represent the probability, particularly within a population so low in the overall disease rate, of an individual's chances of developing the disease. Even though Alpha was tested as positive for future disease using a measure of high accuracy, his chances of getting the disease are quite low at p=0.28.

Two common probability misinterpretations are the **prosecutor's fallacy** and the **defense fallacy**. DNA is known for the stunning odds that are sometimes reported by the experts. For example, suppose that a DNA sample is taken from a crime scene. A DNA sample is also taken from the suspect. It is expected that a random match would occur once in every million people. The prosecutor argues that if there is a DNA match, there is a one in a million chance that the suspect is innocent. However, this is incorrect. The correct interpretation is that if the suspect is innocent, there is a one in 300 million chance of getting this DNA match. Inverting outcomes can lead to serious error. This is called the **prosecutor's fallacy**. A simple example of inverting outcomes would be the following: because it is the month of July, it must be summer. The inverted fallacy would be to say that because it is summer, it must be the month of July.

Another problem with interpretation of probability occurs when no consideration is given to the whole array of evidence. Suppose that in a city of ten million people, any one person has a one in ten chance of having a particular DNA characteristic. The defense argues that any one person has a 10% chance of being guilty, and a 90% chance of being innocent. However, this does not take into consideration other evidence such as additional trace evidence, eye witnesses, etc. This is called the **defense fallacy.**

4.12 Misleading or Inaccurate Reports

Finding errors in statistics need not be overly complicated. For example, errors can occur in simple math. Totals might be summed inaccurately, or percentages presented that add up to more or less than 100%. For example, error would be obvious in a study of a state with 58 counties that reports data for 66 counties, and then percentages based on that erroneous total.

It is important not to abandon common sense when reviewing research. For example, the numbers in a study may be accurate but the result defies common sense. A well-known example is the “birth weight paradox.”[16] Here a study was conducted comparing infant mortalities of two or more groups of mothers. The groups included smokers, non-smokers, different races, social status, and other variables. The study describes the relationship between birth weight and infant mortality. According to the results set out in the table below, (Table 4.9) low birth weight babies whose mothers smoked have a lower infant mortality rate than those whose mothers did not smoke.

It is important not to abandon common sense when reviewing research.

“Birthweight Paradox”

Infant Body Weight (kg)	Smoking Mothers (No. of Infants)	Infant Mortality (%)	Non-Smoking Mothers (No. of Infants)	Infant Mortality (%)
.5	0		4	25.00
1.0	4	25.00	20	15.00
1.5	20	10.00	315	7.20

Infant Body Weight (kg)	Smoking Mothers (No. of Infants)	Infant Mortality (%)	Non-Smoking Mothers (No. of Infants)	Infant Mortality (%)
2.0	315	4.20	3115	3.00
2.5	3115	1.80	12050	1.30
3.0	12050	.70	19000	.53
3.5	19000	.30	12050	.22
4.0	12050	.13	3115	.09
4.5	3115	.06	315	.04
5.0	315	.03	20	.05
5.5	20	.04	4	
All	50004	.46	50008	.82

Table 4.9

Here is a result that defies common sense. It is generally accepted that smoking reduces birth weight, and that lower birth weight results in higher infant mortality. It may be that other causes of low birth weight are more influential on infant mortality than smoking. Many statisticians have sought to solve this paradox, primarily by identifying the many other factors besides smoking that might lead to lower birth weight and therefore, higher infant mortality.[17]

Also important are the sources supporting the subject of the research. An expert should be able to cite well established peer-reviewed journal articles related to the subject being studies.

4.12.1 Misleading Charts & Graphs

Charts and graphs can display data in misleading ways. Examples include charts that do not report a baseline, report incomplete data or use erroneous or false data; show numbers that do not add up; or display absurd results.[18]

4.12.2 No Baseline

Suppose a court wanted to illustrate the growth statewide in the pro se population of litigants. It counted the total number of pro se filings in 2005 through 2007 and made the following chart (Chart 4.8).

