Basic Descriptive Statistics

A Brief Introduction by Chuck Easttom

Introduction

Statistics is a branch of mathematics designed to allow people to accomplish two goals. The first is to accurately describe data and trends in data. The second is to make predictions on future behavior, based on current data. The first goal is simply called descriptive statistics. Any method or formula which yields some number which tells you about a set of data is referred to as descriptive statistics. Any method or formula which discusses a probability of some event occurring is predictive statistics.

In this paper we will discuss descriptive statistics. The goal is to summarize the current level of understanding of basic descriptive statistics and to give some general guidelines for using descriptive statistics. This will be followed by a second paper which discusses probability.

Basic Terminology

Before we can proceed certain terminology must be covered. Without a thorough understanding of these terms it is impossible for any person to be able to study even rudimentary statistics. **Descriptive statistics**: Collection, classification, analysis, and interpretation of data. **Hypothesis**: The idea you are testing. Statistics are usually done in an attempt to confirm or refute some idea. Often in statistics you confirm or refute the null hypothesis, denoted as H_0. It is the hypothesis that essentially the results you get are random and are not due to some real relationship. In other words if the null hypothesis is true, then the apparent relationship is really simply a random coincidence.

Predictive statistics: Using statistics generated from the sample in order to make predictions, this is also often called *inferential statistics.*

Parameter: This is a descriptive number about a population. A statistic is a descriptive number about a sample.

Population: The target group you wish to study, such as all men aged 30 to 40.

Sample: The subgroup from the population you select to study, in order to make inferences about the population.

Types of Measurement Scales[1]

1. Nominal: For qualitative data with distinct categories. For example the categories German, French, and Italian are categories but are not ordered in any way.

2. Ordinal: For qualitative data with distinct categories in which ordering (or ranking) is implied. A good example is the Likert scale that you see on many surveys: 1=Strongly disagree; 2=Disagree; 3=Neutral; 4=Agree; 5=Strongly agree.

3. Interval: For quantitative data with an ordered scale in which the interval between data values is meaningful. For example the categories of rank in the military. Clearly a major is higher ranked that a captain, but how much higher? Does he have twice the authority of a captain? It is impossible to say. You can only say he is higher ranked.

4. Ratio: For quantitative data which have an inherently defined zero and the ratio of data values is meaningful. Weight in kilograms is a very good example since it has a definite ratio from one weight to another. 50kg is indeed twice as heavy as 25 kg.

Data Collection

Normally statistics are done with only a fraction of the actual group being considered. That fraction is called the sample, and the group in question is the population. For example if you

wish to find out if men who are over 40 and more then 30 lbs over weight have an increased risk of heart attack, you might select 1000 men to study. The 1000 selected would be your sample, all men over 40 who are more than 30 lbs over weight would be your population.

This leads to two obvious questions. Is the sample size you selected large enough and is the sample truly representative of the population you are attempting to measure? The first question is always a controversial one. Obviously the larger the sample size the better. However it is often impractical to get very large sample sizes. For example when political polls attempt to predict the outcome of an election, it is almost impossible to get more than a few thousand peoples opinions. Considering that the United states has a population of 270 million, and tens of millions of eligible voters, it is questionable whether or not such a sample provides is accurate. One way around this is to periodically repeat the study. For example with political polls you may only poll 1000 people, but if you do this many times with the same or similar results each time your results have greater validity.

The second question, whether or not your sample is actually representative of the population you are trying to measure, is much easier to answer. There are some very specific ways in which you should select a sample. Using proper sampling techniques will give your statistical analysis credibility. The Statistics Glossary[2], lists several sampling techniques, each is described here:

Independent Sampling: This occurs when multiple samples are taken, but each sample has no effect on any other.

Random Sampling: This occurs when subjects for your sample are picked totally at random with no other factors influencing their selection. For example when names are drawn from a hat, you have random sampling.

