

Revised February 20, 2006

Chapter 1. Introduction

Statistics is the science of collecting, organizing, analyzing and drawing inferences from data. The goal scientific and engineering investigations is to develop models for a system of interest. Experiments and/or surveys are conducted to answer questions such as which factors are important in a manufacturing process or how can we optimize the process. The results of these experiments are data which often correspond to a list of numbers, such as readings from a gauge, temperature settings, material compositions etc. The data is the evidence in a scientific enquiry and statistical ideas are needed to transform the list of numbers into something meaningful. Statistical ideas are used to summarize a long list of numbers by extracting from the data pertinent information in an efficient manner. Graphical techniques are very useful for summarizing the information in data sets. The data can be used to build and test models that help explain the process and solve the problems of interest to the engineer.

We need to define a few terms for further discussion. Statistical applications deal with trying to solve problems pertaining to a group of objects, such as a product that is being manufactured. The entire group of objects is called the *population*:

Definition. A *population* is the set of all possible observations of interest to the problem at hand.

Populations tend to be very large in practice. If we are interested in performing an experiment in the design of a new product, then the population is hypothetical. In other words, suppose we want to determine how a newly re-tooled engine performs. We may produce a sample of prototype engines to test. The population in this case would be all future engines based on this prototype if this prototype is put into production. The power of statistics is that we can infer things about a very large or even hypothetical population based on a relatively small sample obtained from the population.

Definition. A *Sample* is a subset of a population.

Usually samples are obtained at random from the population. This is discussed further below. Once a sample of units from the population is collected, we measure some quantity of interest from these units. These measurements are often numerical and they represent the data. We shall consider another example where a plant produces cup-a-soups. This particular plant was having trouble filling the cups with the correct volume of soup. The population in this example is all cups that will be filled by the filling tanks at the plant. In order to solve the problem with filling the cups, a sample of cups were obtained from the manufacturing process. Interest lies in the volume of soup in the cups. However, it is easier to measure the weight of the soup, the fill-weight which provides essentially equivalent information about the volume. The

fill-weight is an example of a *variable*:

Definition. A *variable* is a (numerical) value of interest obtained from the elements in the population. Sometimes engineers may measure more than one variable from elements in the population. For example, if a factory produces bolts, we can measure the length and the diameter of each bolt, i.e. record data on two variables for each item. Data consisting of measurements on more than one variable puts us in the realm of *multivariate statistics*.

In order to understand the population of interest, we attempt to model the population by a mathematical expression involving the variables of interest. In many cases, very simple models can provide a succinct understanding of the population. The models are usually defined in terms of parameters:

Definition. A *parameter* is a numerical descriptive of the population of interest.

In the cup-a-soup example, a parameter of interest is the average weight of soup that is placed in all cups. Averages for an entire population are generally denoted by the Greek letter μ (pronounced “mu”).

We give examples of models below. Because a parameter is a descriptive of the entire population, the true value of parameters are usually unknown. However, using the data, we can obtain estimates of the parameters of interest. These estimates computed from samples are known as statistics:

Definition. A *Statistic* is a numerical summary computed from the data alone.

A commonly used statistic is to take the average of the observations in a sample. If we denote the variable of interest by the letter y , then the data can be represented by y_1, y_2, \dots, y_n , where n is the sample size, i.e. the number of observations in the sample. The average (or mean) of the sample is denoted by \bar{y} (pronounced “y-bar”) which is found by adding up all the y values and dividing by n .

Statistics is a very broad field of study and plays an essential role in every scientific endeavor because almost all information comes to us in the form of data. Of course, before data can be analyzed, it must first be collected either by an experiment or a survey.

1.1 Obtaining Data.

Running an experiment or doing a survey can be very expensive and time consuming. If there is a problem with a product from a manufacturing process, an experiment may need to be conducted in order to determine how to fix the problem. One needs to determine which factors in the production process effect the product. It may be necessary to shut down a production line to perform the experiment and this will cost the company money in the short term but hopefully a better product will result. A company may want to develop or refine a product and again experiments are needed to determine the optimal way to produce the product.

Therefore, it is very important that great care and thought go into the planning of an experiment. If a survey is to be conducted, again great care is required to get the most information possible. There are two areas of statistics dealing with the collection of data: *experimental design* and *sampling design*. Statistical ideas allow experimenters to design experiments and surveys in ways to obtain the optimal amount of information given the resources available which can result in great savings of time and money.

Unfortunately, it is not uncommon for an experiment or survey to be poorly planned resulting in data that has little value. This is a great waste of resources. If the data is no good, then there are no statistical techniques that can salvage the problem. One essential question is to determine how much data is required. Remember that all the evidence for making decisions lies in the data. If we do not have enough data, then we may not have enough evidence to make a decision.

Before the data is collected, a well-defined problem should be formulated which involves dialogue between the statistician and the engineer. What characteristics are of interest to solve the problem? A model needs to be postulated. Models are defined in terms of *parameters*. The parameters of interest need to be estimated. Often, there will be physical constraints that may cause us to modify the data collection procedure.

There are basically three types of data collection methods:

1. *Retrospective study* based on historical data. In these situations, the data is already collected which is convenient, but this is also a big impediment since there could be problems with the data that we cannot control (perhaps an important variable was not recorded, such as time-order of observations).
2. *Observational Studies* – data from observational studies are obtained by observing a population or system and recording values. In many situations, observational data is the only data available such as wildlife studies or opinion polls. With observational data it is difficult to access *causality*, that is, it may be the case that two variables of interest are related (*correlated*), but can we say that one variable causes the change in the other variable? A classic example is whether or not smoking causes lung cancer in humans. We only have observational data for this problem and there is very strong evidence these two variables are related, but statistics cannot definitively tell us if smoking is the cause of lung cancer.
3. *Designed Experiments* - Experiments can be conducted where the experimenter has complete control over the *factors* that may effect the *response* of interest, such as in a manufacturing process where we control the amounts of ingredients, the temperatures, etc. to see the effect on the product. Suppose you can obtain a certain amount of information based on a sample of $n = 50$ observations using an inefficient method and that you can obtain the same amount of information using an efficient experimental design using on $n = 25$ observations, then it would make good economic sense to use the efficient experimental design since it would save money and time.

1.1.1 Sampling.

Observational studies obtain data by sampling the population of interest. There are many different sampling plans. They each involve collecting data from a population in some randomized way. In order to infer something about a population or process of interest based on a sample taken from that population, we need to obtain a sample that is representative of the population or process. To ensure that a representative sample is obtained, we usually use chance to our advantage and obtain a *random sample*.

It is very easy and tempting to obtain a sample by collecting observations that are easily obtainable (a *convenience sample*). The resulting data is not necessarily representative of the population of interest. If we do not have a well-thought out sampling plan, then our biases can damage the quality of the data collected. Often times these biases are not things we will be consciously aware of. For example, if performing a survey of contracts that your company has with other companies, you may unconsciously select contracts associated with large sales more so than contracts associated with smaller sales resulting in a sample that is not representative of the population of contracts as a whole. One of the most common examples of observational studies in the news media are opinion polls. Typically these are conducted by picking at random people from the general population to solicit their opinion. There are many examples of opinion polls that gave very bad information because the data was not collected in a statistically valid way.

