Producing Data

Lesson 1 focuses on methods of gathering information (data) for statistical use. Every day, we encounter some statistical claim, be it in the form of a political poll or a toothpaste advertisement. We tend to accept these claims as true, without questioning the source or method by which the information was obtained. This chapter explores some reliable methods of retrieving data, as well as some unreliable methods used to produce statistics.

A population is considered to include every individual about whom the information is collected. It is usually impractical to attempt to gather information about every single individual, especially if the population is large. Statistics are based on data gathered from a portion of the population, called a sample. This sample is assumed to represent the population. The level of accuracy with which the sample actually represents the population depends on how the sample is selected. For example, a call-in opinion poll will generate a sample of the population. However, that sample is made up of individuals who volunteer to participate in the poll. Generally, the individuals who volunteer to respond to such polls are often individuals who have very strong opinions, which may not represent the opinion of the population. When a data collection process tends to favor a certain outcome, it is called a biased study.

The best way to avoid bias is to randomize the selection of the sample. Choosing a sample by chance allows all possible groups of individuals the possibility of participating in the study. Such a sample is called a Simple Random Sample, or SRS. Randomization is often achieved by means of a random-digits table. After the individuals in the population are assigned numbers, the table produces a list of numbers so that the sample is chosen randomly. This is similar to drawing numbers out of a hat. Randomization is one of the most important aspects of gathering reliable data.

Observational studies record data as it occurs, without imposing any influence over the outcome. However, experimental studies actually impose some treatment on the individuals in order to observe its affect on the individual. While an observational study is simply attempting to describe a population, an experiment is actually seeking a cause and effect relationship. Essential to any experiment is the presence of a control group. The control group is not subjected to the same treatment as the non-control group. The response of the control group can then be compared to the response of the treatment group, and any significant differences will be observed. Randomization is also important in the selection of samples in an experiment. If the treatments are administered to groups that are very different to begin with, the results will be biased. One must be careful to select sample groups that are similar to each other. This is called a Randomized Comparative Experiment.

Differences observed between two experimental groups may often be misunderstood. When comparing two or more different groups, one would naturally expect some variation among the results. So we must ask the question, at what point are the observed differences meaningful? Is it likely that the observed difference would occur simply by chance? If the answer to the last question is ‘no,’ then we must attribute the difference to some cause or treatment. At this point, we say that our evidence is “statistically significant.”

Exploring Data

Once data has been gathered, it must be organized in a way that makes it easy to understand the data’s implications. For this we use graphs and charts to help visualize the relationships between the pieces of information. A distribution of a variable is simply a list of the values that occur and how often they occur. It is sometimes called a “frequency distribution” because it tells how frequently a certain value occurs. A histogram is a graph that gives the same information as the distribution. Instead of just listing the values, the histogram provides the values that occur, the frequency with which they occur, and a visual comparison between the frequencies of the different values. Looking at a histogram, you can immediately see which value(s) occurred most frequently, and which values occurred infrequently. Stemplots are similar to histograms in that they indicate which values tend to occur more frequently. The histogram and the stemplot for a single set of data should appear to have the same general shape.

Describing a Single Variable

CENTER: The center of the data and the spread of the data are two of the most common descriptive statistics. The center can be described using either the mean (average) or the median. The mean is calculated by adding the values of all the observations and dividing the sum by the number of observations. The median is the value that splits the data into two equal halves. If the data is arranged in order from lowest to highest, exactly half of the data values fall below the median and the other half are above the median. The median is not influenced by outliers (extremely high or extremely low values). However, the mean tends to be pulled toward very high or very low numbers.

SPREAD: The spread of the data can be described either by the quartiles (Q_1 and Q_3) or by the standard deviation. The quartiles are similar to the median as a measure spread. The first quartile (Q_1) is the value for which ¼ of the data falls below and ¾ of the data falls above. The third quartile (Q_3) is the value that leaves ¾ of the data below it and ¼ of the data above. Just as the median splits the data set into two equal halves, the quartiles split the data into four equal quarters.

Minimum Value Q$_1$ Median Q$_3$
Maximum Value
I________________________I________________________I_____
____________I________________________I

25% 25% 25%
25%

The standard deviation is a single number that describes how far the observations fall from the mean. It is roughly based on the average distance all the values are from the mean value. A large standard deviation indicates that the data values are spread out over quite a broad range. However, a smaller standard deviation means that all of the values are within a small distance of the mean value.

