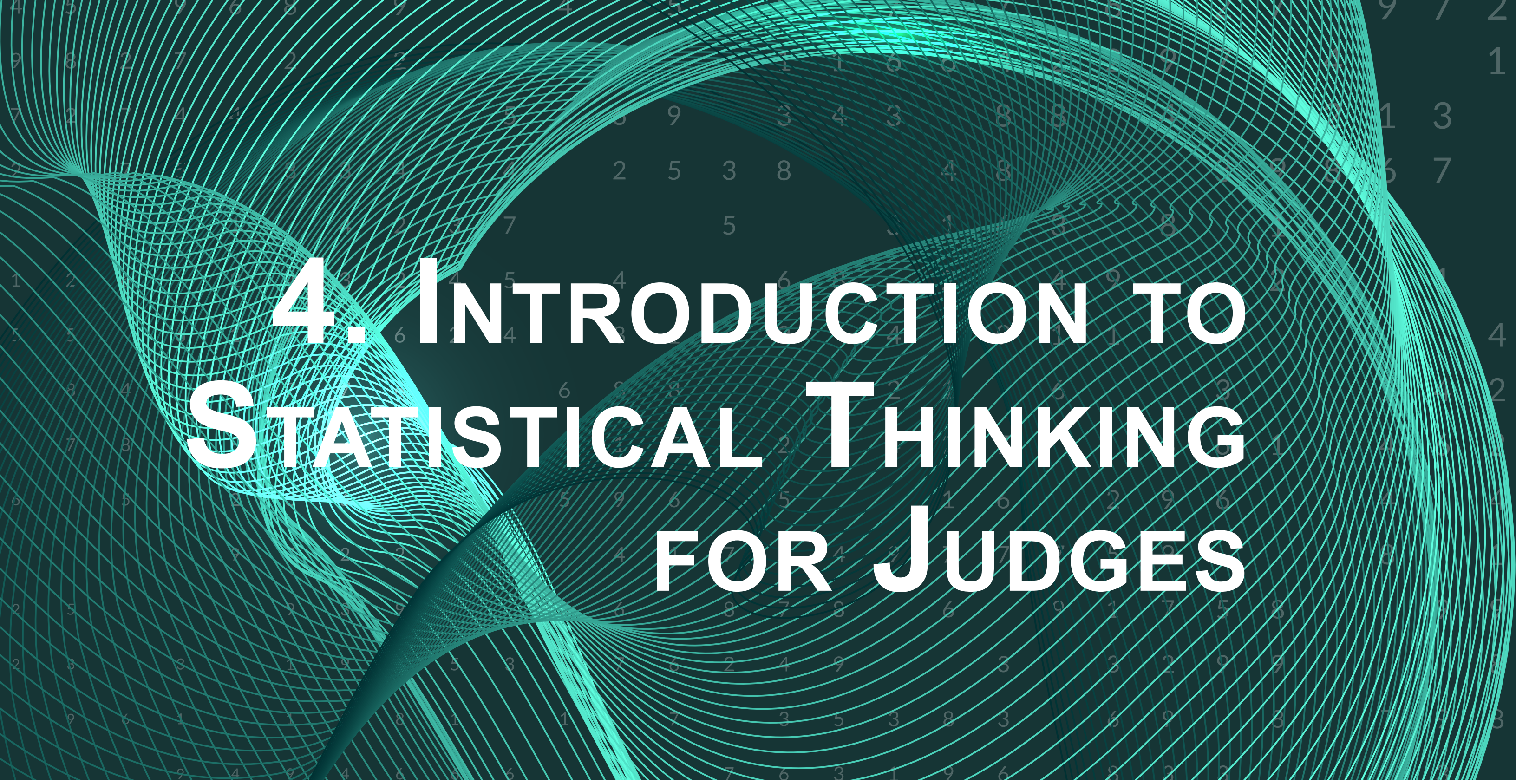

4. Introduction to Statistical Thinking for Judges

Sections 4.1 - 4.9

Eryn Blagg[1] &
Alicia Carriquiry,[2,3] PhD

4.1 What is Statistics?

Statistics is the science of collecting, analyzing and interpreting data. Statisticians develop, test and implement tools to display empirical data, to extract information from those data, and more generally, to draw inferences about populations using samples drawn from populations. Data arise in every discipline, so statistical methods are useful to almost everyone who wishes to use data to answer questions. The civil and criminal justice systems are no exceptions. Questions of interest might include:

- What was the time of death of the victim?
- Did the suspect's shoe leave the print at the crime scene?
- Are hiring practices in company X discriminatory?
- Is the defendant the father of the child?

> *Statistics is the science of collecting, analyzing and interpreting data.*

These are just a few examples of the many questions that may arise in court, and for which the judge or a jury must produce an answer. Ideally, the answer is accompanied by some measure of *uncertainty* to reflect the confidence of the juror or judge on the answer. The idea of uncertainty plays a critical role in statistics. Uncertainty arises when we do not know the outcome of some process, yet decisions must be made in the face of uncertainty. Evidence may suggest the defendant committed the crime, but unless we were there to see the crime in real time, there is always some chance someone else may have be guilty instead. Statistics provides the means to address

[*] Some of the examples in this chapter are reproduced with permission from *Statistical Thinking for Forensic Practitioners*, a set of CSAFE training notes authored by Hal Stern, Naomi Kaplan-Damary, and Alicia Carriquiry.

[**] We are grateful for the constructive and helpful suggestions from our editors. We also wish to thank Joy Lyngar from the NJC for her guidance and for providing the opportunity to contribute to the *Science Benchbook for Judges*.

these types of questions and to produce an estimate of confidence around the answers. In order to do this, statisticians make use of mathematical and computational tools.

The rest of the chapter will expand on some important statistical topics. We start by defining some basic ideas of statistics, including populations and samples, then move on to talking about the different types of data that may arise in the context of legal proceedings. Next, we discuss various approaches to collecting data and talk about the design of studies, including how those factors affect the type of inference that can be drawn. Following that, we talk about describing and summarizing sample information, and present some key ideas associated with statistical inference, or making conclusions about a population using information from a sample. We finish by briefly discussing how to assess the quality of the data arising from a sample or from a study, and of the study itself. We close with a summary of key issues.

4.2 Probability, Statistics and Data

4.2.1 Populations and Samples

When data are used in the courtroom, it is important to establish where the data came from. The data may come from a *sample* of a population, or in rare cases, may include the entire population. The latter happens infrequently because, unless a census is conducted, the complete population of interest is rarely known. Depending on whether data comprise the population, or only a subset of the population (sample), the statistics and statistical analysis that is used are different. Thus, the first step is to determine whether we are working with a population or with a sample.

A *population* is the universe of objects of interest. In the legal context, a population may be every promotion decision made by every manager of a large employer in California, it may be the outsole pattern of every shoe sold in the United States last year, or perhaps every baggy containing some white powder in a container arriving from Asia. Sometimes, the population of interest is a sub-set of the larger population. For example, we may be interested in promotions only among entry-level employees in California. It is important to clearly state what is the population of interest in every case.

A *sample* is a set of objects obtained from the population that are available to us for study. In practice, populations can be large, and it can be impractical, or even impossible, to take measurements on each population object. In the case of the container or baggies, we may select a small number upon which to carry out a chemical test. The goal of a sample is to represent the population without having to test every single baggy in the container.

The tools of *probability* allow us to anticipate what we might observe in the sample. For example, if we know a dice is fair, we can anticipate we will obtain an even number in about half of the rolls. In other words, if we know the probabilities associated with the various possible outcomes from the population, we can *deduce* what we will observe in a sample from the population. The tools of statistics on the other hand, are *inductive*, i.e., we make inferences about the population using what we observe in a sample from that population. For example, we infer that among Caucasians, the gene allele 15 at locus D3S1358 is present on the chromosome of 24.6% of the population.[4] Of course, a genotypic test was not implemented on every possible Caucasian person in the world to reach this conclusion. Instead, this inference was based on the information obtained from a relatively small (in the thousands) sample of Caucasian persons whose DNA was analyzed.

4.2.2 Probability

Probability is invoked often in court cases, from the probability the company in question is discriminatory to chances the gun found on the suspect was the source of the bullets from the crime scene. Different types of probability statements have different interpretations. It is important to distinguish what kind of probability statement is being made in order to make sure the interpretation is correctly presented. When dealing with probability statements from an expert witness, it is imperative to determine if the interpretation matches the relationship being addressed.

Probabilities describe how often an event is likely to occur; *odds* are a ratio of these probabilities. When working with probabilities it is important to determine if the event is conditional upon another event. *Conditional probabilities* give us a way to calculate probabilities of an event “A” given that another event “B” has occurred. For *independent events*, the probability of A is unchanged whether or not B occurs, whereas for dependent events, *Bayes' theorem* can be used to switch between event “A” given “B” and event “B” given “A”. In this section, we will describe the different types of probability statements and the interpretation that corresponds to each.

> *Probability is the mathematical language of uncertainty. The probability of an event is a number between 0 and 1 that reflects the likelihood that an event occurs.*

4.2.2.1 What is probability?

Probability is the mathematical language of uncertainty. The probability of an event is a number between 0 and 1 that reflects the likelihood that an event occurs. Examples of events include:

- A fair die lands on a 6.
- A randomly chosen baggy from the container contains fentanyl.
- The Chicago Cubs win the World Series.

An event with probability of 1 always occurs. An event with a probability 0 never occurs. In most cases, the event probability is somewhere in that interval, i.e., between 0 and 1.

4.2.2.2 Where do probabilities come from?

There are different interpretations of probability, but the two most widely accepted are what are known as the *long-run frequency* interpretation and the *subjective belief* interpretation. The long-run frequency interpretation, as the name suggests, establishes the probability of an event by the frequency with which the event occurs in a very large number of trials. For example, if we toss a fair coin a million times, the probability of heads is estimated as the proportion of tosses resulting in a head. This frequency interpretation of probability is reasonable when the "experiment" (e.g., the coin toss) is repeatable. The *subjective belief* interpretation refers to the expected likelihood an event will occur. This interpretation can be applied in those cases where repeating a trial is not possible. As an example, we might believe the Cubs have a 0.7 chance of winning the World Series. Subjective beliefs can be informed by empirical data or other information. Since there is only one 2021 World Series, we cannot use replication, and must use other methods for determining subjective beliefs. My personal probability the Cubs will win the World Series may be based on the results of pre-season games, on my knowledge of the players that the Cubs have and on information about team injuries. In this sense, the term "subjective" does not necessarily mean "arbitrary." When an expert in court presents a subjective probability, he or she should also describe the information used to establish that probability.