Increase in pro se Litigants (2005-2007)

No Baseline

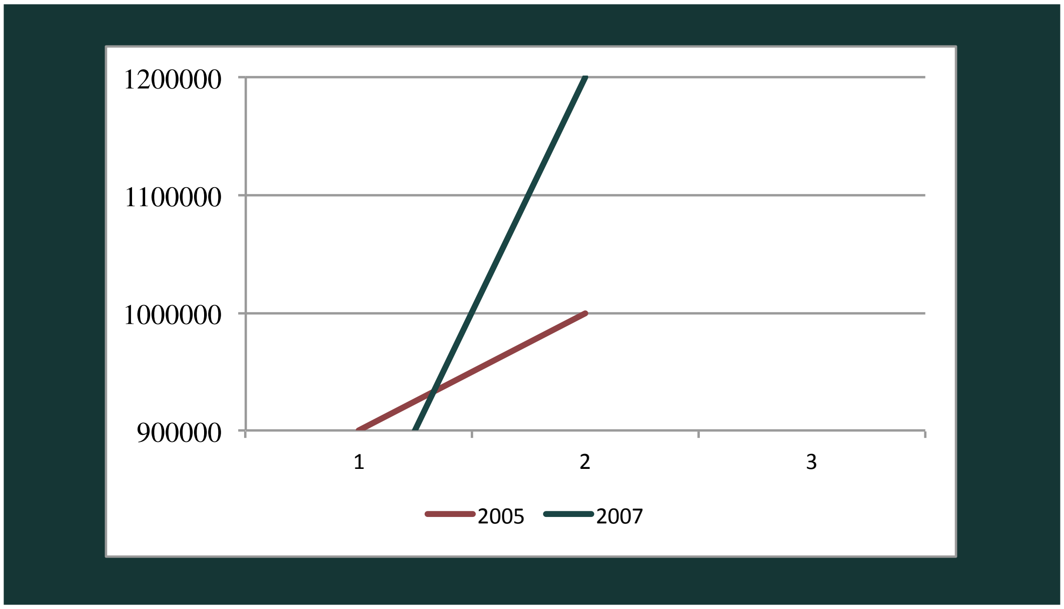

Chart 4.8

There appears to be a large difference between 2005 and 2007. However, note that the Y-Axis starts at 900,000, not at zero. There is no baseline. The bars on this chart make it look like the Y-Axis displays only about 25% of the litigants counted. With no baseline, this chart presents a misleading visual image that makes the increase appear quite large. Checking the chart's numerical data clarifies that there is only about a 20% difference in the numbers of pro se litigants between 2005 and 2007. If the chart had started the Y-axis at a baseline of zero (Chart 4.9) and charted all the data from that baseline, it would be visually accurate.

Increase in pro se Litigants (2005-2007)

