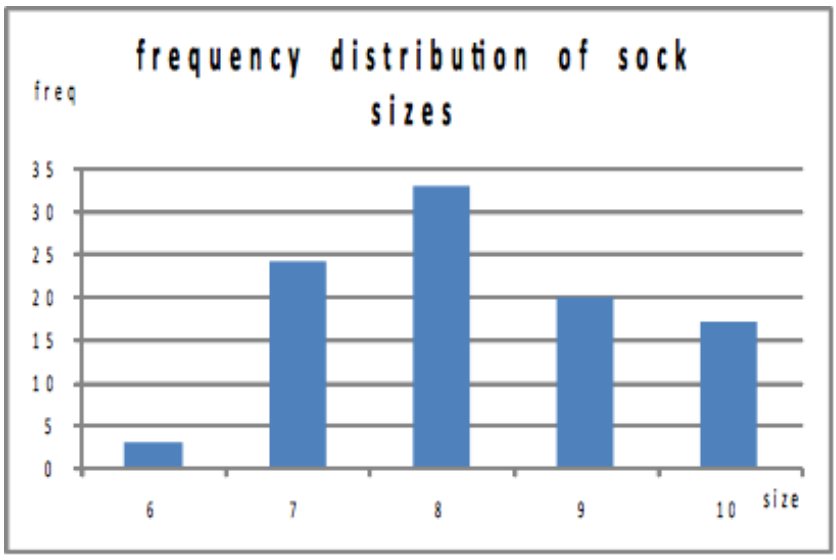
1. Descriptive statistics and frequency distributions

This chapter is about describing populations and samples, a subject known as descriptive statistics. This will all make more sense if you keep in mind that the information you want to produce is a description of the population or sample as a whole, not a description of one member of the population. The first topic in this chapter is a discussion of "distributions", essentially pictures of populations (or samples). Second will be the discussion of descriptive statistics. The topics are arranged in this order because the descriptive statistics can be thought of as ways to describe the picture of a population, the distribution.

Distributions

The first step in turning data into information is to create a distribution. The most primitive way to present a distribution is to simply list, in one column, each value that occurs in the population and, in the next column, the number of times it occurs. It is customary to list the values from lowest to highest. This is simple listing is called a "frequency distribution". A more elegant way to turn data into information is to draw a graph of the distribution. Customarily, the values that occur are put along the horizontal axis and the frequency of the value is on the vertical axis.

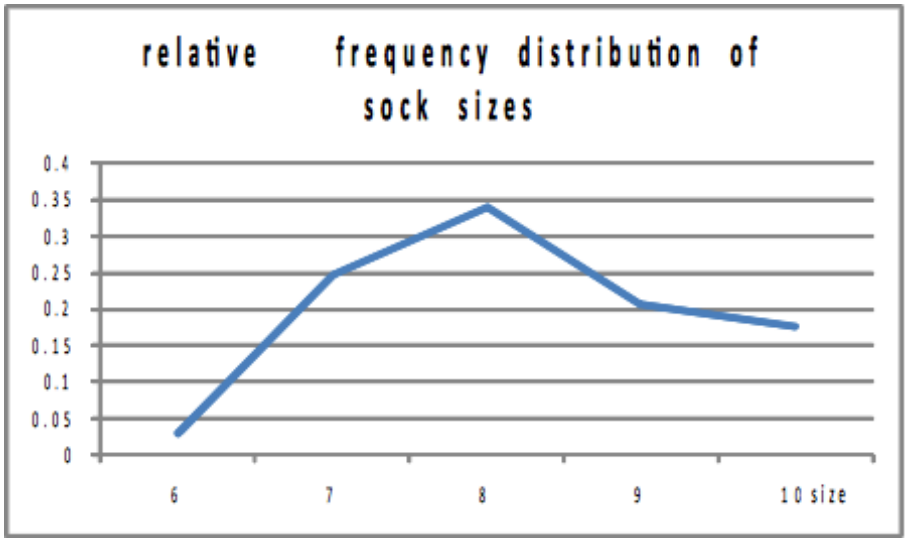
Ann Howard called the equipment manager at two nearby colleges and found out the following data on sock sizes used by volleyball players. At Piedmont State last year, 14 pairs of size 7 socks, 18 pairs of size 8, 15 pairs of size 9, and 6 pairs of size 10 socks were used. At Graham College, the volleyball team used 3 pairs of size 6, 10 pairs of size 7, 15 pairs of size 8, 5 pairs of size 9, and 11 pairs of size 10. Ann arranged her data into a distribution and then drew a graph called a Histogram:



