Lesson 1. Observations

A basic activity of researchers is describing and analyzing observations. Medical experimenters observe patients' reactions to taking new medications. Political pollsters do surveys on the preferences of voters. Psychologists record the play behavior of children. In each of these cases, the researchers are concerned with describing and reaching general conclusions from observations.

In these lessons, we will consider only observations that can be represented by numbers. Numbers can be manipulated using the rules of mathematics and this is essential for statistical analysis. This is not as big a restriction as would at first seem to be the case. Many observations are taken directly as numerical values, for example, scores on psychological tests, like IQ measures, and scores on medical tests for blood pressure and cholesterol level. Other observations that do not start as numbers can be classified using numerical schemes; for example, children's aggression on the playground can be classified from low to high using numerical ratings made by trained observers. Numbers also can be used as labels to designate different categories, e.g., gender can be coded as 1 for men and 0 for women.

In this first lesson, we will learn about two different types of observations, systematic sequences and random sequences. Both types play a major role in statistical analysis.

Systematic Sequences

In a famous historical study, Galileo observed balls rolling down an inclined ramp. He started a ball rolling from the top of the ramp and measured how far the ball traveled down the ramp after different amounts of time had elapsed from the starting time. The intervals between his observations were all the same. His observations were similar to the ones below.

Time	Distance
1	1
2	4
3	9
4	16
5	25
6	36
7	49

Galileo discovered how distance was systematically related to the time when the observation was made. In each interval of time, the ball traveled farther than in the previous interval by a constant amount, two units. From time 1 to time 2, the ball's distance increased by 3 units; from time 2 to time 3, it increased by 5 units; from time 3 to time 4, it increased by 7 units. The increases, 3, 5, 7, 9, 11 and so on, follow a simple pattern.

In general, systematic sequences of observations, like Galileo's, can be described by giving the rule that determines the value of an observation from previous values in the sequence, e.g., in Galileo's results, add 2 to the previous increase, or by an equation that specifies how to calculate the value given the position of the value in the sequence.

The equation for Galileo's observations is

$$\text{Distance} = (\text{Time})^2$$

The distance is equal to the time squared. This equation lets you quickly calculate the distance for any time, e.g., the distance for a time of 12 is 12 times 12 or 144.

Galileo's observations could be represented visually on a graph, with each dot in the graph representing a pair of numbers - ?the distance and its corresponding time. The distance would be shown on the vertical axis and the time on the horizontal axis. The graph would show a systematic increase in distance as time increases.

Another well-known example of a systematic sequence is the Febonacchi sequence, shown below. Here each value is the sum of the two previous numbers; for example, the value 21 in the 8th position is the sum of 8 and 13, the values in positions 6 and 7. The Febonacchi sequence describes aspects of biological growth.

Position	Value
1	1
2	1
3	2
4	3
5	5
6	8
7	13

Random Sequences

In a systematic sequence, the values in the sequence change systematically, according to a rule. In a random sequence, there is no such regularity. Consider a simple experiment that generates a random sequence. Take two dice and roll them and then add up the numbers showing on each die. Each die has the numbers 1 through 6 on its sides, so the sum has possible values of 2 through 12. The sum you get when you roll the dice is the first number in the sequence. Do this again and again until you have 100 observations. The observations below are in order from the 1st roll to 100th roll.

6, 5, 10, 9, 5, 7, 3, 12, 6, 8, 10, 6, 9, 10, 6, 7, 10, 6, 9, 9, 7, 8, 8, 12, 5, 5, 7,
5, 10, 6, 3, 6, 6, 6, 9, 10, 8, 5, 5, 4, 3, 11, 7, 5, 8, 3, 2, 6, 9, 10, 5, 8, 12, 7,
5, 3, 10, 3, 6, 9, 7, 7, 9, 6, 9, 8, 6, 10, 4, 7, 8, 3, 5, 9, 6, 5, 7, 7, 10, 9, 7, 11,
9, 8, 6, 9, 10, 5, 7, 6, 4, 8, 12, 5, 8, 8, 4, 9, 7,12

Don't spend time trying to figure out a systematic pattern for this sequence. No one can figure out what the next value in the sequence will be given the previous values. Also, no one can write an equation that gives the sum for each position. The graph of the sequence would show no regular form for the relationship between the sum of the dice and the position in the sequence.