Note: The measures of center and spread are used in pairs. If you use the mean to describe the center, you would need to describe the spread using the standard deviation. On the other hand, if you use the median to describe the center, then the quartiles would be used to describe the spread of the distribution. Mean and standard deviation are best used when the distribution is symmetric, while median and the quartiles should be used for a distribution that is skewed. Here we see the value of generating a histogram for a set of data, as qualities like symmetry and skew are easily observed in a graph.

Describing Two Variables at One Time

It is often desirable to compare two variables among a set of individuals. Again, we rely heavily on graphical displays to reveal relationships between the variables. One commonly used graph is called a scatterplot. In a scatterplot, each point represents a single individual. As any set of data is sure to be based on many observations, a scatterplot will include that many points. In some cases involving two variables, a cause and effect relationship is being sought. If this is the case, it is important to distinguish which variable is considered the “cause” and which is considered the “effect.” The “cause” variable is called the explanatory variable, and is always situated along the horizontal axis of the graph. The “effect” variable is called the response variable, and it is situated along the vertical axis of the graph. If no cause and effect relationship is being sought, the variables can be placed on either axis.

The scatterplot demonstrates qualities of the relationship between the variables. These qualities include form, direction and strength. The form of a scatterplot might be linear (following a straight-line pattern), or it might have a distinct curve. The direction of a relationship may be positive or negative. A positive direction occurs when both variables seem to increase together. A negative direction occurs when one variable increases as the other one decreases. The strength of the relationship may be evident on the scatterplot if the points very close together (strong relationship) or very far apart (weaker relationship). However, the graph may be misleading in this area. It is better to determine the strength of the relationship using the correlation, r. The correlation is a number between –1 and 1. A correlation close to 0 indicates a weak relationship. However, a correlation close to 1 or –1 indicates a very strong relationship. (Note that the correlation includes the direction of the relationship as well as the strength. Positive correlation = positive direction. Negative correlation = negative direction.)

A numerical method called regression allows us to mimic the behavior of the data points on a scatterplot with a smooth line or curve (depending on the form of the graph). This linear or non-linear function can then be used to predict the value of the response variable for a specific value of the explanatory variable.

Probability: The Mathematics of Chance

The word probability refers to the predictability of a certain event occurring. Probability affects the way we dress (Is there a probability of rain today?), the way we drive (What route should I take to work in order to avoid traffic tie-ups?), the way we invest our money, and many other decisions we encounter in life. Probability is based on previous observations. Probability describes what has happened in the past and projects that occurrence into the future.

Although probability is often referred to in terms of percents (the probability of rain today is 30%), it is more often expressed in decimal form as a number between 0 and 1. A 30 percent chance of rain equates to a probability of 0.3. An event with probability close to 0 is unlikely to occur. The closer the probability is to 1, the more likely the event. In general, the probability of an event is calculated as follows.

$$P(event) = \frac{number \quad of \quad ways \quad the \quad event \quad can \quad occur}{total \quad number \quad of \quad events}$$

A list of all possible outcomes in a random situation is known as the sample space. A probability model includes a sample space and the probabilities of each of the events in the sample space. The sum of the probabilities of all possible outcomes in the sample space must equal 1. Consider the random situation of tossing a coin. There are only 2 possible outcomes – heads or tails. We can represent the sample space as {H, T}. Since both outcomes are equally likely, our probability model would assign the probabilities of each event as ½. The probability of tossing heads is ½, and the probability of tossing tails is ½. Notice that the sum of the probabilities of all possible outcomes is ½ + ½ = 1. Any probability model can be displayed graphically as a probability histogram, where the highest bar would indicate the outcome with the highest probability.

The mean of a probability model is actually the average of the possible outcomes, and its calculation is based on the probability of each outcome. For example, let's consider the probability model for a spinner in a child's game. The spinner is divided into 6 equal segments labeled with the following numbers: 1,3,2,3,2,3

The Probability model for the spinner would be…

Value on Spinner	1	2	3
Probability	1/6	1/3	1/2

(Note: The sum of all possible outcomes is 1/6 + 1/3 + 1/2 = 1.)

To calculate the mean of the probability model, we multiply each outcome by its probability and add these products together.

Mean of the Probability Model: μ = 1(1/6) + 2(1/3) + 3(1/2) = 2.33333
(The symbol μ is the Greek letter "mu," and is used to represent a mean.)

This tells us that if we were to continue spinning this spinner over and over, and record the value of each spin, the average of all of our spins would be 2.33333. Suppose we only spin the spinner 10 times and the obtain the following numbers:

1, 3, 3, 2, 3, 1, 1, 3, 1, 2
The average value of this sample is 2.

Now let's imagine spinning the spinner another 10 times. Do you expect to get a mean value of 2 again? This time the values of the spinner are:

3,2,1,2,3,1,2,3,3,2,3.
The average value this time is 2.5.