Size	Frequency
6	3
7	24
8	33
9	20
10	17

Exhibit 1: Frequency graph of sock sizes

Ann could have created a relative frequency distribution as well as a frequency distribution. The difference is that instead of listing how many times each value occurred, Ann would list what proportion of her sample was made up of socks of each size:



Size	freq.	rel. freq
6	3	.0319
7	24	.2553
8	33	.3510
9	20	.2127
10	17	.1808

Exhibit 2: Relative frequency graph of sock sizes

Notice that Ann has drawn the graphs differently. In the first graph, she has used bars for each value, while on the second, she has drawn a point for the relative frequency of each size, and the "connected the dots". While both methods are correct, when you have a values that are continuous, you will want to do something more like the "connect the dots" graph. Sock sizes are **discrete**, they only take on a limited number of values. Other things have **continuous** values, they can take on an infinite number of values, though we are often in the habit of rounding them off. An example is how much students weigh. While we usually give our weight in whole pounds in the US ("I weigh 156 pounds."), few have a weight that is exactly so many pounds. When you say "I weigh 156", you actually mean that you weigh between 155 1/2 and 156 1/2 pounds. We are heading toward a graph of a distribution of a continuous variable where the relative frequency of any **exact** value is very small, but the relative frequency of observations between two values is measurable. What we want to do is to get used to the idea that the total area under a "connect the dots" relative frequency graph, from the lowest to the highest possible value is one. Then the part of the area under the graph between two values is the relative frequency of observations with values within that range. The height of the line above any particular value has lost any direct meaning, because it is now the area under the line between two values that is the relative frequency of an observation between those two values occurring.

You can get some idea of how this works if you go back to the bar graph of the distribution of sock sizes, but draw it with relative frequency on the vertical axis. If you arbitrarily decide that each bar has a width of one, then the area "under the curve" between 7.5 and 8.5 is simply the height times the width of the bar for sock size 8: 0.3510 x 1. If you wanted to find the relative frequency of sock sizes between 6.5 and 8.5, you could simply add together the area of the bar for size 7 (that's between 6.5 and 7.5) and the bar for size 8 (between 7.5 and 8.5).

Descriptive statistics

Now that you see how a distribution is created, you are ready to learn how to describe one. There are two main things that need to be described about a distribution: its location and its shape. Generally, it is best to give a single measure as the description of the location and a single measure as the description of the shape.

Mean

To describe the location of a distribution, statisticians use a "typical" value from the distribution. There are a number of different ways to find the typical value, but by far the most used is the "arithmetic mean", usually simply called the "mean". You already know how to find the arithmetic mean, you are just used to calling it the "average". Statisticians use average more generally—the arithmetic mean is one of a number of different averages. Look at the formula for the arithmetic mean:

$$\mu = \frac{\Sigma x}{N}$$

All you do is add up all of the members of the population, ∑x, and divide by how many members there are, N. The only trick is to remember that if there is more than one member of the population with a certain value, to add that value once for every member that has it. To reflect this, the equation for the mean sometimes is written :

$$\mu = \frac{\Sigma f_i x_i}{N}$$

where f_i is the frequency of members of the population with the value x_i .

This is really the same formula as above. If there are seven members with a value of ten, the first formula would have you add seven ten times. The second formula simply has you multiply seven by ten—the same thing as adding together ten sevens.

Other measures of location are the median and the mode. The median is the value of the member of the population that is in the middle when the members are sorted from smallest to largest. Half of the members of the population have values higher than the median, and half have values lower. The median is a better measure of location if there are one or two members of the population that are a lot larger (or a lot smaller) than all the rest. Such extreme values can make the mean a poor measure of location, while they have little effect on the median. If there are an odd number of members of the population, there is no problem finding which member has the median value. If there are an even number of members of the population, then there is no single member in the middle. In that case, just average together the values of the two members that share the middle.