A random sequence has two properties:

1) the value associated with each position is within a range of possible values, and

2) it is impossible to write a rule for calculating the next value in the sequence from knowing the previous values or to write an equation for calculating the value from the position.

Games of chance, like tossing coins, rolling dice, and playing roulette, generate random sequences, but, as we will see in the next lessons, so do experiments and surveys.

Lesson 2. The Histogram

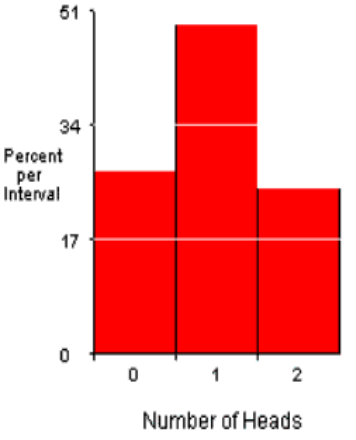
Systematic sequences can be described by given the generating rule for the sequence or the equation relating value and position, but these descriptions won't work for a random sequence. Random sequences can't be described as accurately as systematic sequences since in a random sequence the position gives no information about its corresponding value; i.e., there is no regularity in the way the values change throughout the sequence. Random sequences can be described by giving the range of possible values and the relative frequency of each value in the sequence. This information, the relative frequency of occurrence for all possible values, is called the distribution of the sequence.

Random sequences are described as well as possible by their distribution.

For example, consider the sequence of values that would be generated from throwing two pennies and counting the number of heads. On each toss, the possibilities are 0, 1 and 2. The numbers below show the results of a sequence of 250 tosses of the two coins. 67 tosses resulted in 0 heads, 122 tosses resulted in 1 head, and 61 tosses gave 2 heads. This is as complete a description of the sequence as it is possible to make.

2, 1, 2, 1, 1, 0, 2, 1, 2, 2, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1, 2, 1, 1, 0, 2, 0, 0, 1, 0, 1, 1, 2, 1, 0, 0, 2, 2, 1, 0, 2, 0, 2, 0, 0, 0, 2, 0, 2, 1, 1, 2, 1, 1, 2, 2, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 2, 2, 0, 0, 1, 0, 2, 2, 0, 1, 0, 1, 0, 0, 1, 1, 2, 1, 2, 1, 0, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 2, 0, 1, 1, 2, 1, 1, 2, 2, 2, 1, 0, 1, 1, 0, 0, 2, 1, 0, 1, 2, 0, 1, 0, 1, 1, 0, 1, 2, 0, 2, 2, 1, 1, 1, 1, 1, 2, 0, 1, 1, 0, 1, 0, 1, 0, 2, 2, 2, 0, 1, 1, 0, 2, 2, 0, 2, 0, 1, 0, 1, 1, 1, 2, 2, 1, 1, 0, 0, 2, 1, 1, 1, 2, 0, 1, 1, 1, 1, 1, 1, 0, 1, 2, 0, 2, 1, 1, 1, 0, 0, 2, 1, 1, 1, 1, 1, 2, 0, 1, 1, 1, 1, 1, 1, 2, 0, 0, 0, 2, 1, 1, 1, 0, 1, 1, 1, 1, 2, 1, 0, 2, 1, 2, 2, 1, 1, 0, 0, 2, 0, 1, 2, 1, 2, 0, 1, 1, 1, 0, 0, 1, 2, 1, 2, 1, 0, 0, 2, 0, 1

The distribution of the sequence can be shown in a graph called a histogram. A histogram is a graph showing the distribution of a random sequence. The graph shows the relative frequency for all possible values in the sequence. The histogram for the 250 tosses of two coins is shown below. In the graph the relative frequency is expressed as a percent.