4.2.2.3 Probability and odds

We often talk about the odds of something occurring. For example, the odds we will win the lottery are negligibly small. The odds two DNA samples will match if they belong to the same person are very high. *Odds* are simply ratios of probabilities; they are not probabilities. The odds in favor of event “Y” is defined as the probability that event “Y” occurs divided by the probability that event “Y” does not occur.[5]

Similarly, the odds against “Y” is defined as the ratio of the probability that “Y” does **not** occur to the probability that it does. If we are given the odds for or against an event, then we can derive the probability of the event.

Although probabilities and odds are related to each other, their interpretation is different. For example, if the probability of an event is 0.5, we have odds 0.5/0.5=1. As the probability increases, the odds get larger and larger. For example, for an event with probability 0.99, the odds in favor of the event are 0.99/0.01=99.

4.2.2.4 Conditional probability

The concept of *conditional probability* arises often in the legal context but must be distinguished from the concept of probability as described above. Consider a pathologist trying to determine how long ago a victim died. Based on the body’s temperature, the pathologist concludes the victim died between 18 and 20 hours ago, with probability 0.9.[6] The detective tells the doctor that the victim appears to have been killed outside, and the ambient temperature was 30 degrees Fahrenheit at the time the body was found. Would the pathologist revise the probability? Given the body was outside, it is likely its temperature decreased faster than the pathologist had estimated earlier when there was no information about the body’s location. With the additional information, the pathologist may now decide the probability the victim died between 18 and 20 hours ago, given that the body was outside, is no larger than 0.2.[7]

When dealing with conditional probabilities, it is important to be certain we have accounted for the relevant events.

Conditional probability changes the population to which we refer. When the doctor did not know where the victim was found, the relevant population was all cadavers. With the additional information, the new relevant population is only those cadavers subject to temperatures around freezing. When dealing with conditional probabilities, it is important to be certain we have accounted for the relevant events. From the example above, the probabilities of an event occurring changes drastically depending on the conditioning event, i.e., the temperature where the body was located. It is usually the case that inverting conditional probabilities leads to different results. Because of this, we need to be careful about both identifying the population of interest and including the relevant information about the population.

Another example to illustrate the point: suppose we have 300 pairs of 9mm bullets, 200 of which were fired from the same gun and 100 of which were fired by different guns. For each pair, we measure the number of consecutively matching striae or CMS (a quantitative method of describing an observed pattern match) and find that:

	Number of CMS								
	1	2	3	4	5	6	7	8	Total
Same Gun	0	5	11	21	32	40	49	42	200
Different Gun	6	12	29	32	10	9	2	0	100
Total	6	17	40	53	42	49	51	42	300

Table 4.1 Number of CMS by source of bullets[8]

The probability of observing 6 CMS in this study is Pr (CMS=6) = 49 divided by 300 which is 0.16. However, when we look at the conditional probability, the probability that CMS is 6, given that bullets were fired by the same gun is higher, viz. 40 over 200 or 0.20. In the first case, the population of interest were all pairs of bullets; in the second case, we restricted interest to the population of pairs of bullets fired by the same gun.

Conditional probabilities occur often in the legal and forensic context, and it is important to differentiate them from other forms of probability statements.

The inverted conditional corresponds to a different question: Given I observe that CMS is equal to 6, what is the probability the bullets were fired from the same gun? Now we have 40/49= 0.82. This is one of the reasons why it is important to ascertain the specified population to be used, based on the question being asked. We will see later in this section that the “likelihood ratio” is a ratio of two conditional probabilities, but for now realize conditional probabilities occur often in statistics and it is important to differentiate them from other forms of probability statements.

4.2.2.5 Conditional probability and independence

Sometimes, additional information does not change the probability of an event. This is referred to as *independence*. Suppose in addition to CMS, we also know the firearms examiner was born in Texas. This additional piece of knowledge does not change the probability of observing 6 CMS given that the bullets were fired by the same gun. We say that the place of birth of the examiner is independent of the number of CMS, so the probability of observing a 6 is still 0.16. as before. When two events are independent, the probability that both events occur simultaneously (or jointly) can be computed using the *product rule*: if events A and B are independent, then their *joint probability* is the probabilities of the two events multiplied together.[9]

A well-known example of independent events in the legal and forensic context is the independence of DNA markers located on different chromosomes. This is one of the reasons there is a low probability that two humans share the same alleles at the loci typically used in forensic genotyping. To illustrate, consider two DNA markers, D3S1358 and vWA. Assume that the sample from a crime scene has alleles 16,16 and 15,17 at each locus, respectively. What, then, is the probability of that particular genotype at the two loci? From published allelic frequency tables,[10] we know the probability that a Caucasian person is homozygous 16,16 at the D3S1358 locus is 0.0943 and the probability a Caucasian person has genotype 15,17 at the vWA locus is 0.0866. The probability that a Caucasian person will match the crime scene sample at both loci can then be calculated. Using the product rule, and our knowledge of independence of DNA markers located on different chromosomes, we have 0.094 x 0.0866 or 0.0082. Thus, only about 8 in 1000 Caucasian persons would be expected to match the crime scene sample at both markers. In forensic DNA analysis, scientists examine the genotype at 21 loci. Then, to compute the probability of a match, they apply the product rule using the 21 published allelic frequencies corresponding to the observed genotype. This is how negligibly small match probabilities, perhaps in the order of 1 in a trillion, are obtained and why DNA evidence is so probative.

4.2.2.6 Conditional probability and Bayes' Theorem

When events A and B are not *independent*, typically the probability of event "A" given event "B" is not the same as event "B" given event "A". *Bayes' Theorem*[12] tells us how to "invert" the conditional and go from the probability of event "A" given event "B" to the probability of event “B” given “A” in such instances. If the probability of event “A” and the probability of event “B” are known, then one can find the probability of event “A” given “B” by taking the probability of “B” given “A” times the probability of “A”, then dividing by the probability of “B”. More generally, it allows us to use information in a sample to make inferences about a population given that we know the probabilities of both “A” and “B”.

For example, assume you leave work one day having a sore throat and a headache. You remember that last week one of your coworkers had strep throat. Does this mean you now have it? You know that 95% of people afflicted with strep throat have both a sore throat and headache as symptoms. After some “Googling” you find that about 5% of people in your location get strep every year, but also that about 30% of people experience sore throats and headaches without suffering from strep throat. Using this knowledge, and Bayes’ Theorem, you can find the probability that, given you exhibit the symptoms, you have strep throat. That is, taking the probability that someone has a sore throat and a headache, given they have strep, multiplied by the percentage of people in your location who get the virus each year, then dividing by the percent of people who have headaches and a sore throat without being sick, we get the probability that you have strep given you have the symptoms: 0.95 x 0.05/0.3 = .158. So, the probability you have strep, given you have the symptoms, is about 16%.

Conditional probabilities get reversed in Court so often that this mistake has a name: the prosecutor’s fallacy.

Note that the probability of strep throat given the symptoms (16%) is very different from the probability of symptoms given strep throat (95%). Also note, that in order to go from probability of symptoms given strep throat to probability of strep throat given symptoms, we need two additional pieces of information: the background probability of strep throat and the background probability of the symptoms in the population.

Conditional probabilities get reversed in Court so often that this mistake has a name: the *prosecutor’s fallacy*. The prosecutor’s fallacy occurs when the following two probabilities get equated: the probability of observing the evidence if the suspect is innocent, and the probability that the suspect is innocent given the evidence we have observed. For example, suppose that a witness reports seeing a blond woman with a ponytail and with only one arm at the crime scene. Further, suppose that the prosecution argues that only one in one ten thousand women in the surrounding areas is blond, wears a ponytail and has a single arm. Even if the suspect is a blond woman who is missing an arm, it is still incorrect to conclude that the probability that she is not the criminal is only one in ten thousand.

4.3 From Probability to Statistical Inference: Collecting Data

In order to develop, test and validate instruments and other technologies, or to assess the value of forensic evidence in general, it is necessary to collect *data*. The type and quantity of data we collect determines the type of information we can extract from the data, so it is important to think carefully about the provenance of the data upon which we rely. Statisticians have important knowledge to contribute when it comes to data collection. In this section, we describe two fundamental approaches for data collection—experimentation and sampling—and discuss the uses and limitations of the resulting information. Before we address study design issues, we first talk about the types of data that can be collected.

The two types of data are qualitative and quantitative. Qualitative data refers to data that have different categories; these can be ordinal or not. Quantitative data describes numeric data on a continuous or discrete scale. Depending on the type of data being used, different statements and analysis can be made. Thus, when working with data, the first step is always to determine the type of data we have. In order to collect data, either experimental or sampling studies must be performed. The most common goals of both types of studies are to collect a random and representative sample, even if the mechanisms are quite different. If those two goals are not accomplished, then one cannot generalize about the population from the data. With sampling, there are always some shortcomings, which may skew any results that come from the data.

4.3.1 Types of Data

Statisticians distinguish between various types of data:

- *Qualitative data* represent attributes of an object such as gender, color, zip code or genotype. We distinguish between two types of qualitative data:
    - *Categorical*, where there is no ordering of the categories. An example is blood type, which have values A, B, AB, or O.

    - *Ordinal*, where there is a natural ordering of the categories. An example is the response to a question in a judicial survey that may take on values between 1 and 5, with 1 being "strongly disagree" and 5 being "strongly agree". The assignment of ordinal categories is sometimes arbitrary. It is important to realize that, although ordinal categories are numeric, one cannot take the average, i.e., the *mean*, and assign it meaning. The mean of the responses to two questions in a survey, one being 1=strongly disagree, and the other being 5= strongly agree, in a survey does not mean the judge has average views, but rather that the judge has very different responses for the two questions.

- *Quantitative data* typically arise as the result of some measurement process and is expressed in numerical values. These values normally have units as well, such as inches, years, or miles. Again, we distinguish between two types of quantitative data:

    - *Discrete*, where the measurements can take only integer values, i.e., whole numbers. Examples include the number of consecutively matching striae or CMS, or the number of children in a family.