With Baseline

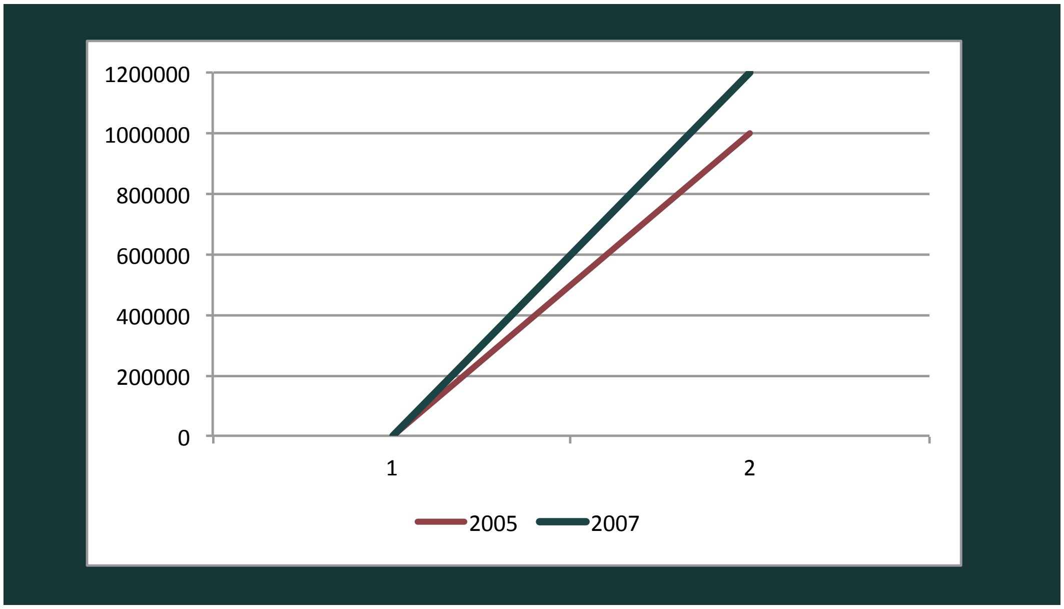

Chart 4.9

4.12.3 Missing or Incomplete Data

Assume a consumer advocacy group wants to illustrate their concern about rising average utilities costs to residential consumers. To do so, it publishes the following chart (Chart 4.10). In fact, the chart makes no sense, there is too much missing data. It basically shows that utility bills increase in cost as the weather gets colder.

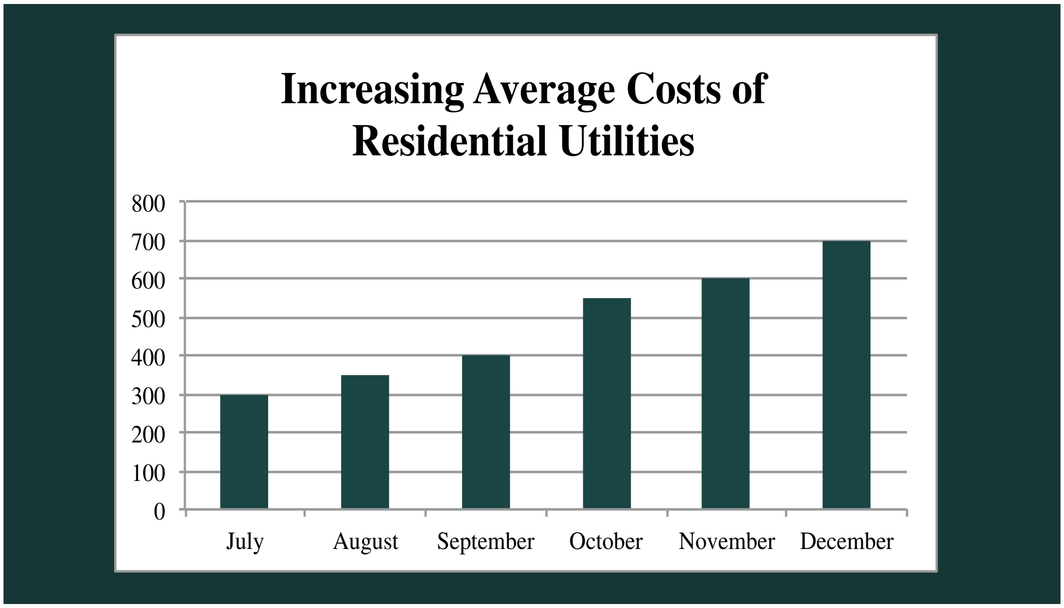

Chart 4.10

4.12.4 Numbers Don't Add Up

Suppose a cable news network did a quick survey of party affiliation of registered voters, and published the results below (Chart 4.11). This chart makes no sense. The numbers add up to more than 100%.