The third common measure of location is the mode. If you have arranged the population into a frequency or relative frequency distribution, the mode is easy to find because it is the value that occurs most often. While in some sense, the mode is really the most typical member of the population, it is often not very near the middle of the population. You can also have multiple modes. I am sure you have heard someone say that "it was a bimodal distribution". That simply means that there were two modes, two values that occurred equally most often.

If you think about it, you should not be surprised to learn that for bell-shaped distributions, the mean, median, and mode will be equal. Most of what statisticians do with the describing or inferring the location of a population is done with the mean. Another thing to think about is using a spreadsheet program, like Microsoft Excel when arranging data into a frequency distribution or when finding the median or mode. By using the sort and

distribution commands in 1-2-3, or similar commands in Excel, data can quickly be arranged in order or placed into value classes and the number in each class found. Excel also has a function, =AVERAGE(...), for finding the arithmetic mean. You can also have the spreadsheet program draw your frequency or relative frequency distribution.

One of the reasons that the arithmetic mean is the most used measure of location is because the mean of a sample is an "unbiased estimator" of the population mean. Because the sample mean is an unbiased estimator of the population mean, the sample mean is a good way to make an inference about the population mean. If you have a sample from a population, and you want to guess what the mean of that population is, you can legitimately guess that the population mean is equal to the mean of your sample. This is a legitimate way to make this inference because the mean of all the sample means equals the mean of the population, so, if you used this method many times to infer the population mean, on average you'd be correct.

All of these measures of location can be found for samples as well as populations, using the same formulas. Generally,μ is used for a population mean, and $\bar{x}$ is is used for sample means. Upper-case N, really a Greek "nu", is used for the size of a population, while lower case n is used for sample size. Though it is not universal, statisticians tend to use the Greek alphabet for population characteristics and the Roman alphabet for sample characteristics.

Measuring population shape

Measuring the shape of a distribution is more difficult. Location has only one dimension ("where?"), but shape has a lot of dimensions. We will talk about two,and you will find that most of the time, only one dimension of shape is measured. The two dimensions of shape discussed here are the width and symmetry of the distribution. The simplest way to measure the width is to do just that—the range in the distance between the lowest and highest members of the population. The range is obviously affected by one or two population members which are much higher or lower than all the rest.

The most common measures of distribution width are the standard deviation and the variance. The standard deviation is simply the square root of the variance, so if you know one (and have a calculator that does squares and square roots) you know the other. The standard deviation is just a strange measure of the mean distance between the members of a population and the mean of the population. This is easiest to see if you start out by looking at the formula for the variance:

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{N}$$

Look at the numerator. To find the variance, the first step (after you have the mean, μ) is to take each member of the population, and find the difference between its value and the mean; you should have N differences. Square each of those, and add them together, dividing the sum by N, the number of members of the population. Since you find the mean of a group of things by adding them together and then dividing by the number in the group, the variance is simply the "mean of the squared distances between members of the population and the population mean".

Notice that this is the formula for a population characteristic, so we use the Greek σ and that we write the variance as σ^2, or "sigma square" because the standard deviation is simply the square root of the variance, its symbol is simply "sigma", σ.

One of the things statisticians have discovered is that 75 per cent of the members of any population are with two standard deviations of the mean of the population. This is known as Chebyshev's Theorem. If the mean of a

population of shoe sizes is 9.6 and the standard deviation is 1.1, then 75 per cent of the shoe sizes are between 7.4 (two standard deviations below the mean) and 11.8 (two standard deviations above the mean). This same theorem can be stated in probability terms: the probability that anything is within two standard deviations of the mean of its population is .75.

It is important to be careful when dealing with variances and standard deviations. In later chapters, there are formulas using the variance, and formulas using the standard deviation. Be sure you know which one you are supposed to be using. Here again, spreadsheet programs will figure out the standard deviation for you. In Excel, there is a function, =STDEVP(...), that does all of the arithmetic. Most calculators will also compute the standard deviation. Read the little instruction booklet, and find out how to have your calculator do the numbers before you do any homework or have a test.