    - *Continuous*, where the measurements can take on an infinite number of different values in some range. An example is the concentration of some chemical element in a glass fragment.

Different types of data call for different types of statistical analyses, as we will discuss later. Before we think about statistical analyses, we briefly discuss the two fundamental data collection paradigms.

4.3.2 Collecting Data Via Sampling Studies

Unless we are dealing with a small population of interest, we must use sampling, because it is typically too costly, or too time consuming, to study the entire population. Sampling simply consists in selecting – in some principled way – a sub-set of the objects in the population. The idea behind sampling is simple: We attempt to draw a sub-set of the population that looks enough like the population itself, so that the results of statistical analysis using measurements from the objects in the sample are generalizable to the population itself.

There are two major types of sampling approaches: 1) Those based on some random selection of the objects in the population, and 2) Those that select objects using some systematic (non-random) approach. The samples that result from random sampling are called *probability samples*. There are different types of probability samples. Three commonly used sampling methods are:

Sampling simply consists in selecting – in some principled way – a sub-set of the objects in the population.

- *Simple Random Sampling*: Characterized by the idea that every member of the population has an equal chance of being selected for the sample.

- *Stratified Random Sampling*: Often large populations will be made up of smaller homogenous groups. We may want to make sure each group is represented in the sample. For a population which can be divided into strata, a stratified random sample is a sample which is obtained by drawing random samples from each stratum. Often, the number of items sampled from each of the strata corresponds to the size of the stratum. When sampling glass fragments for analysis, for example, we might stratify glass into architectural, automotive, and other.

- *Cluster Sampling*: Similar to a stratified random sample, a population can be separated into clusters. A cluster sample is obtained by randomly selecting a number of clusters and sampling each member in those selected clusters. Population surveys often use cluster sampling. For example, a city block is a cluster and a resident in every household in the block is then included in the sample. In the legal context, the population may consist of 1000 containers arriving from abroad in a month, each filled with boxes supposedly containing stuffed toys. Each container is a cluster, and a reasonable sampling approach might be to select a subsample of the containers and from each, inspect every box.

Non-probability samples are used extensively in qualitative research and the social sciences. They can be useful in studying some social phenomenon in depth. They are also used when implementing a *bona fide* random sampling method is impractical, as in the case of sampling populations that do not wish to be found such as drug users or undocumented migrants. Three commonly used approaches for non-probability sampling are:

- *Convenience sampling:* Occurs when the investigator selects objects from the population that is most handy. An example would be a study where we sample only co-workers, or patrons in a mall.
- *Snowball or network sampling:* These types of samples are useful when members of the population of interest do not identify themselves as such. This might include, for example, users of illegal substances, under-age drinkers, or HIV-positive persons. Network sampling consists of finding one or a few members of the population and then using their connections to continue building the sample.
- *Purposive sampling:* In this type of sampling, the data collector selects the objects to be included in the study using some selection criterion. This type of sampling is sometimes implemented when the attribute to be studied is very expensive to measure and the researcher cannot afford to measure it in a large sample. An example might be measuring the effect of exposure to a pesticide on the functioning of the brain of persons exposed. In this type of study, the researcher may select a small number of agricultural workers for example in a limited number of farms known to have low, medium and high exposure to the pesticide of interest.

It is always important to understand how the sample was selected in order to be sure that the statistical findings obtained from the sample are generalizable, and if at all, to the population of interest. For example, suppose that in a study of gun ownership in the US we purposively select 100 counties from which to collect information. Even if the individuals sampled within each of the counties comprise a probability sample, results will be generalizable only to the 100 counties included in the study. If instead the 100 counties are also randomly selected from the 3,141 counties in the US, then results are generalizable to the entire country.

It is always important to understand how the sample was selected in order to be sure that the statistical findings obtained from the sample are generalizable, and if at all, to the population of interest.

Probability sampling is the gold standard and should be used whenever we wish to make statistical inferences about the population from which the sample was drawn. However, not all probability samples allow unbiased and reliable inference about the population. Probability samples are obtained by applying some form of random selection of items from a population. Regardless of the selection method, the important idea is that each member of the population has a known probability of selection. In a simple random sample, defined above, each population item has a probability of selection that is equal to 1/N, where N is the size of the population, and all possible samples of the same size also have a known and equal probability of selection. In the usual classroom example, if I have a bag with 100 identical balls labeled 1 to 100, the ball numbered 57 has a probability of selection of 1/100.

Probability sampling is the gold standard and should be used whenever we wish to make statistical inferences about the population from which the sample was drawn.

For the sample to be representative of the population, a simple random sample may need to be very large. Suppose that we wish to test a new risk assessment tool for predicting recidivism. The tool's performance is likely to depend on individual attributes, including gender, race, age, and offense type. If we consider two genders, five races, four age categories and six different offense categories, that results in 2 x 5 x 4 x 6 = 240 different combinations, some of which may be rare. In order for the sample to include at least a few cases in each of the categories of interest, so that it is representative of the population, the sample size would need to be enormous.

To illustrate, assume a rare category comprises 0.1% of the population of criminals. To include at least one case, the sample would need to be at least of size 1000, and even then, there is a sizeable chance that the simple random sample would not include this combination of attributes. This is a case in which a more effective random sampling approach might be a stratified random sample consisting of strata made up of the different combinations of criminals by sex, age, etc., and then randomly selecting a certain number of cases from within each of those strata. Of course, the resulting sample would not be representative of the population because it would include a higher proportion of the rare cases than exist in the population. But if the selection probability of each sampled person is known, then statisticians can construct survey weights for each sampled person or object so that, after weighting, the sample is once again representative.

The biggest difference between probability and non-probability sampling is that, in probability samples, each sampled object has a known probability of selection, whereas in non-probability sampling, the probability of selection of each item in the population is unknown. In fact, non-probability sampling is often used when we do not even know the size or the composition of the population of interest. Consequently, probability samples allow us to make inferences about the population from which the sample was drawn, but non-probability samples most often do not. There are many different approaches for selecting random samples from large, complex, populations, but as long as the design of the sample or survey is known and the probability of selection of each population item is also known, it is always possible to ensure that the results of analyzing the sample measurements will generalize to the population.

4.3.3 Potential Shortcomings of Sampling

Probability samples are not without issues. Some of those issues include:

- *Incomplete coverage/Undercoverage*: This occurs when a proportion of the population is not represented or is underrepresented. A famous example of this was the political survey carried out by the Gallup organization when Dewey and Truman were running for President of the United States in 1948. Gallup used a method called *quota sampling*, where the idea is to create a sample that equals the population in terms of proportion of genders, races, rural/urban living and so on. Inevitably, some population attributes that affect voting preferences are left out. Famously, Gallup predicted Dewey would defeat Truman by a large margin, but Truman ended up winning.

Image 1: Truman showing the headline of the Chicago Tribune that, following Gallup's forecast, had mistakenly anticipated a win by Dewey[13]

- *Self-selection bias:* Samples that consist of participants who self-select for the survey/study are typically not representative of the population. Self-selection occurs when individuals have a choice of whether to participate in the survey. Examples include surveys carried out by a company such as Survey Monkey on behalf of a client.

- *Non-Response Bias*: People selected for the sample may decide not to participate. Well-designed surveys aim for a sample size large enough to guarantee desirable precision of sample estimates. When non-response is higher than designers anticipated, the resulting estimation error is larger than desired. If, in addition, the non-response is not uniform across all respondent types, then the estimates obtained from the sample can be biased, in addition to exhibiting high error. As an example, suppose we are surveying crime labs to find out about their backlog in cases. The sample was designed so it would be representative of the population of crime labs of a certain size. Now, imagine only the small sized labs respond to the survey. The likely outcome is we would be underestimating the size of the backlog because the sample included no medium-sized or large labs.

- *Response bias:* Response bias occurs because, although a subject may agree to respond to a survey, he or she may not always tell the truth. For example, a worker might not tell her boss how she feels about his actions because of fear of how it may impact her job.

This is not an exhaustive list of the problems that may afflict samples. However, it does include the most commonly observed poor sampling practices, and issues that one should be aware of, as they can strongly impact the quality of the findings obtained from the selected data. One last comment is that when samples are drawn for the purpose of eliciting a political opinion, they are often called "polls." This is just another name for a sampling study or survey. Just like any other survey, polls can be well designed and conducted, or not.

4.3.4 Observational Studies versus Randomized Experiments

Statisticians and other scientists may collect data to compare "treatments" in order to answer a question or test a theory. The two most common types of designed studies are "observational" and "randomized" studies.

Observational studies are studies in which the researcher has no control over the experimental units, or what/who is receiving the treatments. Observational studies are seen frequently when looking at the health effects of exposure to a chemical or the consequences of implementing a new policy. In this type of study, we attempt to establish the effect of exposure to a substance by sampling individuals from populations that were, and were not, exposed then measuring the prevalence of the health outcome of interest. This method of experimentation has its limitations, however. There may be factors contributing to a disease other than exposure rates. For example, the exposed population may live near a polluting site, and consequently be poorer and have worse access to health providers, than those

Observational studies are studies in which the researcher has no control over the experimental units, or what/who is receiving the treatments.

who live in “clean” areas. As a result, the type of inference that can be drawn from observational studies is limited. We might find, for example, that higher exposure is associated with higher prevalence of the disease, but we cannot establish a *causal relationship* between the two.

Despite this limitation, observational studies are often used when traditional studies are not an option because, ethically or logistically, we would be unable to assign individuals to treatments we know (or suspect) will be harmful to them. For example, it would be unethical to assign participants to a smoking group (if they do not already smoke) to study the relationship between cigarettes and cancer. Similarly, if we wish to understand the relationship between race and probability of a traffic stop, it is not logistically possible to reassign, i.e., change, a person’s race.

Randomized studies are the gold standard of experiments.

Randomized studies are the gold standard of experiments. In a randomized trial, participants are randomly assigned to treatments. The random assignment ensures that all other differences between participants, both observed and unobserved, are balanced across treatment groups. In this way, we can be confident that the only differences between participants across groups is the treatment itself. As a result, randomized trials are essentially the only type of study that permit establishing a causal relationship between a factor and an outcome. An example of a randomized study might be a black-box study, where the “treatments” consist of different levels of quality of latent fingerprints, and where participating examiners are randomly allocated a latent print for analysis.