Humans are often skeptical of leaving things to chance, but in order to obtain representative samples, allowing chance to play a role is essential. Because randomness plays an important role in the collection of data, the theory of probability is the foundation to any statistical analysis. We shall discuss probabilistic ideas as we progress through the book.

Here are three popular methods of sample design.

1. *Simple Random Sampling* - this is the most simple method where-in each sample of a given size (say n) has the same probability of being selected. This method guarantees no bias in which observations from the population are selected but there are usually more efficient sampling plans.
2. *Stratified random sampling* - a useful and efficient method when the population can be divided into natural set of strata. The idea then is to obtain a simple random sample from each of the individual strata. The problem then becomes how to choose the sample sizes for individual strata. One choice is to base sample size in individual strata on the overall size of the strata, but other alternatives exist as well.
3. *Systematic Random Sampling* - a useful method in manufacturing processes where you sample every m th item and the first item is chosen randomly from the first m items.

In each of these three methods, random samples are obtained. In practice, to obtain a random sample, one method is list the elements in the population as $1, 2, \dots, N$

where N is the total number. Then write the numbers on slips of paper, put them in hat, mix them up, and then select a sample at random. This is quite cumbersome though. An easier method is to use a computer to randomly generate a sample of numbers to use for our random sample.

1.1.2 Experimental Design.

Unlike observational data, experimental data is obtained from running an experiment that the experimenter controls. As in sampling designs for observational data, randomization also plays an essential role in the design of experiments. Suppose we want to test how well two different seals work on a compressor. We can randomly select a collection of compressors and assign half to get one type of seal and the other half to get the other type of seal. In order to wash out differences in the compressors that may exist, we would want randomly assign compressors to the seal types.

Experiments are conducted for a variety of reasons:

1. *Screening* factors which influence the response.
2. *Predict* the behavior of the response over a specified range of the factors. For example, running a production process at various temperatures, can we predict the outcome based on a given temperature.
3. *Optimize* - In the temperature example, what is the optimal temperature? It may very well depend on the *level* of other factors that affect the response.
4. Make products robust to sources of variability. We may not necessarily want to optimize some quantity, but instead need to minimize the variability. A process is called *robust* if it achieves a target condition for a characteristic of interest with minimal variability.

1.2 Graphical Displays of Data and Distribution Shapes.

Once data from an experiment or survey has been collected, it is time to extract information from the data. Staring at a long list of numbers in a data file is usually not very illuminating. A first step in this direction is to graphically display the data – let the data tell us what is happening by means of a picture. Computer software packages can generate many different types of plots for us. We shall discuss two commonly used plots now: the stem-and-leaf plot and a histogram. Both of these plots are useful for accessing the “shape” of the distribution of data points. We shall introduce a few other types of plots later as we progress.

1.2.1 Stem-and-Leaf Displays.

A very simple type of plot that can be constructed fairly easily by hand is the *stem-and-leaf* plot. This plot can give us a picture of the “shape” of the data distribution

and the shape can tell us a lot about the population from which the data came. The first step in the stem-and-leaf plot is to arrange the data in order from smallest to largest. To illustrate, we present the first 20 observations (the fill-weight in the cups) from the first filling lane in the cup-a-soup example.

Cup-a-Soup Example The first 20 observations in the cup-a-soup example arranged from smallest to largest:

236.39, 236.93, 237.26, 237.39, 237.46, 237.71, 237.83, 237.98, 238.00, 238.03,
238.52, 238.66, 238.68, 238.87, 239.08, 239.27, 239.32, 239.61, 239.67, 239.67.

Next we look for a natural way to break each observation into two pieces: one piece representing the “stem” of the number and the other piece representing the “leaf”. There may be more than one way of doing this. As a first illustration, let’s treat the units digits as the stem and the tenth digit as the leaf and round each number to the nearest 10th.

| | |
|------|---------|
| 236. | 49 |
| 237. | 34578 |
| 238. | 0005779 |
| 239. | 133688 |

This plot seems to indicate that the data is *skewed* to the left. However, it is difficult to discern the shape well because we have so few categories. If we have too few categories we cannot see the shape. Similarly, if we have too many categories, most lines will either zero or one observation each which also will not allow us to see the shape of the data. We must determine a reasonable value that allows us to see the shape of the data.

For the cup-a-soup example, let us split each stem in two based on values 0.0 – 0.4 and 0.5 – 0.9.

| | |
|------|------|
| 236. | 4 |
| 236. | 9 |
| 237. | 34 |
| 237. | 578 |
| 238. | 000 |
| 238. | 5779 |
| 239. | 133 |
| 239. | 677 |

Note: if there are gaps in the data, you still need to include the stems in between these gaps (illustrate with an additional value of 241.2 say) for these gaps provide important information about potentially unusual observations (outliers).

Note – the stem-and-leaf plot retains all the information in the plot (up to rounding) since the actual data is in the plot itself.

1.2.2 Histograms.

Histograms are very similar to stem-and-leaf plots and are useful for seeing the shape of the data distribution, particularly for large data sets where it would be tedious to construct a stem-and-leaf plot. The idea is to divide the range of data into a suitable number of intervals *all with the same length* called measurement classes. Then compute either the number or proportion of observations in each interval and draw a rectangle over each interval whose height is proportional to the number or proportion of observations in each interval. We shall use the $n = 20$ fill-weight observations to illustrate the construction of a histogram. Let us use measurement classes of width $1/2$. The following table breaks the data down into the measurement class:

| Measurement Class | Frequency |
|------------------------|-----------|
| $236 \leq y < 236.5$ | 1 |
| $236.5 \leq y < 237$ | 1 |
| $237 \leq y < 237.5$ | 3 |
| $237.5 \leq y < 238$ | 3 |
| $238 \leq y < 238.5$ | 2 |
| $238.5 \leq y < 239$ | 4 |
| $239 \leq y < 239.5$ | 3 |
| $239.5 \leq y < 240.0$ | 3 |

The histogram for this data is shown in Figure 1.

It is difficult to access the shape of a distribution with only a few observations. If we consider the entire data set on fill-weights from the first filling lane ($n = 969$), we obtain the frequency histogram shown in Figure 2. This figure shows a roughly symmetric looking distribution centered around 239, but with many observations in the “left tail” indicative of an underfilling problem with this lane.

The shape of a distribution can tell us a lot about the population of interest. Figure 3 shows four different shapes commonly encountered with real data sets. The bell-shaped distribution (top-right panel of Figure 3) was constructed based on facial measurements of Swiss soldiers. This symmetric bell-shape distribution is actually quite common in many applications. The distribution plotted in the bottom-right panel was constructed with data on the length of an electrode produced by two different machines. Ideally there should be no differences between the electrodes produced by the two different machines. However, the bottom-right panel in Figure 3 shows a *bimodal* pattern indicative of two distinct subpopulations. Bimodal patterns occur with other types of populations. For example, data collected males and females may yield a bimodal distribution. The distributions plotted in the top-right and bottom-left panels of Figure 3 show skewed distributions. Data on home prices and salaries tend to be skewed to the right due to some very expensive homes and high incomes. If you look at test scores from an easy exam, then you would expect to see a distribution skewed to the left.