The other measure of shape we will discuss here is the measure of "skewness". Skewness is simply a measure of whether or not the distribution is symmetric or if it has a long tail on one side, but not the other. There are a number of ways to measure skewness, with many of the measures based on a formula much like the variance. The formula looks a lot like that for the variance, except the distances between the members and the population mean are cubed, rather than squared, before they are added together:

$$sk = \frac{\Sigma(x - \mu)^3}{N}$$

At first it might not seem that cubing rather than squaring those distances would make much difference. Remember, however, that when you square either a positive or negative number you get a positive number, but that when you cube a positive, you get a positive and when you cube a negative you get a negative. Also remember that when you square a number, it gets larger, but that when you cube a number, it gets a whole lot larger. Think about a distribution with a long tail out to the left. There are a few members of that population much smaller than the mean, members for which $(x-\mu)$ is large and negative. When these are cubed, you end up with some really big negative numbers. Because there are not any members with such large, positive $(x-\mu)$, there are not any corresponding really big positive numbers to add in when you sum up the $(x-\mu)^3$, and the sum will be negative. A negative measure of skewness means that there is a tail out to the left, a positive measure means a tail to the right. Take a minute and convince yourself that if the distribution is symmetric, with equal tails on the left and right, the measure of skew is zero.

To be really complete, there is one more thing to measure, "kurtosis" or "peakedness". As you might expect by now, it is measured by taking the distances between the members and the mean and raising them to the fourth power before averaging them together.

Measuring sample shape

Measuring the location of a sample is done in exactly the way that the location of a population is done. Measuring the shape of a sample is done a little differently than measuring the shape of a population, however. The reason behind the difference is the desire to have the sample measurement serve as an unbiased estimator of the population measurement. If we took all of the possible samples of a certain size, n, from a population and found the variance of each one, and then found the mean of those sample variances, that mean would be a little smaller than the variance of the population.

You can see why this is so if you think it through. If you knew the population mean, you could find $\sum(x-\mu)^2/n$ for each sample, and have an unbiased estimate for σ². However, you do not know the population mean, so you will have to infer it. The best way to infer the population mean is to use the sample mean $\bar{x}$. The variance of a sample will then be found by averaging together all of the $\sum(x-\bar{x})^2/n$.

The mean of a sample is obviously determined by where the members of that sample lie. If you have a sample that is mostly from the high (or right) side of a population's distribution, then the sample mean will almost for sure be greater than the population mean. For such a sample, $\sum(x-\bar{x})^2/n$ would underestimate σ². The same is true for samples that are mostly from the low (or left) side of the population. If you think about what kind of samples will have $\sum(x-\bar{x})^2/n$ that is greater than the population σ², you will come to the realization that it is only those samples with a few very high members and a few very low members—and there are not very many samples like that. By now you should have convinced yourself that $\sum(x-\bar{x})^2/n$ will result in a biased estimate of σ². You can see that, on average, it is too small.

How can an unbiased estimate of the population variance, σ², be found? If is $\sum(x-\bar{x})^2/n$ on average too small, we need to do something to make it a little bigger. We want to keep the $\sum(x-\bar{x})^2$, but if we divide it by something a little smaller, the result will be a little larger. Statisticians have found out that the following way to compute the sample variance results in an unbiased estimator of the population variance:

$$s^2=\frac{\sum(x-\bar{x})^2}{n-1}$$

If we took all of the possible samples of some size, n, from a population, and found the sample variance for each of those samples, using this formula, the mean of those sample variances would equal the population variance, σ².

Note that we use s² instead of σ², and n instead of N (really "nu", not "en") since this is for a sample and we want to use the Roman letters rather than the Greek letters, which are used for populations.

There is another way to see why you divide by n-1. We also have to address something called "degrees of freedom" before too long, and it is the degrees of freedom that is the key of the other explanation. As we go through this explanation, you should be able to see that the two explanations are related.