As in the case of surveys and sampling, the size of the study is directly proportional to the precision of the estimates obtained from the data and with the power of the study to detect differences between treatment groups. In this regard, intuition is accurate, i.e. the more information we have the better we can more accurately describe what the data are showing.

4.3.5 Describing Data

Once you have collected data, from either a survey, an observational study or a randomization study, the next step is to describe the data. There are two common ways of describing data: graphically or numerically. Each different type of data requires its own type of visualization, and of numerical descriptions. When these descriptions arise in a courtroom, it is important to make sure that the graphical or numerical summaries presented are correctly matched to the datatype.

4.3.6 Graphical Displays for Describing Data

The appropriate form of graphical display used for describing a collection of observations depends on the type of data (described above) and on what we are trying to summarize.

A *bar chart* is used to look at the frequencies of qualitative variables. When reading a bar chart, the length or height of the bars show which of the categories occur most often. For example, if we were interested in looking at which crimes are most often committed in the United States, we see that larceny or theft is the most frequent crime category, while rape is the least frequent.

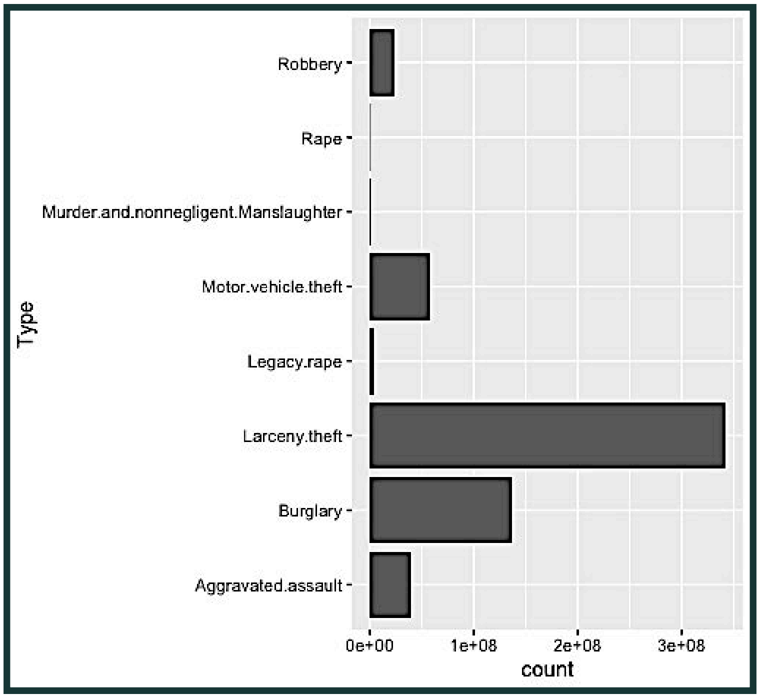

Figure 4.1: Bar chart showing the type of crimes committed in the US[14]

When the data are quantitative, a *histogram* is used. A histogram is a graphical summary of the distribution of a quantitative variable, which can be either continuous or discrete. A histogram has an X-axis, that covers the range of the variable, and a Y-axis showing the frequency at the given range. For example, a histogram displaying the discrete numbers of CMS that were shown in Table 4.1 would appear as follows:

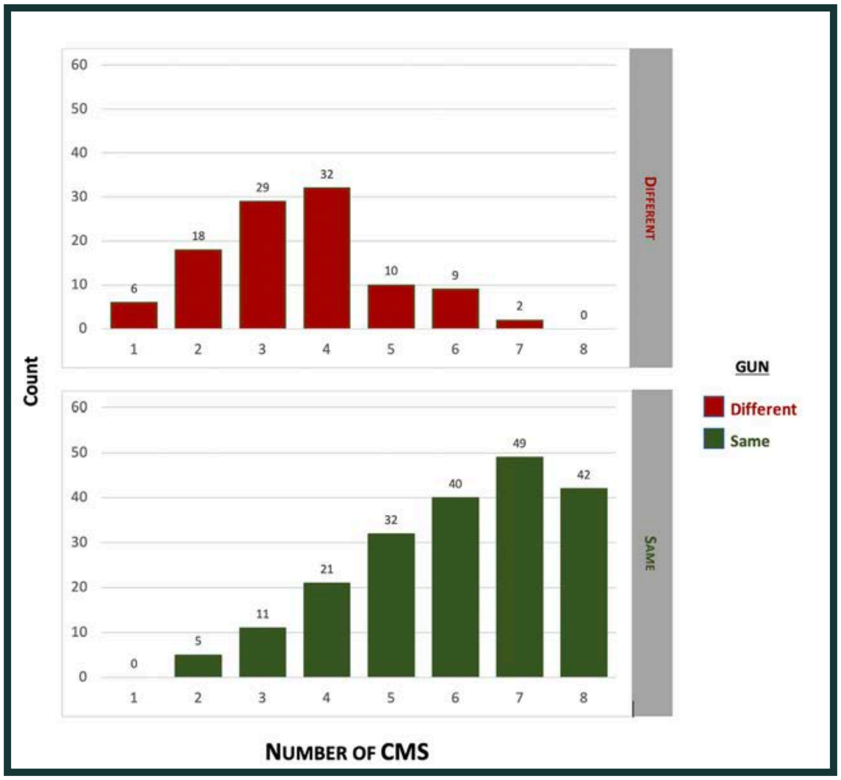

FIGURE 4.2: A STACKED HISTOGRAM OF THE CMS DISPLAYED IN TABLE 4.1

Histograms can also be used for displaying continuous measurements after we first group the measurements into bins. For example, we saw above that larceny or theft is the type of crime committed most often in the U.S, at least between 1960 and 2018. If we want to look at the distribution of larceny or theft crimes, we can draw a histogram as shown in Figure 4.3. The histogram shows the distribution of number of larceny or thefts per state and per year, as recorded by the FBI between 1960 and 2018. The highest peak of the histogram approximately corresponds to the value 10,000 to 20,000, meaning that over the 58 years reported, the most frequent number of larceny or thefts in a state was about

27 to 54 reported per day. When histograms have "tails" of different lengths, we call the distribution skewed. The direction of the tail corresponds to the direction of the skew. In Figure 4.3, we have a right skew:

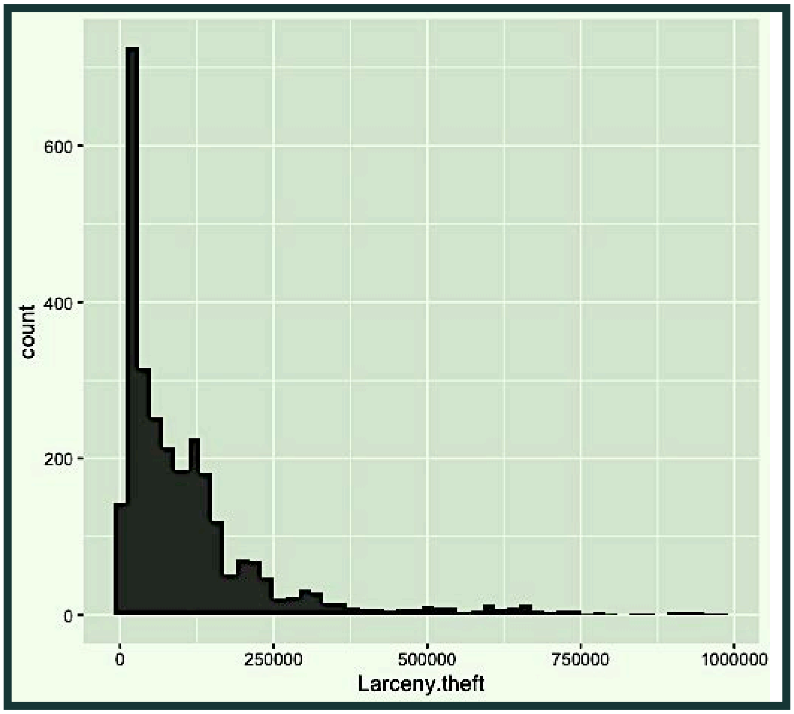

Figure 4.3: Histogram of the number of larceny theft crimes committed.

When we wish to visualize the relationship between two or more different variables, we can use a *boxplot*. For example, assume we obtain glass fragments from manufacturers A and B, both located in the Midwest. Over a range of dates, we then measure the chemical concentration of some element "Y" in parts per million. In this example, the element of interest is zirconium (Zr). A boxplot is useful for displaying the range of values of Zr by day of manufacture of the fragments. In addition, we look at fragments from the different companies.

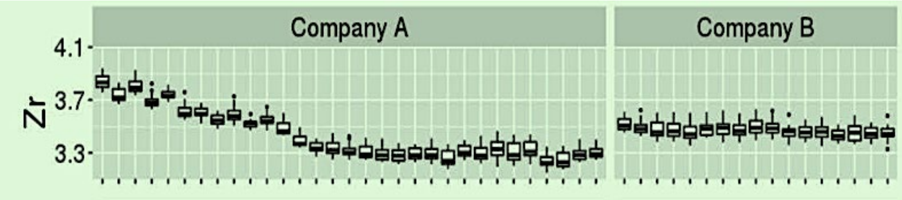

Figure 4.4: Concentration of Zr in glass panes manufactured by companies A and B over 31 days (A) and 17 days (B).

For each day, Zr concentrations were measured on 24 fragments obtained from each pane. Each small box summarizes measurements made on a different pane of glass.

Boxplots provide a lot of information: The *median* value of the measurements on each pane, is shown as the line in the center of the box: the box itself, which shows the middle 50% of the data. In addition, the dots denote outliers or unusual values. From Figure 4.4, we see that the concentration of Zr on glass produced by company A appears to decrease over time, where it looks approximately constant over time for company B glass. The *height* of the box is an indication of the *variability* in the Zr measurements within glass produced on the same day in each of the companies.

The final, most frequently used, figure to describe data is a scatterplot. *Scatterplots* display the relationship between two quantitative variables. The two variables can be associated in three different ways: the association can be positive, negative or none.