Stratified Random Sampling: In this process the population is divided into layers based on some criteria and a number of random subjects are taken from each strata. In our example of studying men over 40 and over weight you might break the population into strata based on how much over weight they are, or how old they are. For example you might have men that are 25 to 50 lbs over weight in one strata and those who are 50 to 100 lbs over weight in another, then finally those who are more than 100 lbs over weight.

There are other sampling methods but these are very commonly used. If you wish to learn more about sampling methods, the following websites will be helpful:

- Stat Pac http://www.statpac.com/surveys/sampling.htm
- Statistics Finland http://www.stat.fi/tk/tt/laatuatilastoissa/lm020500/pe_en.html
- Australian Bureau of Statistics http://www.abs.gov.au/websitedbs/D3310116.NSF/4a255eef008309e44a255eef00061e57/116e0f93f17283eb4a2567ac00213517!OpenDocument

When evaluating any statistical analysis it is important to consider how the sampling was done, and if the sample size seems large enough to be relevant. It might even be prudent to never rely on a single statistical study. If multiple studies of the same population parameter, using different samples, yield the same or similar results, then one has a compelling body of data. A single study always has a chance of being simply an anomaly, no matter how well the study was conducted.

Measures of Central Tendency

The first and simplest sort of descriptive statistics involves measures of central tendency. This is simply a way of seeing what the aggregate of the data tells us about the data. The three most simple measures of central tendency are the mean, median, and mode. The *mean* is simply the

arithmetic average, the *mode* is the item in the sample that appears most often, and the *median* is the item that appears in the middle. Let me illustrate. Assume you had a set of test scores as follows:

65, 74, 84, 84, 89,91,93,99,100

The mode is easy, 84 is the only score that appears more than once.

The median is the score in the center, which in this case is 89

The mean is found by adding the scores and dividing by the number of scores (in this case 9). The formula for that is mean $x = \Sigma x/n$. In this case it would be 86.55

Another important term is *range*. The range is simply the distance from the lowest score to the highest. In our example the highest is 100, the lowest is 65, thus the range is 35.

You will see these four numbers ubiquitously presented in statistical studies. However what do they really tell us. In this case the arithmetic mean of the scores was actually about dead center of the scores. In our case, all must two of our scores are grouped in a narrow range from 84 to 100. This clustering means that our measures of central tendency probably tell us a lot about our data. But what about situations with much more variety in the numbers? In such cases, the mean may not tell us much about the actual data. This leads to other measures we can do, which can indicate just how accurate the mean is. The standard deviation is a measurement that will tell you this. To quote a popular statistics website[3]

"The **standard deviation** is a statistic that tells you how tightly all the various examples are clustered around the mean in a set of data. When the examples are pretty tightly bunched together and the bell-shaped curve is steep, the standard deviation is small. When the examples are spread apart and the bell curve is relatively flat, that tells you, that you have a relatively large standard deviation."

The standard deviation (denoted s) is the square root of the sum of the variance divided by the number of elements – 1. Put in simpler terms you take each item in the sample, and see how far it varies from the mean. You then square that value and divide it by the number in the sample – 1. Take the square root of that number (which is called the variance) and you have the standard deviation4. Something like this

$$s = \sqrt{\left(\Sigma(x^{i} - \text{mean})2\right)/\ n\text{-}1}$$

In our example we would take each item in the sample minus the mean of 86.5 and square that difference and total the results. Like this:

$(65 - 86.5)^2 + (74 - 86.5)^2 + (84 - 86.5)^2 + (84 - 86.5)^2 + (89 - 86.5)^2 + (91 - 86.5)^2 + (93 - 86.5)^2 + (99 - 86.5)^2 + (100 - 86.5)^2$

which is equal to:

462.25 + 156.25 + 6.25 + 6.25 + 6.25 + 20.25 + 42.25 + 156.25 + 182.25 = 882

Now divide that by n – 1 (n = 9 so divide by 8) and you get 110.25 which is the variance.
The square root of the variance, in this case 10.5 is the standard deviation. That means, in plain English, that on average, the various scores where about 10.5 units from the mean.
So you can see that standard deviation and variance are simply arithmetical computations done to compare the individual items in the sample, to the mean. They give you some idea about the data. If there is a small standard deviation that indicates that values where clustered near the mean and that the mean is representative of the sample.