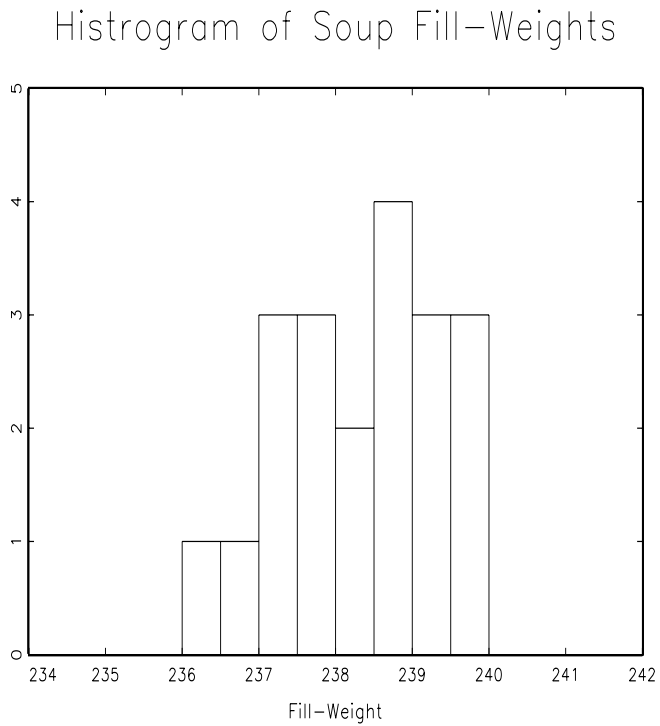


Figure 1: Histogram of fill-weights of the first 20 observations of the cup-a-soup example.

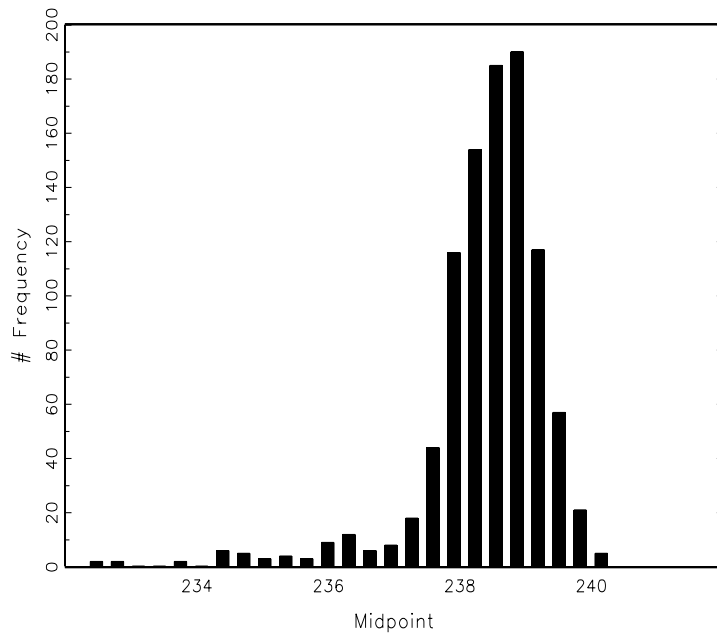


Figure 2: Histogram of fill-weights for lane 1 of the cup-a-soup production sample of $n = 969$ cups.

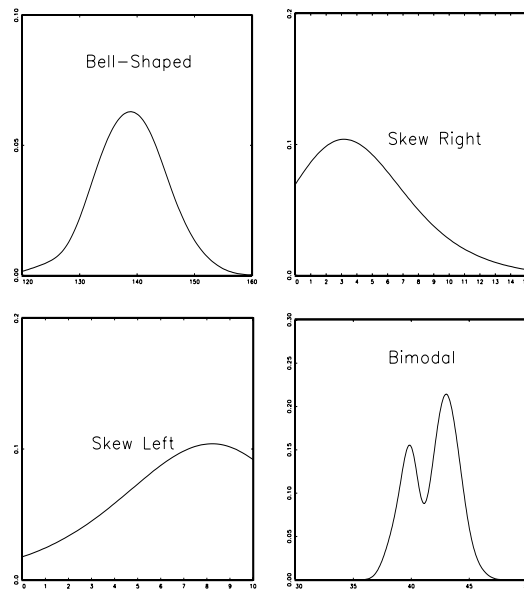


Figure 3: Four common shapes of data distributions: top-left: symmetric and bell-shaped (or mound-shaped), top-right: skewed to the right, bottom-left: skewed to the left, bottom-right: bimodal

1.3 Some Important Descriptive Statistics.

Plots of data provide pictures of our data. In addition to plots, it is also useful to compute some common *statistics* to provide an overall summary of the data. The most common statistics used in practice are the mean (or average), the median, and the standard deviation. The mean and median are measures of “central tendency.” These two statistics give an indication of where the data is centered. The standard deviation tells us how spread out our observations are.

1.3.1 The Sample Mean (or average). Given a sample of data y_1, y_2, \dots, y_n , the sample mean is simply the average of the observations and is denoted by \bar{y} :

$$\bar{y} = (y_1 + y_2 + \dots + y_n)/n = \sum_{i=1}^n y_i/n.$$

In the cup-a-soup example, considering the first $n = 20$ measurements only, the sample mean (i.e. average) fill weight is

$$\bar{y} = (236.39 + 236.93 + 237.26 + \dots + 239.67)/20 = 238.32.$$

A convenient way to think of the mean is that it is the *center of gravity* for the data points. If we draw a number line on a board and place a ball on the board corresponding to its value on the number line, then the center of gravity of the board with the balls will occur exactly at the sample mean.

In the cup-a-soup example, we computed the sample mean based on a sample of $n = 20$ cups. In practice, one would be interested in the mean fill-weight of all cups

produced by the filling machine. The mean fill-weight for the entire population of cups that can be produced by the filling machine is called the *population mean* and it is denoted by the Greek letter μ (“mu”). In typical applications, the population mean μ is almost always unknown and we use the sample mean \bar{y} to estimate the population mean. The statistical question then becomes – how good is \bar{y} as a point estimate of the true population mean μ ? We shall address this question later in the book.

1.3.2 The Sample Median. If we arrange our observations in order from smallest to largest, then the sample median is simply the middle observation. If the sample size n is odd, then there is a unique middle observation. However, if the sample size is even, then the median is computed as the average of the two middle observations. To illustrate, consider the cup-a-soup example again. The data, arranged in ascending order is:

236.39, 236.93, 237.26, 237.39, 237.46, 237.71, 237.83, 237.98, 238.00, 238.03,
238.52, 238.66, 238.68, 238.87, 239.08, 239.27, 239.32, 239.61, 239.67, 239.67.

Because the number of observations ($n = 20$) is even, the median is computed as the average of the 10th and 11th observations:

$$\text{Median} = M = (238.03 + 238.52)/2 = 238.28.$$

If a distribution is symmetric, then the mean and median are exactly equal. However, if the distribution deviates from symmetry, then the mean and median may not be equal. The mean is the statistic most often used for central tendency. However, there are many examples where the median is a better statistic for central tendency, particularly if the distribution is strongly skewed or has extreme outlying observations. Consider the following extreme example:

Example. A sweatshop used to manufacture athletic shoes pays their 19 employees \$1 per hour and pays the manager \$100 per hour. A human rights organization decides to investigate the factory. The factory claims (correctly) that the average wage is \$5.95 per hour which is considered a very good wage in the country where the factory is located. Where did the figure \$5.95 come from? The data is $y_1 = 1, y_2 = 1, \dots, y_{19} = 1, y_{20} = 100$, and the sample mean is

$$\bar{y} = \sum_{i=1}^{20} y_i / 20 = (1 + 1 + \dots + 1 + 100) / 20 = 119 / 20 = 5.95.$$

Clearly, the mean is a very poor measure of central tendency in this example since all but one of the employees earn less than the average. The median on the other hand is $M = 1$ which is a much better reflection of the distribution.