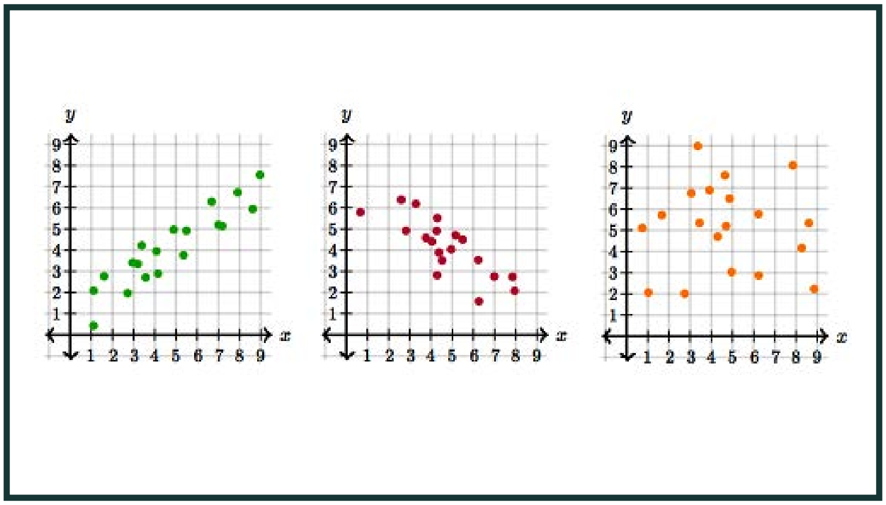

Figure 4.5: A positive association (left), negative association (middle) and no association (right)

In some applications, the variables shown on the x-axis (the horizontal axis) and on the y-axis (the vertical axis) are called the *explanatory* and the *response variables*, respectively. The scatterplots shown in Figure 4.5, are examples of good plots that show information in a concise and direct way. However, this is not always the case. Bad plots occur more often than statisticians would like to admit. Such plots convey inaccuracies and false information.

In sum, the goal of a graph is to be simple and easy to read, while still accurately conveying information. However, it is important to make sure the graphic displays accurately depict the relevant information.

4.3.7 Numeric Ways of Describing Data

Data can also be described numerically. When describing data numerically, there are two different measures used – measured of center and measures of spread.

Measures of center are measures that show where the center of a group of data points is. They include mean and median. The *mean* of a group of data points is what is commonly referred to as the "average." Mathematically, it is the sum of the observations divided by the total number of observations:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i .$$

One characteristic of a mean is that it can be affected by *outliers*, or observations that are unusual.

The mean of a group of data points is what is commonly referred to as the "average."

The other most used measure of center is the median. The *median* is the middle number in a group of observations (If you have an even number of observations, the median is defined as the mean of the two numbers in the center). Unlike a mean, outliers do not affect the median. For example, if we have a set of 15 measurements: 5, 16, 19, 24, 25, 25, 26, 30, 33, 33, 34, 34, 37, 37, and 40, the mean and the median are 27.8 and 30, respectively. If we add one more measurement equal to

100 to the data set, the median changes, to 31.5, a difference of only 1.5 units. But the value 100 is an outlier relative to the other values, and it pulls the mean up, to 32.3, or 4.7 units. The median is a more robust measure of the center of a group of numbers in that it is less susceptible to the presence of outliers. While the mean and median are not the only measures of center, they are the most often used in statistical analysis.

Measures of spread explain how much variation is in the data. Small variation implies that the observations are all concentrated around a central point, while large variation implies that the data are spread out over a large range. The *range* is the difference between the lowest and highest values in the data set. It measures total variability of the observations. In our example above, the range is 95, i.e., 100 – 5. The range is highly affected by outliers, as it is the maximum minus the minimum values in the dataset.

Quartiles divide the observations into four equally sized groups, and the *interquartile range (IQR)* is defined as the middle 50 percent of the data. This measure of spread allows us to see the variability of the data without the extreme values; thus, without the impact of outliers.

When the median is used as the measure of center, IQR and range are the most often used measures of spread. When the mean is used, the measure of spread used is the standard deviation. The *standard deviation* squared is called the *variance* and is computed as the average of the squared distances between the observations and the sample mean. The positive square root of the variance is the standard deviation. The standard deviation is typically denoted SD or s. Mathematically, s is

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^{\,2}}{n-1}}.$$

The standard deviation is always a positive value.

Because the standard deviation contains the mean in its formula, the standard deviation of a data set is highly affected by outliers. In any case, if a data set has a high standard deviation, the data are very spread out, while a low standard deviation suggests that the data are clumped together around the mean.

When reporting statistics, it is important to report both a measure of center and a measure of spread in order to get the full picture of the set of observations. When the observations are very spread out, the mean is not a good summary of the data. Therefore, if only a measure of center is reported, it is not possible to determine whether the mean is an informative summary.[15]

Unfortunately, it is common for non-scientists to report only a mean (or a median) without a measure of spread, let alone a graphical data summary.

In addition, including a visualization of the data set, along with a numeric summary, helps with understanding other aspects of the data. For example, assume we report the mean number of fatal crashes in Iowa per year to be 648 over the last 10 years. If we then determine that in eight of the 10 years the number of crashes was below 600, but there were two years with over 900 incidents, than we realize the mean is high because of these two high-fatality years. Thus, if just the mean is reported, there is no way to know if there are outliers in the data. However, if a graph were to accompany the numeric summary, a skew can be seen, showing more information of the whole of the data's structure. Unfortunately, it is common for non-scientists to report only a mean (or a median) without a measure of spread, let alone a graphical data summary.

4.3.8 The critical importance of understanding uncertainty

Every measurement is subject to some degree of uncertainty. If we measure the same object repeatedly, we will not get the exact same answer every time, because there is always some variability in the measurement process. This variability can be due to the measuring instrument, the operator and to changes in environmental conditions.

If we measure the same object repeatedly, we will not get the exact same answer every time, because there is always some variability in the measurement process.

The *magnitude of the measurement variability* (or measurement error) due to instrument is often known by the scientists making the measurements. For example, chemists will typically know the limit of detection of a spectrometer or the accuracy of a thermometer. Other sources of variability may be more difficult to quantify, and some of the variability observed in a measurement may not have a known source.

In statistics, the idea of variability or uncertainty is broad, and encompasses the variation we expect to observe in some measurement due to both known and unknown sources. Uncertainty is quantified using probabilities, probability distributions, or some summary of a probability distribution, depending on the measurement of interest. Two common examples of uncertainty quantification used in every-day life are:

- Weather forecasts, e.g., the chance it will snow tomorrow is 60%.
- The proportion of Iowa voters who plan to caucus for candidate X is 27% ± 3%.
- The current temperature is 50 degrees F, and the measurement is accurate to ±0.5 degree.

In the three examples above, the uncertainty quantifies variability due to different sources. In the case of the political poll, the true proportion of Iowa voters supporting X is unknowable (at least prior to the election), unless we ask every possible Iowa voter. The margin of error is inversely proportional to the number of voters we poll. It reflects the fact that if we were to poll different sets of persons, we would get a different answer each time. This is known as *sampling variability*. In the third example, the uncertainty is related to the precision of the thermometer, which in this case, is half a degree.

In the legal and forensic contexts, we are often concerned with the variability observed when the same object or related objects are measured repeatedly by the same or by different individuals. We might wish to evaluate the variability observed between:

- Repeated measurements of the same object made by the same person.
- Repeated measurements of the same object made by different persons.
- Repeated measurements of different, but similar, objects made by a single person.

- Repeated measurements of different, but similar, objects made by different persons.

We say that measurements are *repeatable* when the same person gets similar measurements over multiple trials. We say that measurements are *reproducible* when two individuals obtain similar results when measuring the same object. Repeatability and reproducibility are both components of the concept of reliability.

Repeatability and reproducibility are both components of the concept of reliability.

4.4 Statistical Inference

Once data are collected, the next step is statistical inference. Statistical inference is the process of drawing conclusions about populations, or scientific truths, from data. Typically, we focus on some summary, such as the mean, of some attribute and draw inference about that *parameter* or population summary. Because we typically do not have measurements from every member of the population, inference about a parameter are almost always based on sample data. From the sample data we compute *statistics*. Intuitively, we might think that the sample mean is a good "guess" for the population mean of some attribute, and in general, our intuition would be correct. Here, we discuss the inferences about population quantities using inferential methods most likely to be introduced in the courtroom.

4.4.1 Point Estimation

As mentioned above, parameters are summaries of some attribute of the population, e.g., the mean, the median or the standard deviation of some variable. Because parameters pertain to the population, unless we obtain measurements from all members of that population, the true value of parameters will always be unknown. As a side note, oftentimes parameters are denoted by θ (theta). *Point estimation* is the process of finding an *estimate*, or a good guess of a parameter—such as the mean—using measurements we obtain from a random sample of members of the population. Because we cannot know the true value of a parameter, it is almost impossible to tell whether the estimate is accurate. However, we can check whether the estimator meets the properties required by good point estimates:

- Unbiasedness: An estimator is unbiased when its expected value is equal to the value of the unknown population parameter it is estimating. As an example, the mean of measurements made on a representative random sample from some population, is an unbiased estimator of the population mean. A biased estimator either overestimates or underestimates the value of a population parameter. Bias can occur from a measurement error (e.g., instrument drift) or from a sampling error (e.g., when the sample does not represent the population).

- Efficiency: Efficient estimators have the smallest variability. Think of it this way: if we were to draw multiple random samples from the same population and from each computed an estimate for the parameter of interest, that estimating method is efficient if the variability of the estimates across the samples is small. The estimator with the smallest possible variance is also called the "best" estimator. That is, the estimator deviates from the true parameter very little. The variability of an estimate is called the *standard error* (SE) of the estimate, and it depends on the sample size. For example, the SE of the sample mean is computed as the *standard deviation* (SD) of the observations divided by the square root of the sample size.

- Consistency: This property states that as the sample size gets larger, the estimate gets closer to the true parameter value. As you get a larger and larger sample, we have more and more information about the population so the statistic we find from our sample will be closer to the population parameter.

Figure 4.6 illustrates some of the ideas discussed above. In this example, we wish to estimate a parameter θ from some population. Suppose that we draw 20 different random samples from the population, each of size *n*, and from each obtain an estimate for θ. The black squares in the circles in Fig. 4.6 represent the 20 sample estimates. The ideal situation is depicted in the top leftmost panel, where all sample estimates are concentrated tightly around the true parameter value shown in the center. In this case, estimators have low bias and low variance, so we can be confident that our guess for θ is reasonable. The worst scenarios are shown on the second row, where in both panels the estimators are biased.