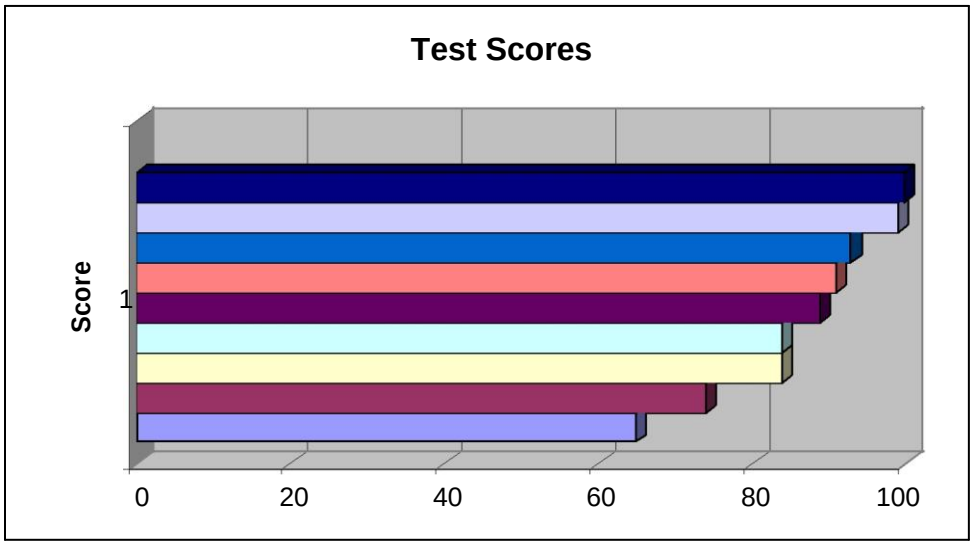
These are not the only means of measuring central tendency but they are the most commonly used. Virtually all statistical studies will report mean, median, mode, range, standard deviation, and variance.

Graphical Representation of Data

There are several different ways to represent data, and each of these has certain advantages and disadvantages. Lets use our example test score data and show some different graphs.

Bar Chart

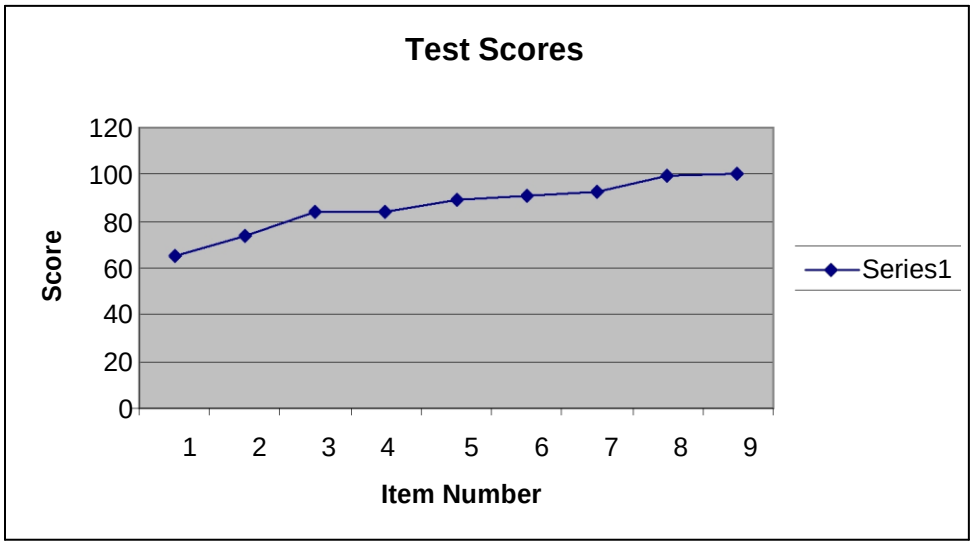
This is probably the most common chart, everyone is familiar with it. Here is a bar chart with our sample test scores:



This sort of chart gives a very obvious comparison of values. You often see it when monetary comparisons are made.