Imagine that you have a sample with 10 members (n=10), and you want to use it to estimate the variance of the population form which it was drawn. You write each of the 10 values on a separate scrap of paper. If you know the population mean, you could start by computing all 10 $(x-\mu)^2$. In the usual case, you do not know μ , however, and you must start by finding $\bar{x}$ from the values on the 10 scraps to use as an estimate of m. Once you have found $\bar{x}$, you could lose any one of the 10 scraps and still be able to find the value that was on the lost scrap from and the other 9 scraps. If you are going to use $\bar{x}$ in the formula for sample variance, only 9 (or n-1), of the x's are free to take on any value. Because only n-1 of the x's can vary freely, you should divide $\sum(x-\bar{x})^2$ by n-1, the number of (x's) that are really free. Once you use $\bar{x}$ in the formula for sample variance, you use up one "degree of freedom", leaving only n-1. Generally, whenever you use something you have previously computed from a sample within a formula, you use up a degree of freedom.

A little thought will link the two explanations. The first explanation is based on the idea that $\bar{x}$, the estimator of μ, varies with the sample. It is because $\bar{x}$ varies with the sample that a degree of freedom is used up in the second explanation.

The sample standard deviation is found simply by taking the square root of the sample variance:

$$s=\sqrt{[\sum(x-x)^2/(n-1)]}$$

While the sample variance is an unbiased estimator of population variance, the sample standard deviation is not an unbiased estimator of the population standard deviation—the square root of the average is not the same as the average of the square roots. This causes statisticians to use variance where it seems as though they are trying to get at standard deviation. In general, statisticians tend to use variance more than standard deviation. Be careful with formulas using sample variance and standard deviation in the following chapters. Make sure you are using the right one. Also note that many calculators will find standard deviation using both the population and sample formulas. Some use σ and s to show the difference between population and sample formulas, some use s_n and s_{n-1} to show the difference.

If Ann Howard wanted to infer what the population distribution of volleyball players' sock sizes looked like she could do so from her sample. If she is going to send volleyball coaches packages of socks for the players to try, she will want to have the packages contain an assortment of sizes that will allow each player to have a pair that fits. Ann wants to infer what the distribution of volleyball players sock sizes looks like. She wants to know the mean and variance of that distribution. Her data, again, is:

size	frequency
6	3
7	24
8	33
9	20
10	17

The mean sock size can be found:

=[(3x6)+(24x7)+(33x8)+(20x9)+(17x10)]/97 = 8.25.

To find the sample standard deviation, Ann decides to use Excel. She lists the sock sizes that were in the sample in column A, and the frequency of each of those sizes in column B. For column C, she has the computer findfor each of $\sum(x-\bar{x})^2$ the sock sizes, using the formula = (A1-8.25)^2 in the first row, and then copying it down to the other four rows. In D1, she multiplies C1, by the frequency using the formula =B1*C1, and copying it down into the other rows. Finally, she finds the sample standard deviation by adding up the five numbers in column D and dividing by n-1 = 96 using the Excel formula =sum(D1:D5)/96. The spreadsheet appears like this when she is done:

A	B	C	D	E
1	6	3	5.06	15.19

2	7	24	1.56	37.5
3	8	33	0.06	2.06
4	9	20	0.56	11.25
5	10	17	3.06	52.06
6	n=	97		Var = 1.217139
7				Std.dev = 1.103.24
8				

Ann now has an estimate of the variance of the sizes of socks worn by college volleyball players, 1.22. She has inferred that the population of college volleyball players' sock sizes has a mean of 8.25 and a variance of 1.22.

Summary

To describe a population you need to describe the picture or graph of its distribution. The two things that need to be described about the distribution are its location and its shape. Location is measured by an average, most often the arithmetic mean. The most important measure of shape is a measure of dispersion, roughly width, most often the variance or its square root the standard deviation.

Samples need to be described, too. If all we wanted to do with sample descriptions was describe the sample, we could use exactly the same measures for sample location and dispersion that are used for populations. We want to use the sample describers for dual purposes, however: (a) to describe the sample, and (b) to make inferences about the description of the population that sample came from. Because we want to use them to make inferences, we want our sample descriptions to be "unbiased estimators". Our desire to measure sample dispersion with an unbiased estimator of population dispersion means that the formula we use for computing sample variance is a little difference than the used for computing population variance.