Other terms often used in connection with point estimates are *accuracy*, *validity*, *reliability*, *reproducibility* and *repeatability*. We provide brief definitions below.

The term *reliability* is similar to the colloquial use of *consistency*, and essentially refers to the ability to measure something well, with little variability.

Validity and *accuracy* refer to the closeness with which our estimator approximates the true parameter value. A biased estimator is not valid or accurate.

Repeatability, as the name suggests, refers to the ability of an analyst to reach the same conclusion when presented with the same problem at a later time.

Reproducibility on the other hand, refers to the case where two analysists reach the same conclusion when presented with the same evidence. Both reproducibility and repeatability are components of reliability.

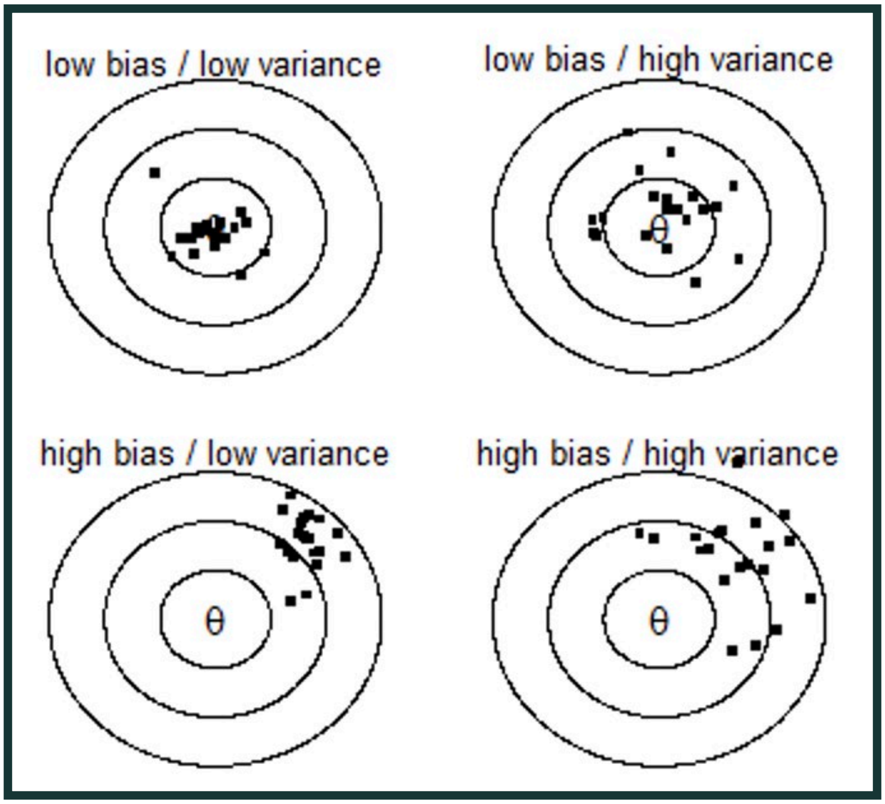

FIGURE 4.6: THE IMPACT OF BIAS AND VARIABILITY OF AN ESTIMATOR.

4.4.2 Interval estimation

As we saw, point estimation results in a single value, our "best guess," for the parameter. A limitation of this approach is that we get no information about the margin of error associated with the estimator. The *margin of error* tells us how far off we can expect our estimate to be given the sample size and the variability of the measurements with which we are working. Thus, often it is useful to report the *range* of likely values of the parameter. This is where intervals come in handy. An interval is constructed by adding and subtracting the margin of error to the point estimate. That is, the general form of an interval estimate is:

Estimate ± margin of error.

The type of interval that is used sometimes depends on the type of data or on the type of data analysis being implemented. Here we focus on the most common type of interval, a *confidence interval*. When computing a confidence interval, we implicitly assume that the sample measurements are distributed more or less symmetrically around their mean. The two most commonly computed confidence intervals are for the mean of a continuous measurement or for a proportion when measurements are discrete.

A *confidence interval* is an estimated range of values that is likely to include the unknown population parameter of interest (e.g., a mean or a proportion), and is computed using the sample data. The level of confidence (C), gives the probability that the interval actually includes the true parameter value. That is, in C% of all samples taken randomly from the population, the population parameter will be contained in the confidence interval calculated using the sample data. For a single sample, we do not know if the interval includes the population parameter value, but we can be C% confident that it does. Common choices for the confidence level C are 0.90, 0.95, and 0.99. This choice of C is often dependent on the type of data and the questions we are trying answer. For example, if we were studying the effects of a lifesaving drug, that may have some negative side effects, we may want to have a higher confidence that it works. However, if we want to find a confidence interval for a drug that has no side effects, we may not need as high of a confidence level.

Confidence intervals can be either one-sided or two-sided. A *two-sided confidence interval* is centered on the sample mean or on the sample proportion, and the width of the interval is such that there is a C% chance that the interval contains (or “covers”) the true parameter value. With a *two-tailed confidence interval*, the sample estimate is directly in the center of the interval. On the other hand, a one-sided interval is not centered around the parameter value but gives more value to the lower or upper region of possible values.

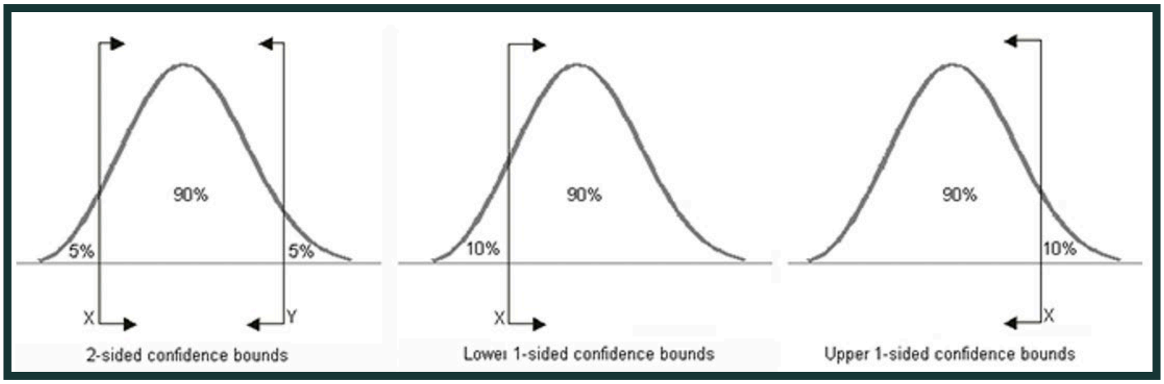
90%
5%
5%
X
Y
2-sided confidence bounds
90%
10%
X
Lowei 1-sided confidence bounds
90%
10%
X
Upper 1-sided confidence bounds

Figure 4.7: A visual representation of the difference between two-sided and one-sided confidence intervals. Notice that both use a 90% confidence intervals; however the two-sided confidence interval is centered, while the one-sided intervals include all possible low values, or all possible high values, depending on direction..

The concept of a confidence interval is to provide some information about the uncertainty associated with a point estimate. The idea is to compute estimates with a small margin of error. One way to achieve this is to increase the sample size, when the realized margin of error is unacceptably large. However, the relationship between sample size and margin of error is not linear (or one to one). To cut the margin of error in half, you would need 4 times as many observations in the sample.

4.4.3 Hypothesis Testing

Hypothesis testing is a standard statistical method for making inferences about an unknown population parameter. When performing a hypothesis test, we postulate two non-overlapping hypotheses, known as the null and the alternative hypothesis. The null hypothesis, denoted H_0, typically reflects our current beliefs, while the alternative hypothesis, denoted H_A, is what we wish to test. For example, assume we wanted to determine whether a coin was fair. The null hypothesis might be half the flips will result in heads. The alternative hypothesis, then, may be the number of heads and tails will be different.[16] The hypotheses are stated in such a way that they are mutually exclusive. That is, if one is true, the other must be false.

Once we have formulated the hypotheses, sample data is used to compute a *test statistic* to help decide between the null or the alternative hypotheses. In the coin flip example, we might toss the coin one hundred times and count the number of heads. Suppose that we get 46 heads and 54 tails. In this case, the test statistic is the sample proportion of heads, or 0.46. The question now is whether 0.46 is close enough to 0.5 to allow us to say the coin is fair or is different enough from 0.5 to lead us to conclude the alternative. Statisticians compute a quantity called the p-value, that can help decide whether to conclude H_0 or H_A given the test statistic we obtained from the sample. A very small p-value (say 0.05 or lower) leads to rejection of the null hypothesis.

The p-value, while used widely, is often mis-understood and mis-used, and with reason.

The p-value, while used widely, is often mis-understood and mis-used. Formally, the p-value is the probability of observing a value of the statistic that is "more extreme" than the observed value *if the null hypothesis is true*. In our coin example, assume we obtain a p-value equal to 0.24. This says that the chance of getting 46 *or fewer heads* even if the coin is fair, is 24%. With this p-value, we would conclude that there is no evidence to say the coin is unfair and would fail to reject the null hypothesis. The rule is: reject the null hypothesis when the p-value is small; fail to reject when it is large.

To decide whether the p-value is small enough to reject the null hypothesis, we must choose a cutoff, or a *level of significance*. This choice is arbitrary, and typically is highly dependent on the context of the problem. When incorrectly concluding H_A is "costly" in some sense, we cautiously set a high level of confidence, and we only reject the null hypothesis when the evidence in favor of the alternative is overwhelming. Common confidence levels include 0.99, 0.95 and 0.90, which lead to cutoffs for the p-values of 0.01, 0.05, and 0.10, respectively. Consider, for example, testing whether a new drug will cure cancer. The null hypothesis is that the drug is no better than what is already on the market, while the alternative is that the drug is more effective than the best treatment available today. If the drug has no bad side effects, then we might not be too worried about incorrectly concluding the alternative and might choose a low confidence level, and a higher cut-off for the p-value of, say, 0.1. This makes it easier to reject the null hypothesis. If, however, the drug has a terrible side effect (for example, it increases the probability of a stroke), then we might want to be more cautious and only reject the null hypothesis if we have overwhelming evidence the drug is effective for cancer. In this case, we would select a higher confidence level, say 99%, which results in a lower threshold for the p-value, 0.01, and therefore make it more difficult to reject the null.