This example points out that the mean can be highly influenced by a few extreme observations. This is why one sees the median used as a measure of central tendency in examples concerning salaries and home prices. The median is said to be a robust

measure of central tendency because the median is not strongly influenced by a few extreme observations.

1.3.3 Variance and Standard Deviation. Suppose a professor gives two tests during the course resulting in the following test scores for the five students:

Test 1 : 50, 60, 70, 80, 90

Test 2 : 68, 69, 70, 71, 72

For both tests, the average score is a 70. However, the distribution of test scores are very different for the two tests. This example illustrates that the mean (or average) itself is not adequate by itself for describing a distribution. The test 1 scores are quite spread out with scores ranging from 50 to 90; the test 2 scores are tightly clustered together and less spread out. Another statistic is needed to characterize the “spread” of the data points and the most common statistic used to measure the degree of spread is the standard deviation. The idea is to look at the individual deviations from the average: $(y_i - \bar{y})$. The deviations always sum to zero which is not informative. Instead, the average of the squared deviations is used to measure the spread. However, instead of dividing by n as we did for the mean, we average the squared deviations by dividing by $n - 1$ which will be explained later. Given a set of observations y_1, y_2, \dots, y_n , with a sample mean of \bar{y} , the *sample variance* s^2 is defined to be:

$$\text{Sample Variance : } s^2 = \sum_{i=1}^n (y_i - \bar{y})^2 / (n - 1). \quad (1)$$

The (positive) square root of the variance is called the *sample standard deviation*:

$$\text{Sample Standard Deviation : } s = \sqrt{s^2} = \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 / (n - 1)}.$$

To illustrate the computation, we compute the sample variance and standard deviation for the two sets of test scores above. Letting s_1^2 and s_2^2 denote the sample variances of the first and second tests, we find

$$s_1^2 = \{(50 - 70)^2 + (60 - 70)^2 + (70 - 70)^2 + (80 - 70)^2 + (90 - 70)^2\} / (5 - 1) = 250$$

and

$$s_2^2 = \{(68 - 70)^2 + (69 - 70)^2 + (70 - 70)^2 + (71 - 70)^2 + (72 - 70)^2\} / (5 - 1) = 2.5.$$

The standard deviations for tests 1 and 2 are

$$s_1 = \sqrt{250} = 15.811$$

and

$$s_2 = \sqrt{2.5} = 1.581.$$

The standard deviation for test 2 is much smaller than for test 1 indicating less spread of test scores for test 2 compared to test 1.

In the cup-a-soup example, the sample standard deviation for the $n = 20$ measurements is $s = 0.96$. The standard deviation for the population of all fill-weights from this filling tank would be the population standard deviation. Just as \bar{y} was an estimator for the population mean μ , the sample standard deviation s is an estimator for the population standard deviation which is denoted by the Greek letter σ (sigma). We shall give a formal definition of σ in the next two chapters.

1.4 The Empirical Rule

Data from many populations form approximately bell-shapes when plotted (in histograms say). For these types of data sets, the information can typically be summarized by the mean and standard deviation without much loss of information. The following rule, known as the *empirical rule*, allows us to interpret the mean μ and standard deviation σ of a distribution:

- Approximately 68% of the observations will lie within 1 standard deviation of the mean: $\mu \pm \sigma$.
- Approximately 95% of the observations will lie within 2 standard deviations of the mean: $\mu \pm 2\sigma$.
- Approximately 99.7% of the observations lie within three σ of the mean. That is, practically all observations will lie within 3 standard deviations of the mean.

If 68% of the distribution lies within one standard deviation of the mean, then we can divide this 68% evenly in two ($68\%/2 = 34\%$) with 34% lying between $\mu - \sigma$ and μ . The other 34% lies between μ and $\mu + \sigma$. Further breakdown of the distribution is illustrated in Figure 4.

The empirical rule also holds for sample data possessing a bell-shaped by simply replacing μ by \bar{y} and σ by s in the above rule.

According to the empirical rule, practically the entire distribution (99.7%) will lie in a range of 6σ (from $\mu - 3\sigma$ to $\mu + 3\sigma$). In a manufacturing process, if we are recording the value of some variable, say the length of a part, the recorded lengths should mostly vary within two standard deviations of the mean. If the measurements start drifting to more than two or three standard deviations from the mean, then that is an indication that the production process is having a problem.

Batting Averages Example. No major league baseball player has had a batting average of .400 or better since Ted Williams in 1941. A player's batting average is basically the number of hits the player gets divided by the number of times the player goes to bat. Many baseball fans have debated the reason for the extinction of .400 hitting in baseball for years. If we look at the data, the empirical rule demonstrates why .400 hitting is non-existent in major league baseball. Below are dotplots (see Section 1.5) for league batting averages in the years 1920 and 1997 – the dots represent batting averages for all players with at least 250 plate appearances for the season. The dotplots show that the spread of batting averages in 1920 is greater than in 1997. The dotplots also show approximately bell-shaped batting average distributions.