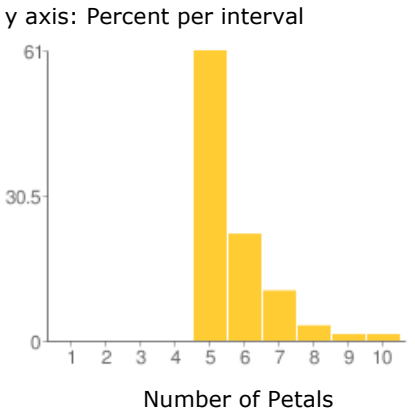
¶

The horizontal axis shows the range of values in the sequence; an interval

on the axis is associated with each value. In this sequence, there are 3 possible values, getting 0, 1 or 2 heads; so the horizontal axis is divided into 3 intervals. The vertical axis is scaled according to the percent of observations per interval. A bar is drawn in each of the intervals on the horizontal axis to a height that corresponds to the percent of the observations in the sequence that fall within the interval. Accordingly, the bar for the 0 interval is drawn to a height of (67/250) times 100 = 26.8%, the bar for interval 1 is drawn to a height of 48.8%, and the final bar, for interval 2, has a height of 24.4%. Histograms have the power of revealing characteristics of random sequences that are not apparent to the naked eye. The numbers below were recorded by a botanist walking through a field of buttercups. Each value is the number of petals on a flower.

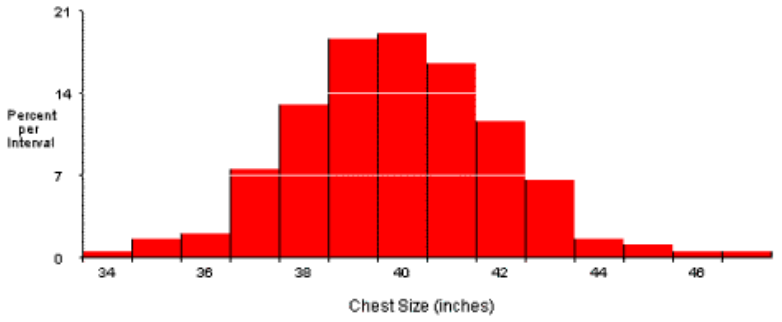
5, 5, 5, 7, 5, 6, 6, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 5, 5, 5, 5, 5, 5, 5, 5, 5, 7, 5, 7, 7, 6, 5, 5, 5, 5, 5, 5, 5, 6, 7, 5, 5, 6, 5, 5, 7, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 5, 5, 5, 5, 6, 5, 5, 5, 5, 5, 5, 6, 5, 6, 5, 7, 5, 5, 6, 6, 5, 5, 7, 5, 5, 5, 5, 5, 6, 5, 5, 6, 6, 5, 5, 5, 5, 5, 6, 5, 8, 5, 5, 5, 7, 6, 5, 6, 8, 5, 5, 6, 6, 6, 9, 6, 5, 6, 6, 5, 8, 5, 5, 6, 6, 7, 6, 10, 5, 6, 5, 6, 5, 6, 5, 7, 5, 5, 5, 5, 5, 5, 7, 5, 7, 5, 5, 7, 5, 10, 5, 6, 5, 6, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 5, 6, 5, 5, 5, 7, 5, 6, 7, 6, 5, 5, 8, 6, 8, 5, 6, 5, 5, 6, 7, 5, 6, 6, 5, 5, 8, 6, 5, 5, 6, 7, 5, 5, 5, 5, 5, 7, 5, 5, 7, 7, 5, 5, 6, 9, 9, 5, 5, 8, 5, 6, 5, 6, 7, 6, 6, 5, 5, 5, 5, 5, 7, 5, 5, 10

The histogram for this set of observations has a distinctive shape.



The most common number of petals is five, followed in decreasing popularity by six, seven, eight, nine, and ten petals per flower. The shape of the distribution is a property of buttercups as much as the color of the flower or its scent.

Perhaps the most famous distribution is the bell curve. The histogram below shows the distribution of chest sizes for about 6000 men. This distribution has a shape like the cross section of a bell. This histogram, drawn by Quetelet in the 19th century to illustrate the circumference of soldiers' chests, marked the first time that a characteristic of people was shown to have a bell curve or a normal distribution. The publication of this remarkable finding encouraged statistical analyses in the social sciences and led to the development of psychological measurement.


21
14
7
0
Percent per Interval
34
36
38
40
42
44
46
Chest Size (inches)

Lesson 3. Drawing the Histogram

The numbers below are the scores on the verbal SAT exam for 50 college students.