The p-value is often incorrectly interpreted as representing the probability that the null hypothesis is true.

As mentioned above, the p-value is often **incorrectly** interpreted as representing the probability that the null hypothesis is true. However, the p-value says nothing about the probability of H_0 (or H_A). This is one of the reasons why statisticians are moving away from p-values, and from these artificially selected cutoffs, encouraging instead the use of strength of evidence indicators, that may be better suited to the context. One such indicator is what is known as "effect size"; in the cancer drug example, how much improvement does the new drug effect? By focusing on the size of the effect, we emphasize the importance of practical, rather than statistical significance.

4.4.4 Errors in Testing

Errors may occur when we decide between one of the two hypotheses. There are two types of errors: *Type I* and *Type II* errors. A type I error, also known as a *false positive*, occurs when the null hypothesis is rejected even though it is true. In other words, this is the error that consists of accepting an alternative hypothesis when the results we observed were due to chance. We can control the probability of committing a type I error by selecting the confidence level for the test. A type II error, also known as a "false negative" is the error we make when we fail to reject a null hypothesis when the alternative hypothesis is true. The type II error is associated with what is known as the *power of the test*. A powerful test has a low probability of a type II error, meaning that when the alternative is true, we will likely conclude that it is. There is a trade-off between the two types of error, and we cannot minimize them both at the same time; typically, we focus on setting the type I error to an acceptably low value and make sure that the sample size is large enough to ensure acceptable power.

4.4.5 Hypothesis Testing in the Courts

Hypothesis testing is often introduced in legal proceedings in the context of the forensic evaluation of evidence. A question asked in trials is whether the suspect is the source of some evidence found at the crime scene. For example, suppose that glass fragments are recovered from the suspect's clothing and some attribute of the glass – such as its refractive index or RI – is measured. Here, the question of interest is whether the RI of the suspect's fragments are similar enough to the RI of the broken window at the crime scene to suggest that the fragment may have originated from the scene.

The null hypothesis in this particular example is that the $RI_{window} = RI_{fragment}$, and the alternative hypothesis is that the RIs are different. Given measurements of the RI from both sources of glass, a statistician can compute a p-value as described earlier. If the p-value is small enough, the analyst would conclude that the RIs are not similar and therefore, that the fragment found on the suspect is not part of the broken window at the crime scene. If the p-value is not small enough, then the analyst would fail to reject the hypothesis of equal RIs and would be unable to exclude the broken window as the source of the fragment.

While in principle hypothesis testing appears to be well suited to address questions of source, there are two important caveats that we mention even though a thorough treatment is beyond the scope of this chapter:

- The weight of the evidence against the null hypothesis must be overwhelming before we are willing to reject it in favor of the alternative. In the glass example, we begin by assuming that the defendant was at the crime scene unless we can show otherwise. This seems to be backwards in the sense that in the law, a defendant is innocent until proven guilty.
- Failing to reject H_0 does not imply that the fragment was once part of the window. In fact, the RIs of the two glass samples may be indistinguishable, yet the fragment could have come from some other source with the same RI. Thus, testing the hypothesis of equal measurements is the first step. The next step is to demonstrate that if the fragment had come from some other

source, it could not have had an RI that matched that of the broken window at the crime scene. In other words, the analyst should be expected to show that a coincidental match is unlikely before we can conclude they come from the same source. The statistics that have been proposed for this type of analysis include the likelihood ratio (LR) and the coincidental match probability.

4.4.6 Linear Regression

So far, we have talked about inference for a single variable. *Correlation* is an indicator of the relationship between two variables. The correlation coefficient measures the strength of linear association between two quantitative variables. It ranges between -1 and 1. Negative correlations imply a negative association, while positive correlations imply a positive association between the two variables. When two variables are positively correlated, they either increase or decrease together. When two variables are negatively correlated, when one increases the other decreases. The closer a correlation coefficient is to 1 or -1, the stronger the relationship between the two variables. The figure below shows the range of the strength of correlations. Commonly, the range between 1 and .7 is considered a strong relationship, .7 to .3 a moderate relationship, .3 to 0 a weak relationship and 0, no relationship. However, these strengths of relationships often depend on the type of data with which we are working.

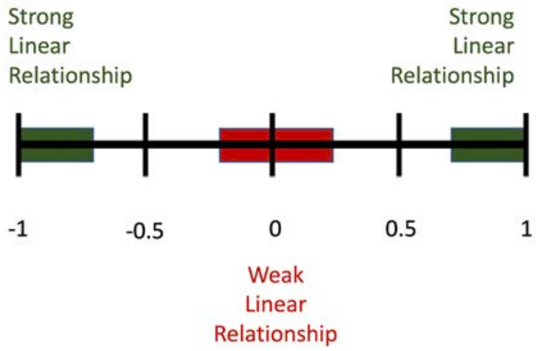

Figure 4.8: Strength of relationships based on correlation coefficients

Correlation does not mean causation. Just because two variables are highly correlated, does not mean one variable causes the other. For example, there is a high positive correlation between number of TV sets per person and average life expectancy. That does not mean one should buy several TV sets to have a long life. Instead, it is more likely that some other variable, or variables, such as wealth, may be creating an association between TVs and life expectancy. These *lurking variables* can have important effects on the associations we observe. A common problem, however, is these lurking variables are often not included as part of the data collection.

Correlation does not mean causation. Just because two variables are highly correlated, does not mean one variable causes the other.

While the correlation coefficient is a useful measure of the association between two variables, sometimes we wish to go further and *model* that association. The simplest statistical model is a straight line, to provide a good representation of the relationship between the variables. Such a line is called a *linear regression line*.[17] A regression line explains how the values of the *response variable* change in relation to changes in the value of the *explanatory variable*. For a response variable y, and an explanatory variable x, the linear regression line is defined by:

$$\hat{y} = b_0 + b_1 x,$$

where b_0 is the intercept and b_1 is the slope of the line. That is, for a one unit increase in the explanatory variable (x), the predicted value of the response variable (y) will change by an amount equal to the slope. This gives us a reasonable way to quantify the relationship between the two variables. When the slope is negative, there is a negative correlation; when the slope is positive, there is a positive correlation.

In most cases, the slope is the parameter we most care about. For example, suppose a town wants to build a new fire station. In order to find a good location, they examine the relationship between the distance from the fire station and the amount of damage to homes from past fires (in thousands of dollars). Figure 4.9 shows the scatter plot with the regression line.

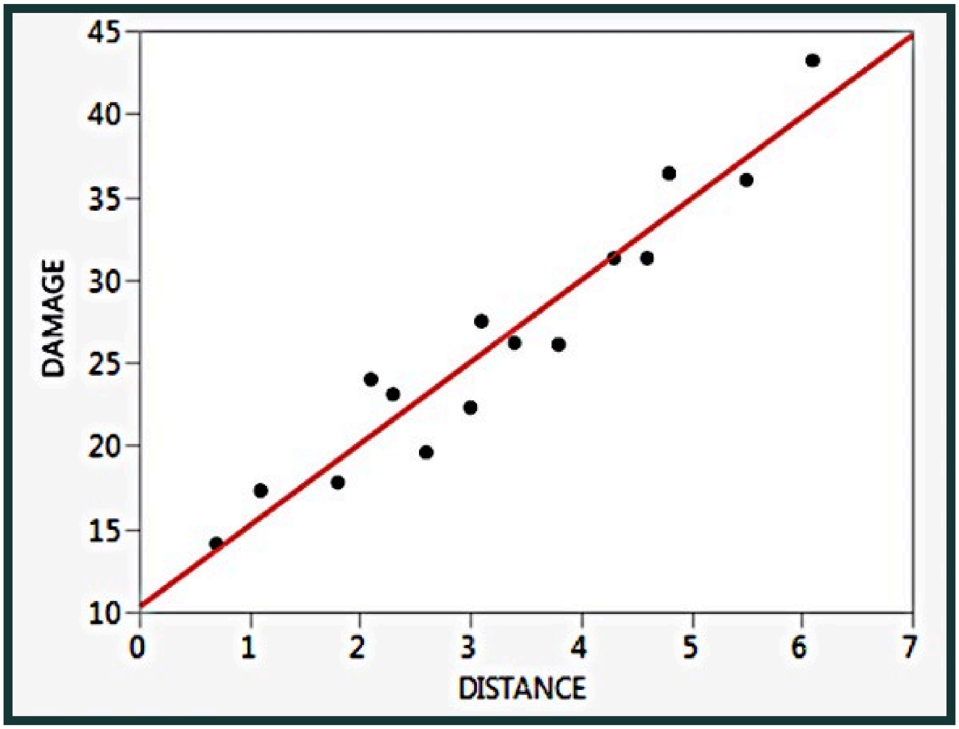

Figure 4.9: The regression line depicting the relationship between distance from a fire station and damage in thousands of dollars.

In this example we find that Damage= 10.28 + 4.92 x Distance. Interpreting this, we would say that, for every extra mile away a property is from the fire station, we expect the damage to increase by 4.92 thousand dollars.

Extrapolation, or predicting a response value for an x-value outside the scope of the data, is risky.

When appropriate, we can use a regression line to predict the expected value of a response variable given the value of the explanatory variable. In our example, we would expect a property five miles from the fire station to sustain damage of approximately $34.88 thousand. However, these predictions can be very inaccurate when we extrapolate beyond the range of the data we used to estimate the regression line. *Extrapolation*, or predicting a response value for an x-value outside the scope of the data, is risky. We really do not know whether the association between y and x continues to be linear beyond the range of our data. Figure 4.10 shows what might happen when we extrapolate. The blue dots represent the sample data, the blue line is the regression line estimated from those data, and the red curve represents the true (but unknown) relationship between x and y. If we only observe the response y for values of x between 0 and x tilde, then we would believe that their relationship is linear. But if we wish to use the estimated regression line to predict the response for a value of x equal to x*, we will make a huge error because beyond x tilde, the relationship between y and x is no longer linear.