If we repeat this process 100 times, we would have 100 averages to compare. Can you guess what the average of all these averages should be? It should be 2.3333, the value we calculated as μ. This notion of collecting many different samples of the same size and analyzing a statistic from all the trials is referred to as a sampling distribution. It is important to realize that the sampling distribution is looking at the average of many averages.

Now if we were to draw a histogram of the distribution of the averages, we should find that after many trials our graph has a symmetric shape and looks like a bell curve. A distribution that is symmetric and bell shaped is called a normal distribution. The center of the distribution is the true mean, μ. The area of a region under the curve represents the probability of the outcomes within that region. Normal distributions have several convenient qualities that make analyzing the data easier. For instance, 68% of the observations in a normal distribution fall within 1 standard deviation of the mean. This is just part of a well-known principal called The 68-95-99.7 Rule. The rest of the rule tells us that 95% of all the observations in a normal distribution fall within 2 standard deviations of the mean, and 99.7% of the observations fall within 3 standard deviations of the mean. Therefore, if you want to find the interval that includes practically all of the observation values, you would only need to calculate 3 standard deviations below the mean, and 3 standard deviations above the mean. Almost all of the values lie somewhere in between. The convenience of a normal distribution cannot be exaggerated. Fortunately, the Central Limit Theorem assures us that if we use many repeated trials, our sampling distribution will be a normal distribution.

Statistical Inference

Recall the difference between a population and a sample. The population includes every single individual being described. However, a sample is just a portion of the population. If we collect data from a sample of individuals, our result is called a statistic. However, if we are able to collect data on the entire population, we call the result a parameter. A parameter is the true value describing the population. A statistic is just an estimate of that value.

When we hear a statistic reported, how could we be sure that the statistical "estimate" is very close to the true parameter? For normal distributions, we rely on the 68-95-99.7 Rule. Assuming we have a large enough number of repeated trials, the Central Limit Theorem guarantees a sampling distribution that is normal. Let the letter *p* represent the percent of the population that possesses a certain quality of interest. We will refer to the mean of our distribution as $\hat{p}$ (Read as 'p-hat). Remember that our observations are only based on samples of the population, so $\hat{p}$ is a statistic – not the actual parameter, *p*. But the 68-95-99.7 Rule tells us that 95% of all observed samples should fall within 2 standard deviations of $\hat{p}$. So, if $\hat{p}$ isn't the exact value of the parameter, *p*, then we can be 95% sure that the value of *p* is within 2 standard deviations of $\hat{p}$. To put it another way, we are 95% confident than $\hat{p}$ is within 2 standard deviations of the true parameter, p. Our estimate is $\hat{p}$, but we recognize that it may be a little bit above or below the actual parameter. We call this "little bit" the margin of error. The combination of our estimate and its margin of error is called a *confidence interval* because we are 95% confident that the true value of the parameter lies somewhere between these two values.

$$\text{estimate} \pm \text{ margin of error}$$
$$\hat{p} \pm 2 \text{ standard deviations}$$
$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(100-\hat{p})}{n}}$$

The formula $\sqrt{\frac{\hat{p}(100-\hat{p})}{n}}$ gives the standard deviation, and n is the sample size.

It is important to realize that in the situation above, we are studying a quality of the individuals. (Perhaps hair color or political party – something that is not numerical in nature.) To produce numerical data, we can only count the number of individuals that possess that quality and compare it to the number of individuals who do not possess that quality. That is why we use *p* to represent a percentage of the population having the quality of interest.

When the data is already numerical in nature (such as height, age or SAT score), it is not necessary to convert our information into percents. We can simply record the value of each observation in the sample, and calculate the mean value of the observations. This mean is based on a sample, so it is a statistic. We represent the sample mean with the symbol $\bar{x}$ (Read x-bar). The sample mean, $\bar{x}$, is an estimate for the true population parameter, µ. A 95% confidence interval for the mean is subject to the same reasoning by which we developed the confidence interval for p. That is …

$$\text{estimate} \pm \text{ margin of error}$$
$$\bar{x} \pm \text{ 2 standard deviations}$$
$$\bar{x} \pm 2\frac{s}{\sqrt{n}}$$

In this formula, *s* stands for the sample standard deviation, and $\frac{s}{\sqrt{n}}$ estimates the true standard deviation of the population.

Some words of caution: Before drawing conclusions from confidence intervals, be sure to consider the source of the data. How carefully were the samples selected? Were the participants randomized, or was they chosen by convenience or voluntary response? The data must come from a Simple Random Sample to produce a valid confidence interval. Also, the margin of error will not account for nonresponse or other selection biases.