530, 490, 540, 550, 490, 460, 650, 480, 470, 370, 440, 560, 480, 430, 340, 480, 530, 540, 390, 550, 440, 420, 540, 550, 490, 480, 490, 480, 360, 490, 610, 420, 540, 410, 450, 360, 370, 320, 410, 500, 430, 360, 550, 450, 420, 480, 410, 530, 510, 480

The scores range from 320 to 650. To construct a histogram showing the distribution of the scores, we need to divide the range of scores into a number of intervals and then determine the percentage of scores in each interval. For each of the three previous histograms we have looked at, the size of the interval was one - one head, one petal, and one inch in chest size. For the SAT data, however, the range of scores is too great to have an interval for every possible value. To cover the range of scores from 320 to 650 would require 650-320 = 330 intervals of size one. Typically, histograms only have from 10 to 20 intervals to cover the whole range of scores. To cover the range of 330 points with, say, 15 intervals would require 330/15 = 22 points per interval. We can round this to 20 points per interval to construct the SAT histogram.

In Quetelet's chest size histogram, each interval was one inch wide and the intervals were labeled by giving their mid-point, the chest size falling in the center of the interval. For example, the interval on the histogram labeled 40 inches included all chest sizes from 39.5 inches to 40.5 inches, an interval of one inch in size.

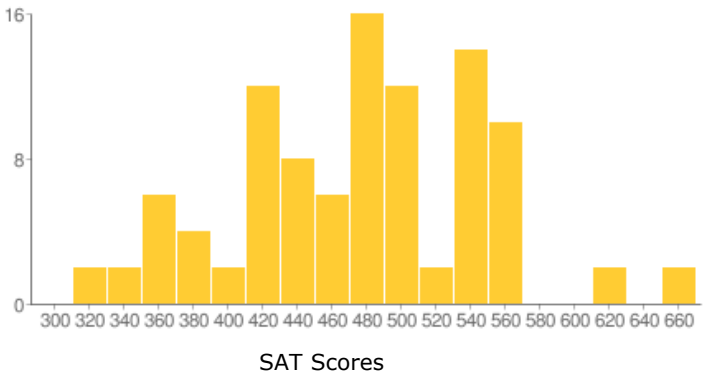
For the SAT scores, we have decided on an interval size of 20 points; so the interval labeled with a mid-point of 400 points will include scores from 390 to 410. Similarly, the interval labeled 420 will include scores from 410 to 430, and so on. A score falling on the upper limit of an interval, say, the score 410, is, by convention, placed in the next higher interval. The table below shows the result of assigning the 50 SAT scores to the different intervals. The table starts with an interval with the mid-point of 300. This was chosen as a convenient whole number less than the lowest score in the set, 320.

Midpoint	Frequency	Percent
300	0	0
320	1	2
340	1	2
360	3	6
380	2	4
400	1	2
420	6	12
440	4	8
460	3	6
480	8	16
500	6	12
520	1	2
540	7	14
560	5	10

580	0	0
600	0	0
620	1	2
640	0	0
660	1	2

The numbers in the Frequency column are calculated by counting the frequency of scores in each interval. The percent of scores is equal to the frequency divided by the total number, times 100. For example, there are 3 scores in the interval with mid-point 460, so the percent in this interval is (3/50) times 100 = (.06) times 100 = 6 percent. The histogram is drawn using the information in the table.

y axis: Percent per interval



The calculations for the histogram are easy, but tedious, and prone to error. It is wise to do them on a computer. The program SPSS does histograms as well as Excel.

The website http://muststudy.com/LearnStat/descriptiveStatHist.html calculates descriptive statistics and draws histograms.

Using the computer to draw histograms has a big advantage over doing it by hand. You can experiment with different interval sizes and starting mid-points to get just the right look for the graph.

Deciding on the starting mid-point is straightforward. Start at a value equal to or less than the minimum value in the set of numbers. It is a good idea to start at some logical or "natural" starting point like 0. Set the mid-point of the first interval equal to this natural starting point.

Deciding on the best interval size is analogous to focusing a microscope. With a microscope, you turn a wheel back and forth until the slide comes into focus. With a histogram, you adjust the interval size higher and lower until the histogram shows the "best" pattern for the data. Below are guidelines, rules of thumb, to help in constructing a good histogram.