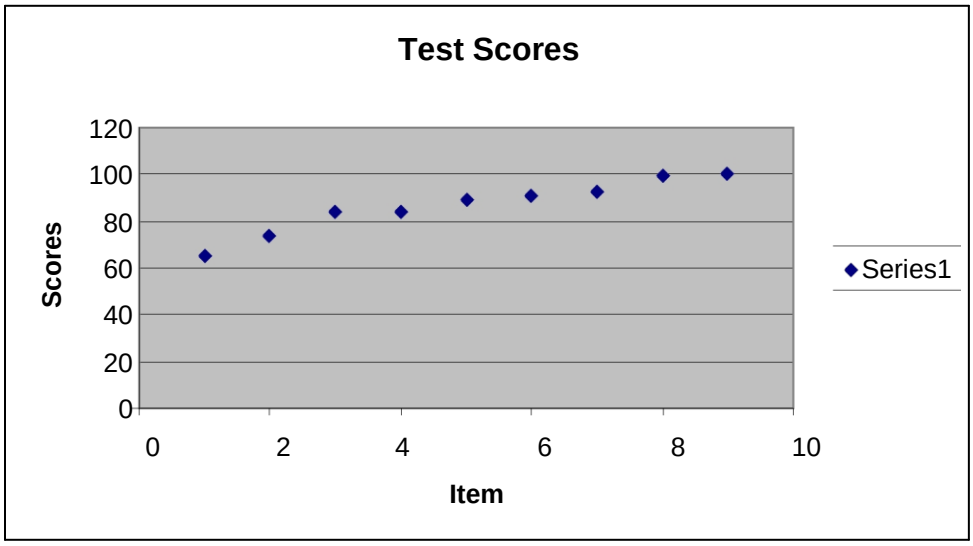
Line Chart

This chart will show individual items in comparison. It is most often used to show changes in some value over time (such as profit per year). However to illustrate we will use it with our test score example:



Scatter Plot

This shows each dot representing each item in the sample. Sometimes a 'best fit' line is drawn to connect the dots. This is very useful in also getting a pictorial view of measures of central tendency. If you have tight grouping of the dots, you probably have low standard deviation. Wide spread dots mean high standard deviation.



Stem and Leaf Display

This is often used in statistics and can be very informative. Essentially you take a series of values, like our scores. You divide them into large numerical groupings which provide the stem. In our example the 10's place would be the stem. What is left are the leaves.

65, 74, 84, 84, 89,91,93,99,100 so our stems are 6,7,8,9,10 and our stem and leaf display is:

6	5
7	4
8	4 4 9
9	1 3 9
10	0

You can see that the benefit of this type of graphic display is that it clearly shows groupings of items in the sample. For example in our group of test scores you can see that much of the sample is grouped in stem's 8 and 9. This gives you some pictorial insight into the measures of central tendency.

Frequency Distribution

A frequency distribution shows the number of items falling into each of several ranges of values. Frequency distributions can show either the actual number of observations falling in each range or the percentage of observations. In the latter case the distribution is called a relative frequency distribution[5]. In essence this means that a frequency distribution is related to a stem and leaf display as it tells you how often a given item in a sample occurs. Sometimes you will see something like a line graph, but instead of the individual items from a sample you will see the frequencies plotted.

Correlation

After a study you have may have two variables, lets call them x and y. The question is how much of a correlation do they have? What is the relationship between them. There are a few statistical methods for calculating this. The psychology statistics page at Southwest Missouri State University puts it this way[6]

"The *Pearson Product-Moment Correlation Coefficient* (r), or correlation coefficient for short *is a measure of the degree of linear relationship between two variables*, usually labeled X and Y. While in regression the emphasis is on predicting one variable from the other, in correlation the emphasis is on the degree to which a linear model may describe the relationship between two variables."