Empirical Rule Illustration

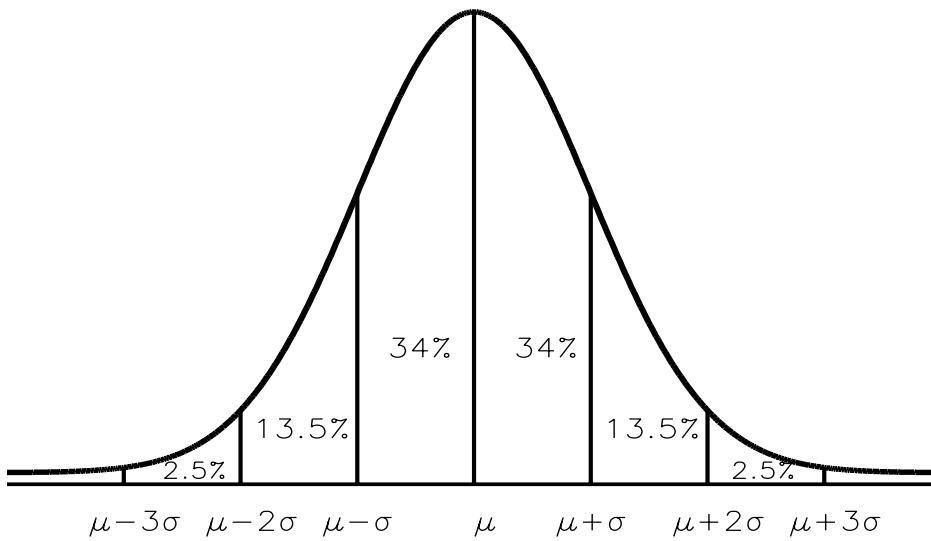
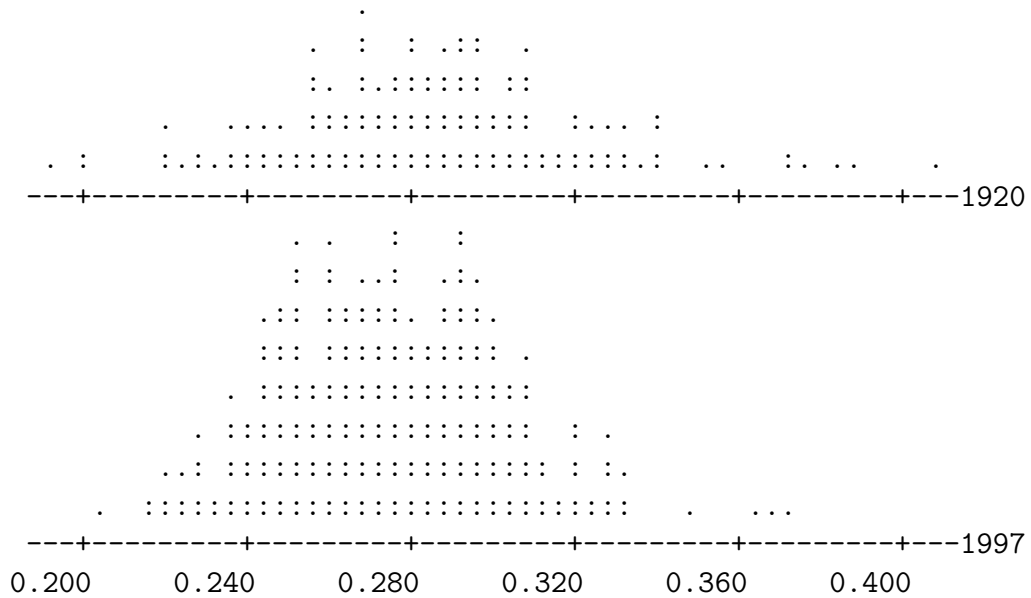


Figure 4: An illustration of the empirical rule.



The mean and standard deviation for league batting averages in 1920 are $\mu_1 = 0.28553$ and $\sigma_1 = 0.03769$ and for 1997, the mean and standard deviation are $\mu_2 = 0.27490$

and $\sigma_2 = 0.02874$. In 1920, a .400 batting average was about three standard deviations above the mean: $(.400 - .28553)/0.03769 = 3.037$. Thus, a .400 batting average in 1920 is pretty extreme but it would not be too unusual to observe one player batting .400 or better and that is what happened in 1920. However, in 1997, a .400 batting average is about 4.35 standard deviations above the mean: $(.400 - .27490)/0.02874$. According to the empirical rule, practically all players will have batting averages within three σ of the mean. In 1997, a .400 batting average was 4.35 standard deviations above the mean indicating that obtaining a .400 batting average that year would have been extraordinarily rare. One of the differences in the statistics between 1920 and 1997 is that the standard deviation of batting averages is smaller. In fact, the standard deviations of batting averages became smaller and smaller from the early 1900's to the end of the century while the overall league batting averages tended to fluctuate up and down. To understand why .400 hitting became extinct, one needs to figure out why the variability in batting averages (and thus the standard deviations) declined as the years went by. Stephen Jay Gould provides a discussion of this in his book *Full House* as well as his explanation for the decreasing variability in batting averages.

1.5 Statistical Models and Statistical Thinking.

What do we mean by a *model* in statistics? To illustrate, consider again the factory that produces cup-a-soups. The factory was having problems with filling the cups consistently to a specified volume. The variable that was measured was the fill-weight, call it y . Let μ denote the specified weight (e.g. $\mu = 238$ oz.). Since we are considering the first $n = 20$ observations, we can denote them y_1, y_2, \dots, y_{20} . Then one of the simplest models possible is

$$y_i = \mu, \quad i = 1, 2, \dots, 20. \quad (2)$$

However, this model is clearly inadequate for the problem at hand. The fill-weights in the cups are not all the same even though the filling machine is set to fill each cup to the same level. That is, there is variability in the process. Variability is a natural consequence of any process and our models need to incorporate the variability. The actual fill-weight of each cup varies randomly due to many factors such as variability in the soup contents, variability in the volume of the filling tank, variability in the exact amount of time the filling tank fills each cup, as well as other factors. The model (2) is insufficient because it does not take the sources of variability into consideration. A more appropriate model incorporates a random error which we shall denote by Greek letter ϵ (“epsilon”) to account for this variability in fill-weights:

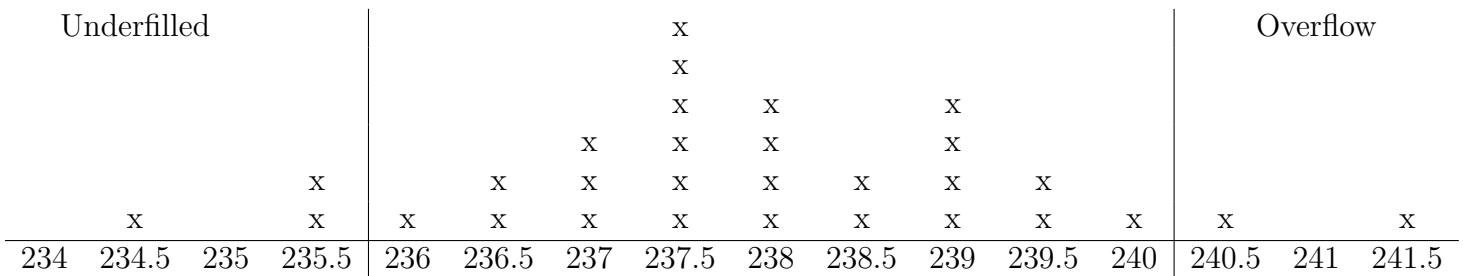
$$y_i = \mu + \epsilon_i, \quad i = 1, 2, \dots, 20. \quad (3)$$

The value μ is known as a *parameter* of the model. Generally model parameters are unknown and must be estimated from the data. The cup-a-soup factory can set the filling machine to fill each cup to a specified volume (and hence weight), but we cannot be sure the machine is doing the right thing until we have collected data and estimated the true value of μ . In (3), μ can be thought of as the “average” or “mean” weight of soup in the cups.