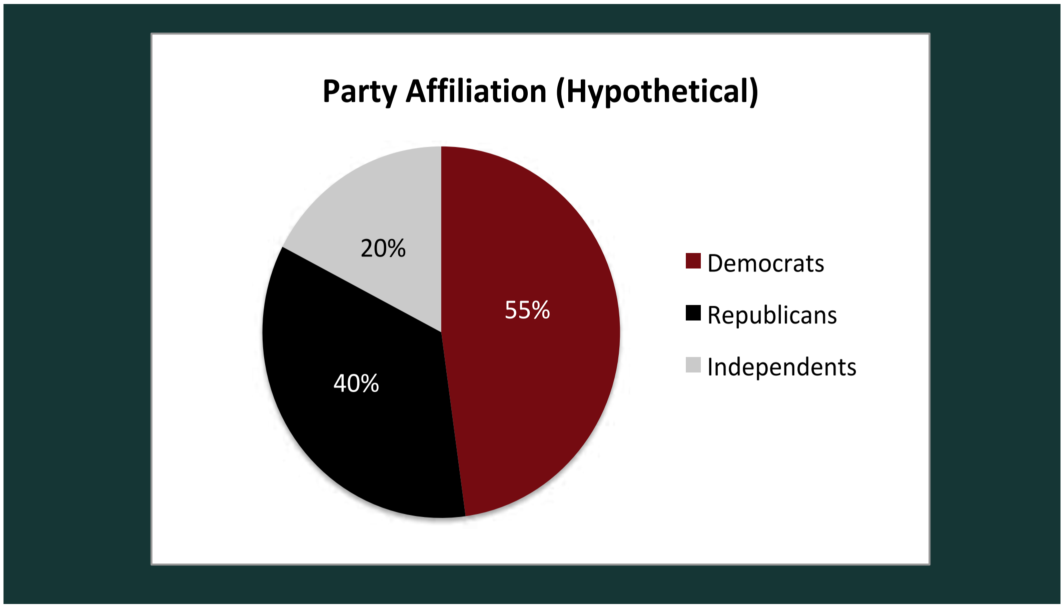

Chart 4.11

4.12.5 Review - Analyzing Expert Reports

In addition to identifying misleading charts, there are several factors to consider when assessing an expert's report or testimony.

1. Does the report rely on random sampling of the population? Is the sample a reasonable representation of the population?
2. Does the report contain unwarranted claims about causation? For example, reporting cause and effect outcomes from observational data such as correlations.
3. Do there appear to be spurious correlations which defy common sense?
4. Is there bias either in sample selection or data collection? Has there been bias in reporting such as only reporting on the results that were significant?
5. Are the outcomes reliable and valid? Did the tools used in the research measure what they were intended to measure?
6. Has complete data been provided? For example, the median, the mode and standard deviation should be provided along with the mean in measuring central tendency. Error rates should also be provided.
7. Are outcomes powerful? That is, is there a large effect size and a low probability that the outcomes happened by chance? The larger the sample size, the more powerful will be the outcomes.
8. If the survey or poll data is being used, what is the response rate? Are the questions competent to obtain the data desired?
9. Has probability been estimated using a frequentist or conditional probability process?
10. Does the report cite articles published in peer reviewed journals to support the results?

Given the large amount of statistical information contained in expert reports, as well as in the daily lives of the general society, the ability to be a competent consumer of scientific reports is challenging. Effective critical review of scientific information requires vigilance, and some healthy skepticism.

4.13 Endnotes

1. Confounding variables are comparable to intervening causes in negligence cases.

2. Examples of settings include individuals undergoing pre-trial assessment; entering a substance abuse program; criminal defendants entering please of not guilty by reason of insanity, parents going through child custody evaluations; candidates for spine surgery and for bariatric surgery; pre-employment screening for law enforcement officers, and more.

3. Yoseph S. Ben-Porath, *Addressing Challenges to MMPI-2-RF: Questions and Answers,* 27 ARCHIVES OF CLINICAL NEUROPSYCHOLOGY 691-705 (November 2012) https://doi.org/10.1093/arclin/acs083

4. An archival study collects data from records. In the study above, the records were court documents. Of course, the study assumes that the court records are accurate.

5. Derek Wood & Sir David Spiegelhalter, *Statistics and Probability for Advocates*: Understanding the Use of Statistical Evidence in Courts and Tribunals, THE INNS OF CT. COLLEGE OF ADVOCACY & THE ROYAL STATISTICAL SOC. (2017) https://www.statsref.com/ICCA-RSS-guide.pdf,

6. TYLER VIGEN, SPURIOUS CORRELATIONS, http://tylervigen.com/spurious-correlations. Reprinted with permission.

7. Although regression analysis does affect predictability, it is not a probability analysis.

8. The saga of the MMPI from the 1940s until the present is an illustration of ongoing validation research to establish that this instrument can completely and accurately measure its domain, the personality characteristics of American adults.