A more clear and concise definition can be found at the BMJ (a medical journal) website[7]

"The word correlation is used in everyday life to denote some form of association. We might say that we have noticed a correlation between foggy days and attacks of wheeziness. However, in statistical terms we use correlation to denote association between two quantitative variables."

One valuable way to calculate this is via Pearson's correlation coefficient, usually denoted with a lower case r. The formal for this is

Or re-arranged this yields

$$r = \frac{\Sigma xy - \bar{n}\bar{x}\bar{y}}{(n-1)SD(x)SD(y)}$$

In short you follow these steps

1. Take each value of x and multiply it by each value of y.
2. Multiply n times the mean of x and the mean of y.

3. Subtract the second number from the first and you have the numerator of our equation.
4. Now multiple n-1 times the standard deviation of x and the standard deviation of y and you now have the numerator.
5. Do the division and you have Pearson's correlation coefficient.

Lets work out an example. Lets assume you have x variable of years of post secondary education, and y variable as annual income in tens of thousands. Lets do a small sample to make the math simpler:

Years of post secondary education	Annual salary in 10's of thousands
2	4
3	4
4	5
4	7
8	10

Now follow our steps

1. $\Sigma xy = 148$
2. mean x = 4.2 mean y = 6 n =5 so 6 * 5 * 4.2 = 126
3. 148-126 = 22
4. The standard deviation of x is 2.28 the standard deviation of y is 2.54 and n-1 = 4 so we have 4 * 2.28 * 2.54 = 23.164
5. 22/23.164 =.949 as our Pearson's correlation coefficient.

Now we have calculated the value of r, but what does it mean? Values of r will always be between –1.0 and positive 1.0[8]. A –1.0 would mean a perfect negative correlation where as a

+1.0 would indicate a perfect positive correlation. So a value of .949 indicates a very strong positive correlation between years of post secondary education and annual income.

In our limited sample we found a high positive correlation but how significant is out finding? That is where the T test comes in, and fortunately the arithmetic operations for it are much simpler. The equation is given here:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

It simply states that t is equal to the Pearson correlation coefficient multiplied by the square root of the quotient n-2/1-r^2 The value r^2 represents the proportion of common variables between the dependent and independent variables. This value is sometimes referred to as the coefficient of determination[9].

In our scenario we have n = 5 and r = .949 so we have $\sqrt{}$(3/ .099399) which is 5.493 which we multiply by r to get 5.213 for T. This T- score gives us some idea of how significant our correlation coefficient is.

P-Value

Another way to determine if the relationship between our variables is statistically significant is the P-value. According to the Math World[10] website a p-value is "The probability that a variate would assume a value greater than or equal to the observed value strictly by chance:"

$$P(z \geq z_{\text{observed}})$$

The six sigma site[11] explains p-values this way

"Each statistical test has an associated null hypothesis, the p-value is the probability that your sample could have been drawn from the population(s) being tested (or that a more improbable

sample could be drawn) given the assumption that the null hypothesis is true. A p-value of .05, for example, indicates that you would have only a 5% chance of drawing the sample being tested if the null hypothesis was actually true.

Null Hypothesis are typically statements of no difference or effect. A p-value close to zero signals that your null hypothesis is false, and typically that a difference is very likely to exist. Large p-values closer to 1 imply that there is no detectable difference for the sample size used. A p-value of 0.05 is a typical threshold used in industry to evaluate the null hypothesis. In more critical industries (healthcare, etc.) a more stringent, lower p-value may be applied."

Z-Test

The Z test is used to make inferences about data. Its formula is shown here

$$z = \frac{\mu - x}{\sigma}$$

- σ (the standard deviation of the population)
- μ (the mean of the population)
- *x* (the mean of the sample)

So the z is equal to the mean of the population minus the mean of the sample divided by the standard deviation of the population. The z test is often used when you wish to compare one sample mean against one population mean. But what does it actually tell us? To begin with you need to select a threshold of validity for your statistical study. The Z score converts your raw data (your sample mean) into a standardized score that can be used (along with your pre selected threshhold of rejection) to determine if you should reject the null hypothesis or not. Z scores alwas have a mean of zero and a standard deviation of 1.