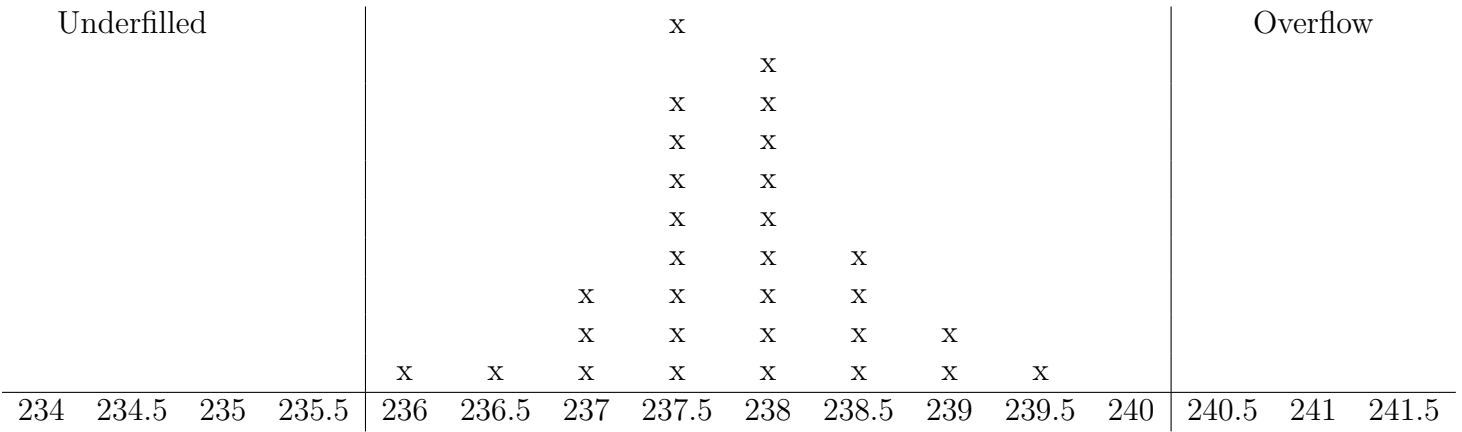
The other part of the model (3) is the random error component ϵ which tells how far the weight in a particular cup differs from the mean weight μ . Ideally all the cups would have exactly the same weight in which case $\epsilon = 0$, but this is never the case in practice. The hallmark of a quality product then is to have very little variability in the product. In the cup-a-soup example, we would want the fill-weights of the cups not to vary much. One of the concepts we shall formally define later is the concept of variability – a mathematical measure of how much items in a population vary from one another.

Below are two plots using hypothetical data for fill-weight data. These plots are known as *dotplots* and provide yet another convenient and easy way of visualizing data. Suppose for illustration that $\mu = 238$ oz. and that we have recorded the fill-weight in $n = 30$ cups. To construct a dotplot, simply draw a horizontal axis and place dots above the axis corresponding to each of the data points in the sample. The first dotplot shows points scattered about $\mu = 238$ but there is quite a bit of variability among the points. There are two types of problems that can occur with the cup-a-soup example: overfilling the cup so that the soup spills over, or you underfilling the cups so that there is too little soup in the cup. Suppose a cup overfills if the weight exceeds 240 and is underfilled if the weight is under 236. The vertical lines in the two plots show the over- and under-fill cut-off points. In the first plot we see some of the cups falling in the problem area of over- and under-filling. However in the second dotplot, the points once again scatter about $\mu = 238$ but they are more tightly clustered about the mean indicating less variability and hence fewer cups with over- and under-filling problems. Because the second plot has less variability in the fill-weights, there are fewer problems with over- and under-filling the cups.

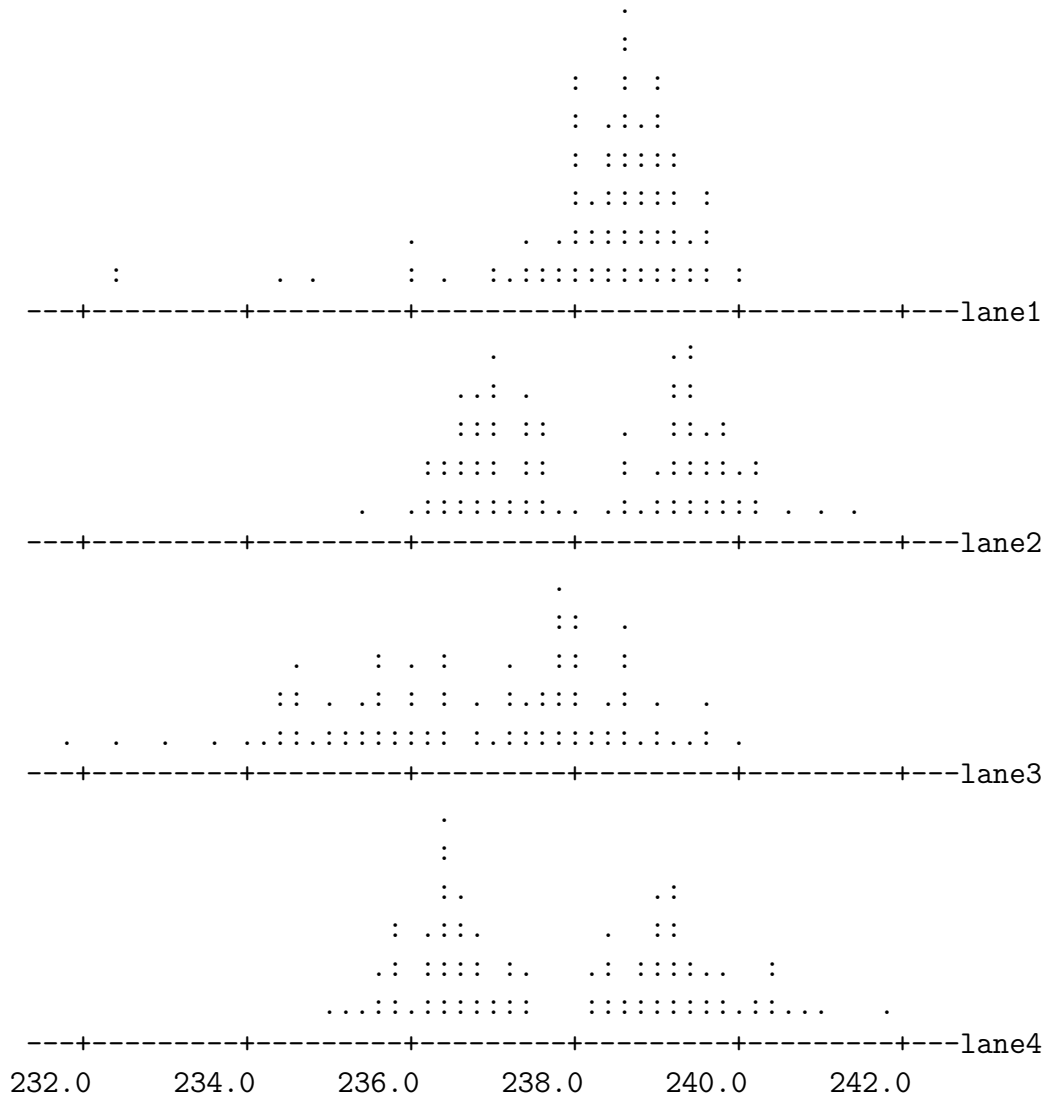
Although the use artificial data in the above plots is useful for illustrating the concept of variability, it is nonetheless somewhat unsatisfactory. The next four dotplots are of the actual data showing the first 100 fill-weights from the four filling lanes in the cup-a-soup example. The four dotplots, generated using the Minitab software, are plotting using the same scale so we can easily compare the four groups.



Fill-weights with high level of variation



Fill-weights with a lower level of variation



Lane 1 appears to have the least variability and in lane 3 the points are spread out quite a bit indicating a high level of variability. Lanes 2 and 4 show a very distinct

bimodal distribution of fill-weights which is a sign that something is seriously wrong with the production process. The average fill-weights of the cups in lane 1 and 2 are 238.29 and 238.21 respectively. These averages are very close together in value, but, as can be seen from the above dotplot, the fill-weight distributions are quite different. Lane 2 has a lot of under and over-filled cups whereas lane 1 does not. This points out that one often has to do more than just look at averages from their data.