9. R. Joober, et al., *Publication Bias: What Are the Challenges and Can They be Overcome?* 37 J. PSYCHIATRY & NEUROSCIENCE 149-152 (2012).

10. *Duke University pays $112M to Settle Faked-Research Lawsuit*, NBC NEWS (March 26, 2019), https://www.nbcnews.com/news/amp/ncna987316

11. Fiona Godlee, et al., *Wakefield's article linking MMR vaccine and autism was fraudulent*, 342 BRITISH MED. J. c7452 (2011), https://www.bmj.com/content/342/bmj.c7452

12. H. GARB, ET AL., WHAT'S WRONG WITH THE RORSCHACH?: SCIENCE CONFRONTS THE CONTROVERSIAL INKBLOT TEST, (Jossey-Bass, 1st ed. 2003) (in which the authors review 50 years of research on the Rorschach). See also J. Hunsley & J. M. Bailey, T*he clinical utility of the Rorschach: Unfulfilled promises and an uncertain future*. 11 PSYCHOLOGICAL ASSESSMENT 266-277, (finding that there is currently no scientific basis for justifying the use of Rorschach scales in psychological assessments), http://dx.doi.org/10.1037/1040-3590.11.3.266. A. I Rabin,. *Statistical problems involved in Rorschach patterning.* 6 J. CLINICAL PSYCHOLOGY 19-21 (1950), (Finding that there are at present no statistical techniques which are really appropriate for analyzing Rorschach patterns), http://dx.doi.org/10.1002/1097-4679(195001)6:1<19::AID-JCLP2270060106>3.0.CO;2-1

13. *See e.g., People v. Jones*, 57 Cal. 4th 899 (2013) in this appeal in a death penalty case, the California Supreme Court left open the question of whether or not an expert's opinion based solely on the Rorschach test would be admissible under *Kelly*, but affirmed the trial court's decision to admit the expert's testimony anyway based on his long history of training and experience in Rorschach testing. *See also Lefkowitz v. Ackerman*, No. 2:16-CV-00624, 2017 WL 4237068 (S.D. Ohio Sept. 25, 2017), in which a malpractice claim against a forensic psychologist was brought in part due to his use of "improper and subjective methods as the Rorschach Test, which is no longer generally accepted in the field of forensic psychology." The case was dismissed on causation grounds since the results of the Rorschach were so intertwined with other psychological testing that it was impossible to tell what parts of the evaluation were based on that test, and the direct effect on the plaintiff's loss of time with his children.

14. *See e.g., Ramona v. Superior Court*, 57 Cal. App. 4th 107 (1997), directing the trial court to exclude the recovered memory testimony of his daughter in his successful suit of her therapist for defamation. *See also Hungerford v. Jones*, 722 A.2d 478 (1998), allowing a father to sue his daughter's social worker for negligence. *See also Sawyer v. Midelfort*, 595 N.W.2d 423 (1999), allowing a father to sue his daughter's therapist for negligence based on recovered memory.

15. Korsakoff's Syndrome is a disease of the brain associated with prolonged alcohol use affecting memory and producing symptoms similar to dementia. See Nat'l Inst. Neurological Disorders & Stroke www.ninds.gov

16. A.L. Wilcox, *Invited Commentary: The Perils of Birth Weight – a Lesson From Directed Acyclic Graphs*, 164 AM. J. EPIDEMIOLOGY 1121-1123 (December 1, 2006).

17. Sonia Hernandez-Diaz et al., *The Birth Weight "Paradox" Uncovered*, 164 AM. J. EPIDEMIOLOGY 1115-1120 (December 1, 2006) https://doi.org/10.1093/aje/kwj275

18. STATISTICS HOW TO: STATISTICS FOR THE REST OF US, MISLEADING CHARTS AND GRAPHS, REAL LIFE EXAMPLES. https://www.statisticshowto.datasciencecentral.com/misleading-graphs/

19. *Id*.