Most statistics text books will have a Z score chart. You can find what percentage of sample items should appear above or below your Z score. You now compare that to your pre selected

rejection level. If you previously decided that a 5% level would cause you to reject the null hypothesis and your Z score indicates 11 %, then you will reject the null hypothesis. So the Z score is essential in helping you to decide whether or not to reject the null hypothesis. For more information on the Z Score try these resources:

- http://www.animatedsoftware.com/statglos/sgzscore.htm
- http://www.stat.sc.edu/~ogden/javahtml/power/power.html
- http://www.stedwards.edu/bss/swinkels/SEUWebpage/CoursePages/StatisticsCourse/ztut/ztutmain.htm

Outliers

A significant problem for any statistical study is the existence of outliers. An outlier is a value that lies far outside where most other values lie. For example if a sample of high school athletes all have heights ranging from 70 inches to 75 inches, except for one who has a height of 81 inches, that one will skew all statistics that you generate. His height will make the mean much higher and the standard deviation much wider. Statistics generated by including that data point in the sample will not accurately reflect the sample. So what can be done?

One solution to outliers is to exclude them from the data set. It is common to exclude data points that are 2 or more standard deviations from the mean. So if the mean is 74 inches, and the standard deviation is 3 inches, then any height that is more than 80 inches or less than 68 inches is excluded from the data set. There are, of course, varying opinions on just how far from the mean constitutes a standard deviation, and whether or not they should be excluded at all. If you are conducting a study and elect to exclude outliers, it is a good idea to indicate that in your study, and what method you chose to exclude outliers.

Statistical Errors

Obviously statistical errors can and do occur. Generally such errors can be classified as either type I or type II errors[12]. A type I error occurs when the null hypothesis is rejected when in fact it is true. A type II error occurs when the null hypothesis is accepted when in fact it is false. The following table illustrates this:

Table 1: Statistical Errors

	Decision		
		Reject H0	Accept H0
Truth	H0 true	Type I Error	Correct Decision
	H1 true	Correct Decision	Type II Error

Power of a test

The power of a statistical hypothesis test measures the test's ability to reject the null hypothesis when it is actually false - that is, to make a correct decision. In other words, the power of a hypothesis test is the probability of not committing a type II error. It is calculated by subtracting the probability of a type II error from 1, usually expressed as:

Power = 1 - P(type II error) =

The maximum power a test can have is 1, the minimum is 0. the ideal situation is to have high power, close to 1.

Foot Notes

1. The Children's Mercy Hospital Statistics Definitions, accessed in September 2004 http://www.cmh.edu/stats/definitions.asp

2. The Statistics Glossary, accessed in October 2004,

   http://www.stats.gla.ac.uk/steps/glossary/sampling.html

3. Robert Niles Statistics Page, accessed in October 2004,

   http://www.robertniles.com/stats/stdev.shtml

4. A Data Based Approach to Statistics, Iman, Duxbury Press 1994

5. Rice University, http://www.ruf.rice.edu/~lane/hyperstat/A26308.html

6. Southwest Missouri State Psychology Department,

   http://www.psychstat.smsu.edu/introbook/sbk17.htm

7. BMJ Journal online site, http://bmj.bmjjournals.com/collections/statsbk/11.shtml

8. Winks Statistics Online, http://www.texasoft.com/winkpear.html

9. Statsoft's basic statistics online textbook,

   http://www.statsoft.com/textbook/stbasic.html#Correlationsb

10. Math World, http://mathworld.wolfram.com/P-Value.html

11. Six Sigma, http://www.isixsigma.com/dictionary/P-Value-301.htm

12. The Statistics Glossary, http://www.cas.lancs.ac.uk/glossary_v1.1/hyptest.html#1err