One of the key aspects of statistical thinking is to realize that variability is part of any process or any population. We see in the cup-a-soup example that no two cups have exactly the same fill-weight. Analogously, if you look out at a group of people, you see considerable variability in their heights as well as weights and hair color etc. Variability in everyday life makes life interesting. If everybody were exactly the same, then life would be quite dull. However, in a manufacturing setting, variability in the product is not welcome. If I open a can of coke today and another one next week, I want it to taste exactly the same. In order to fully understand a process, one needs to understand that variability is part of the process and hopefully the sources of the variability can be identified. Models for processes or populations need to incorporate a random component that is responsible for the variability in the process.

The model $y = \mu + \epsilon$ is a very simple but useful model. Other applications require more complicated models. Figure 5 shows a *scatterplot* of points from an experiment to examine the shear strength of the bonding between rocket propellants versus the age of the propellant (Montgomery and Peck 1992, page 10). The y -axis is the shear strength of the propellant (in psi) and the x -axis is the age (in weeks) of the propellant. The plot shows data from $n = 20$ propellants. Clearly there is a relationship between the shear strength of the propellant and the age of the propellant. In particular, as the propellant ages, the shear strength grows weaker. Letting x equal the age of the propellant and y equal the shear strength, we could formulate a model such as

$$y = f(x)$$

where $f(x)$ is some unknown function. Looking at the scatterplot of points, a reasonable choice for the function f would be a linear function: $f(x) = \beta_0 + \beta_1 x$ since the points fall roughly in a linear pattern. The terms β_0 and β_1 are the y -intercept and slope of the line respectively. However, this model is inadequate because the points do not fall exactly along a line. They appear to be scattered about some imaginary line. Thus, once again, we need to add a random error component to the model to account for this random variation in the data. Our model becomes

$$y = \beta_0 + \beta_1 x + \epsilon \tag{4}$$

where ϵ is the random error component of the model. The model (4) is known as a *regression model*. One of the statistical goals is to use the data to estimate the parameters β_0 and β_1 . Once these parameters are estimated, we can use to model to predict the shear strength of propellants for future use by knowing the age of the propellant.

Another key aspect of statistical thinking is to realize that the data shown in Figure 5 is based on a sample of $n = 20$ propellants and that the estimates of β_0 and β_1 obtained from the data would be different had we obtained a sample of 20 other propellants.

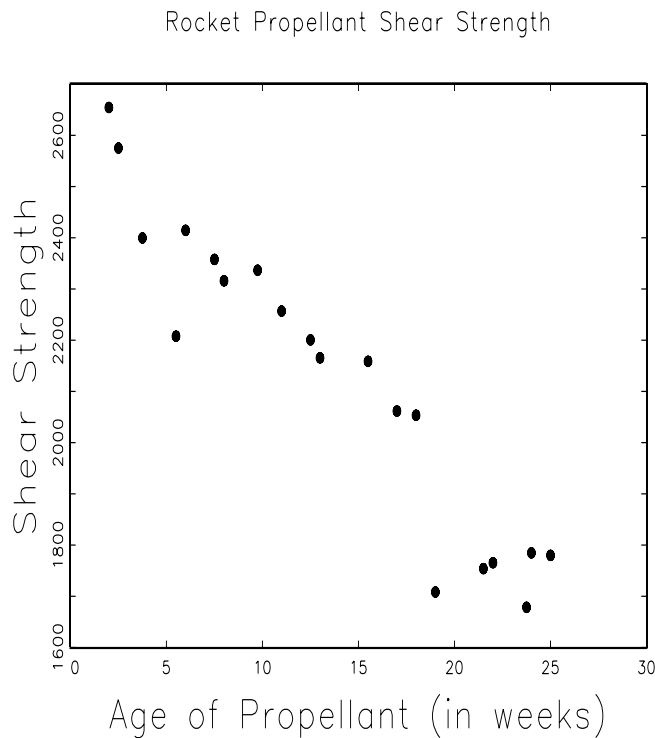


Figure 5: Scatterplot of the shear strength of the bonding of propellants versus age of the propellant.

Thus, the estimates of the parameters can vary depending on the sample that was obtained by random sampling. In order to fully understand a statistical model, we need to know the reliability the parameter estimates. In other words, had we obtained 20 other measurements of shear strength, would we expect the slope and intercept estimates of the parameters in (4) to be nearly the same or to differ substantially from our original estimates? Statistical inference techniques can answer questions like these as we shall see later.

The linear regression model in (4) is still a relatively simple model because linear functions are not complicated. We can consider more complicated models such as a quadratic model:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

or higher order polynomial models as well if the application requires it. Often in practice, the functional relation between two variables x and y is unknown: $y = f(x) + \epsilon$, for some unknown function f . If the function f is fairly well-behaved, then the model may be well approximated by using a Taylor series approximation to f in which case a polynomial model may suffice.

The data shown in Figure 6 indicates that particular applications require very complex modeling strategies. Silverman (1985) analyzed data from a simulated motorcycle accident used to test motorcycle helmets. The acceleration of the crash test dummy's head was measured after impact and the data from the accident is shown in Figure 6. Straight-line models and polynomial models will not suffice for a complicated data

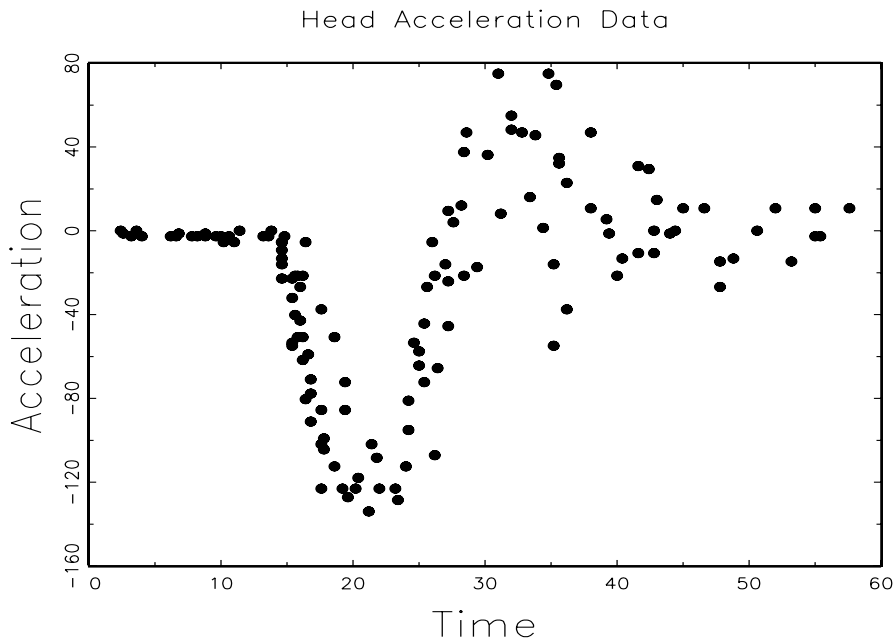


Figure 6: Head acceleration data of a crash test dummy in a simulated motorcycle accident.

set such as this.