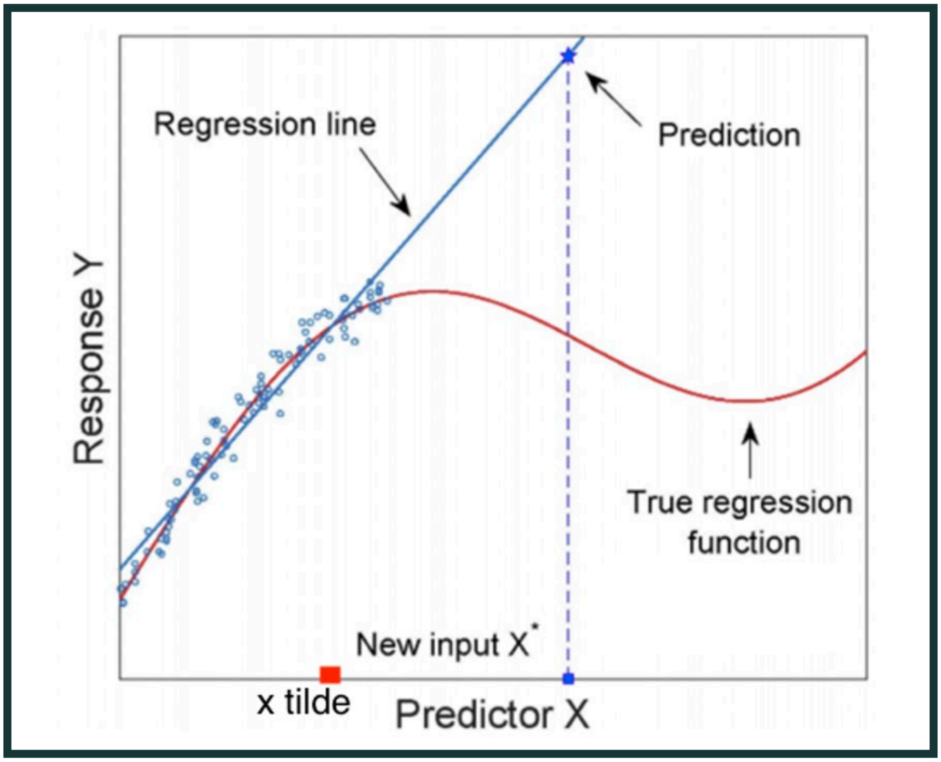

Figure 4.10: A visual representation of the risk that predicting outside the scope of the data may lead to.

For example, consider plotting the height of a sample of persons against their age, but only conducting the study with participants no older than 10 years of age. While this study may predict accurately the height of pre-adolescents, it would not reliably predict the height of a 49-year-old.

Another caution regarding the use of linear regression is that the relationship between the response y and the explanatory variable x needs to be linear. If the relationship between the two variables is not linear, you should not summarize it with a line. For example, income tends to rise almost linearly as years of education increase between 0 and about 16, but the relationship flattens after that point. Thus, whether you went to school for 18 or for 24 years, your income will tend to be unaffected.

Another problem that may arise when using linear regression is known as overfitting. *Overfitting* occurs when a function is too closely fit to a limited set of data points. In the case of linear regression, overfitting can occur when the sample size is small or when the range of the explanatory variable is limited. The consequence of overfitting is a decrease in the accuracy with which we can predict the response for a new value of the explanatory variable.

Statisticians have developed many diagnostic tools that a user of linear regression can implement to decide whether the linear regression model is "good." By "good" we mean the model fits the sample data reasonably well and has good predictive properties, and that the sample data do not violate any of the assumptions implicit in the method. Perhaps the most common approach to carry out a diagnostic for the linear regression model is a *residual analysis*.

For more information about regression modeling, residual analyses and other tools, the reader should refer to any introductory statistics book. Two good references are: *An Introduction to Statistical Learning*[18] by James Gareth, et al., and *Intro Stats* by Richard De Veaux.[19]

4.5 Summary

Statistics, like other fields of study, provides a number of tools that may be of assistance in understanding and interpreting data of many different types. We have sought herein to explain some of the common concepts encountered in statistical analysis with the hope it will aid in evaluating statistical evidence.

4.6 Definitions from Section 4 (in alphabetical order)

Bayes Theorem: a theorem that computes the probability of an event based on prior knowledge about the event and on the probability of conditions that may be related to the event.

Bias: a systematic distortion of a statistical result due to a factor not accounted for in its computation.

Coefficient of Determination: R^2, the proportion of the variance in the response variable that can be explained by the explanatory variable(s).

Conditional Probability: a measure of the probability of an event occurring given that another event has occurred.

Confidence Interval: a range of values around an estimate of a quantity, that reflects uncertainty about the true value of the quantity. In statistics, the quantity we wish to estimate is often called a parameter.

Confidence Level: the probability that the confidence interval covers the true value of a parameter.

Continuous Variable: A variable that can take on any value within an interval.

Correlation: a quantity measuring the extent of the interdependence of two or more variables.

Data: facts and statistics collected together for reference or analysis.

Discrete: A variable that can only take on integer values, i.e., whole numbers, within an interval.

Experiment: a scientific study undertaken to make a discovery, test a hypothesis, or demonstrate a known fact.

Explanatory Variable: The x variable; a variable that explains or predicts changes in another variable, known as the Response Variable.

Hypothesis: a supposition or proposed explanation based on limited evidence as a starting point for further investigation. The statement at the beginning of a hypothesis test explains what is being tested.

Independence: the attribute of a variable whose variation does not depend on the variation of another.

Interquartile Range: the range of the middle 50% of a data set.

Joint Probability: the chance of two events occurring together.

Linear Regression: approach to modeling the relationship between a response (or dependent or response variable) and one or more **Explanatory Variables** (or independent variables) by a straight line.

Long Run Frequency: establishes the probability of an event by the frequency with which the event occurs in a very large number of trials.

Lurking Variables: a variable unknown and not controlled for but which has an important, significant effect on the variables of interest.

Mean: the mathematical average of a collection of observed values.

Median: the midpoint of a frequency distribution of observed values. Half of the data values are below the median and half are above.

Observational Studies: A study in which the study subjects are not randomly assigned to treatments by the investigator.

Odds: ratios of probabilities, describing how likely an event is to occur.

Ordinal Data: statistical data type where the variables have natural, ordered categories and the distances between the categories is not known.

Outliers: a data point on a graph or in a set of results, that does not follow the general pattern of the data.

Overfitting: when a function is too closely fit to a limited set of data points.

Parameter: a numerical or categorical measurement that describes the population.

Population: the universe of objects of interest.

Point Estimate: a single value computed from a sample, used as an "educated guess" of the value of a parameter for a population.

Probability: The probability of an event is a number between 0 and 1 that reflects the likelihood that the event occurs.

Product Rule: if events A and B are independent, then their joint probability is the product of the probability of A and the probability of B.

Qualitative Data: data that are not numerical but fit into categories. An example is marriage status.

Quantitative Data: data that are numeric. An example is annual income.

Response Variable: a variable (often denoted by y) whose value depends on that of another.

Sample: a set of objects that are available for study and that were obtained from the population of interest.

Sampling: the action or process of drawing samples from a population, typically for statistical analysis.

Standard Deviation: a measure of how much variation there is in a set of data.

Statistic: a numerical measurement that describes an attribute of the sample.

Type I Error: in a test of hypothesis, rejecting the null hypothesis, when in fact the null is true.

Type II Error: Failing to reject the null hypothesis, when in fact the null hypothesis is not true.

Variance: a measure of how much variation is in a set of data, computed as the standard deviation squared.

4.7 Bibliography:

Bruce, Peter C., and Andrew Bruce. *Practical Statistics for Data Scientists: 50 Essential Concepts*. OReilly Media, 2018.

D., De Veaux Richard, et al. *Intro Stats*. Pearson, 2018.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. , 2013. Print.

Pishro-Nik, Hossein. *Introduction to Probability, Statistics, and Random Processes*. Kappa Research, LLC, 2014.

4.8 Endnotes

1. Eryn Blagg is a doctoral student in the Department of Statistics at Iowa State University

2. Alicia Carriquiry is distinguished professor of Statistics at Iowa State University, and Director of the Center for Statistics and Applications in Forensic Evidence (CSAFE)

3. Blagg's and Carriquiry's work was partially funded by the Center for Statistics and Applications in Forensic Evidence (CSAFE) through Cooperative Agreement 70NANB20H019 between NIST and Iowa State University, which includes activities carried out at Carnegie Mellon University, Duke University, University of California Irvine, University of Virginia, West Virginia University, University of Pennsylvania, Swarthmore College and University of Nebraska, Lincoln.

4. Butler, J.M. 2014. *Advanced Topics in Forensic DNA Typing: Interpretation*. First Ed., Academic Press. 608 pp.

5. Mathematically the odds of an event occurring is as follows:

$$Odds_f = \frac{\text{Probability that } Y \text{ occurs}}{\text{Probability that } Y \text{ does not occur}}$$

6. Probability in mathematical terms:

$$\Pr(\textit{died } 18\text{- } 20 \textit{ hours ago}) = 0.9$$

7. We write: Pr(A|B) to denote the probability of observing event A given that event B has occurred. In the example: $Pr(\textit{died } 18\text{-}20 \textit{ hours ago} \mid \textit{body was in the cold}) \leq 0.2$

8. This is a frequency table. The goal of a frequency table is to visually display the different counts of each of the categories.

9. $Pr(A \text{ and } B) = P(A) \times P(B)$.

10. Butler, J.M. 2014. *Advanced Topics in forensic DNA Typing: Interpretation.* First Ed. Academic Press, 608 pp.

11. $Pr(A \mid B) \neq P(B \mid A)$.

12. Mathematical formula of Bayes Theorem:

$$Pr(A|B) = \frac{Pr(B|A) \times P(A)}{Pr\,(B)}$$

13. https://www.chicagotribune.com/nation-world/chi-chicagodays-deweydefeats-story-story.html

14. The FBI Data has come from the Uniform Crime Reporting from the US Department of Justice found at https://www.ucrdatatool.gov/index.cfm. These data were collected between 1960 and 2018.

15. A good visual introduction to these topics can be found here : https://seeing-theory.brown.edu/index.html#firstPage

16. Symbolically, these hypotheses would be expressed as Ho: $P_{heads} = 0.5$ and Ha: $P_{heads} \neq 0.5$

17. Here we reference linear regression. There is also polynomial regression. Some good resources for these topics are: https://towardsdatascience.com/5-types-of-regression-and-their-properties-c5e1fa12d55e

18. James, Gareth, Witten, Hastie, and Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. , 2013. Print.

19. De Veaux, et al. *Intro Stats*. Pearson, 2018.