In the material that follows, we first will consider discrete models where the measured variable can take only a discrete number of values (either finite or countably infinite). After that we shall consider continuous models where the variable of interest are continuous such as temperature, pressure, and weight measurements. Next, we consider more complicated regression models and finally we look at the problem of comparing populations.

Problems

1. The data below give 29 consecutive measurements on the inner diameter of a landing gear triunion (Cowden 1957).

101.25, 98, 95.5, 97.25, 93.5, 95.25, 93, 85.25, 96.25, 97.5, 94.5
 103.75, 100.5, 109.5, 97.25, 102, 99, 102, 101.25, 101.25, 98.75
 99.75, 100.25, 92.5, 93.75, 97.5, 95.6, 95.75, 99.75

- b) Make a stem-and-leaf plot of this data.
- b) Make a histogram of these diameters and describe the shape of the distribution.

- c) The average of these $n = 29$ measurements is 97.840. In order to assess how much variability there is in the data, a common approach is to look at the squared difference between each observation and the average value: $(y_i - 97.840)^2$, where y_1 is the first measurement, y_2 is the second measurement etc. These squared deviations are given in the following table:

0.026, 0.116, 0.116, 0.348, 0.348, 0.828, 1.346, 2.528, 3.648, 3.648
 4.368, 5.018, 5.476, 5.808, 6.708, 7.076, 11.156, 11.628, 11.628, 11.628
 16.728, 17.306, 17.306, 18.836, 23.426, 28.516, 34.928, 135.956, 158.508

Make a dotplot of these squared deviations and describe the shape of the distribution.

2. The fastest wind velocity (in miles per hour) were recorded for each year in Washington D.C. for the years 1945–1977. The data are listed here (in consecutive order of years):

42, 52, 61, 56, 41, 52, 60, 60, 49, 78, 50,
 57, 63, 40, 42, 54, 57, 42, 52, 43, 49, 38
 42, 43, 45, 43, 47, 43, 38, 42, 43, 50, 50

- a) Make a stem-and-leaf plot of these data.
 b) Make a histogram of these data.
 c) Describe the shape of this distribution. (Note that this is an example of an *extreme value* distribution).
3. A heat flow meter calibration and stability analysis was conducted at the National Institute of Standards and Technology (NIST). The following data are the computed calibration factors for $n = 120$ observations listed in increasing

order. Make a histogram of this data and describe the shape of the distribution.

| | | | | | |
|------|------|------|------|------|------|
| 9.14 | 9.17 | 9.19 | 9.21 | 9.24 | 9.26 |
| 9.15 | 9.17 | 9.19 | 9.21 | 9.24 | 9.26 |
| 9.15 | 9.17 | 9.19 | 9.21 | 9.25 | 9.26 |
| 9.15 | 9.17 | 9.19 | 9.21 | 9.25 | 9.26 |
| 9.15 | 9.17 | 9.19 | 9.22 | 9.25 | 9.26 |
| 9.15 | 9.17 | 9.19 | 9.22 | 9.25 | 9.26 |
| 9.16 | 9.17 | 9.19 | 9.22 | 9.25 | 9.26 |
| 9.16 | 9.18 | 9.19 | 9.22 | 9.25 | 9.27 |
| 9.16 | 9.18 | 9.19 | 9.22 | 9.25 | 9.27 |
| 9.16 | 9.18 | 9.19 | 9.22 | 9.25 | 9.27 |
| 9.16 | 9.18 | 9.20 | 9.23 | 9.25 | 9.27 |
| 9.16 | 9.18 | 9.20 | 9.23 | 9.25 | 9.27 |
| 9.16 | 9.18 | 9.20 | 9.23 | 9.26 | 9.27 |
| 9.16 | 9.18 | 9.20 | 9.23 | 9.26 | 9.28 |
| 9.17 | 9.18 | 9.20 | 9.24 | 9.26 | 9.28 |
| 9.17 | 9.18 | 9.20 | 9.24 | 9.26 | 9.28 |
| 9.17 | 9.18 | 9.21 | 9.24 | 9.26 | 9.28 |
| 9.17 | 9.18 | 9.21 | 9.24 | 9.26 | 9.28 |
| 9.17 | 9.18 | 9.21 | 9.24 | 9.26 | 9.28 |
| 9.17 | 9.18 | 9.21 | 9.24 | 9.26 | 9.28 |

4. The heights of a sample of Wright State University students were obtained (in inches) and they are given in the following table (in ascending order):

| | | | | | | | | |
|----|----|----|----|----|----|----|----|----|
| 54 | 56 | 59 | 60 | 60 | 60 | 60 | 60 | 60 |
| 60 | 61 | 61 | 61 | 62 | 62 | 62 | 62 | 62 |
| 62 | 62 | 62 | 63 | 63 | 63 | 63 | 63 | 63 |
| 63 | 63 | 63 | 63 | 63 | 63 | 63 | 63 | 63 |
| 63 | 63 | 64 | 64 | 64 | 64 | 64 | 64 | 64 |
| 64 | 64 | 64 | 64 | 64 | 64 | 64 | 64 | 64 |
| 64 | 64 | 64 | 65 | 65 | 65 | 65 | 65 | 65 |
| 65 | 65 | 65 | 65 | 65 | 65 | 65 | 65 | 66 |
| 66 | 66 | 66 | 66 | 66 | 66 | 66 | 66 | 66 |
| 66 | 66 | 66 | 66 | 66 | 67 | 67 | 67 | 67 |
| 67 | 67 | 68 | 68 | 68 | 68 | 68 | 68 | 68 |
| 68 | 68 | 69 | 69 | 69 | 69 | 70 | 70 | 70 |
| 70 | 70 | 70 | 70 | 71 | 71 | 71 | 71 | 71 |
| 71 | 72 | 72 | 72 | 72 | 72 | 72 | 72 | 72 |
| 72 | 72 | 72 | 72 | 72 | 73 | 73 | 73 | 74 |
| 74 | 74 | 74 | 74 | 74 | 74 | 74 | 75 | 75 |
| 75 | 75 | 75 | 76 | 76 | 76 | 76 | 76 | 79 |

- a) Make a histogram of this height data using 2 inch measurement classes.
- b) Describe the shape of the distribution. Can you think of a reason for the shape of this distribution? Do you think the group of students from which this data was obtained consisted of more men or more women?

5. A Charpy machine tests the breaking strength of small metal samples. The following are the recorded absorbed energy (in foot-pounds) from a sample of $n = 8$ metal specimens:

67.4, 65.5, 72.0, 73.6, 65.2, 67.0, 66.3, 67.9.

- a) Find the sample mean \bar{y} .
- b) Find the sample median.
- c) Find the sample variance s^2 .
- d) Find the sample standard deviation s .

References

Cowden, D. J. (1957), *Statistical Methods in Quality Control*. Prentice-Hall, Englewood Cliffs.

Gould, S. J. (1996), *Full House : The Spread of Excellence from Plato to Darwin*, Harmony Books.

D.C. Montgomery D. C. and Peck, E. A. (1992), *Introduction to Linear Regression Analysis* 2nd Edition, Wiley, New York.

Silverman, B.W. (1985) Some aspects of the spline smoothing approach to non-parametric curve fitting. *Journal of the Royal Statistical Society, B*, 47, 152.