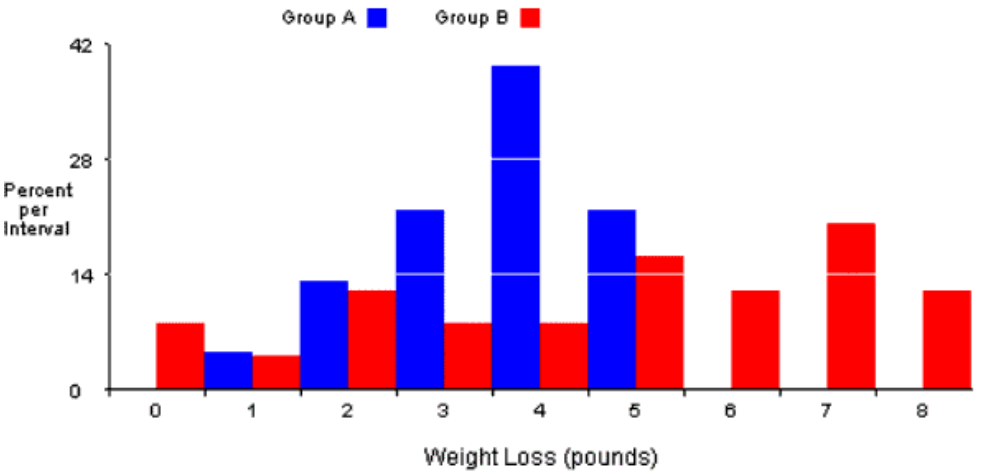
- Consider a starting point equal to or less than the minimum value in the data set.
- If there is a logical minimum value, consider using this as the

starting point. With money, for example, 0 may be a good starting point.

- Pick a "natural sized" interval, one that is easily understood by the reader. For example, 1, 2, 3, 5, 10, 20, 25, 50, and 100 are commonly used interval sizes. Avoid weird interval sizes, like 11 or 3.5, unless there is a good reason.
- For histograms where the data are small whole numbers like 3 and 4, as in the buttercup observations in the last problem set, pick an interval size of 1.
- Don't select too large or too small an interval size. The buttercup histogram used intervals of size one. With an interval size of five, the shape of the distribution of petals would have been lost. Select the number of intervals by trial and error, using the computer. Try different interval sizes to see if a definite pattern or shape emerges from the histogram.

Lesson 4. The Mean and Range

The histogram gives a concise description of a set of observations and also can reveal characteristics of the observations not apparent from inspecting the values themselves. For these reasons, the histogram should be used routinely in data analysis. However, for some purposes, an even more concise summary of the observations is needed. For example, consider an experiment comparing the effectiveness of two diets, Diet A and Diet B. The distribution of weight loss in pounds is shown below for the subjects who were on each of the diets for six weeks. This graph is a variant of the histogram used for comparing two distributions. Each interval has two bars showing the percent of subjects having weight losses in the interval for both diets.



The graph suggests that Diet B may be more effective than Diet A, but it is hard to assess the degree of advantage of Diet B. Some subjects on Diet B lost the most weight, but also some subjects on Diet B lost no weight. To simplify the comparison between the weight losses of the diets, statisticians summarize the weight loss of Diet A with a single number, the average, and compare this value with the average weight loss for Diet B. For the results shown above, the average for Diet A was 3.6 pounds and the average for Diet B was 4.7 pounds; so Diet B has a 1.1 pound advantage in weight loss over Diet A. Summarizing the results for each diet with a single number makes the comparison between diets clear and objective.

In this lesson, we will begin to learn about the most popular summary statistics, including the average (also called the **mean**), range, standard deviation, median, and percentiles. In an actual experiment on diet, researchers overfed 16 volunteers by 1000 calories a day for 8 weeks (Levine, Eberhardt, & Jensen, 1999). The participants, who agreed not to burn off the extra calories by exercising, were weighed-in at the start of the study and again after eating the 56,000 extra calories. Each person's weight gain was calculated resulting in a set of 16 numbers. The sum of the weight gains for all 16 people was 166.6 lb. If everyone had gained the same amount, each person would have gained 166.6 lb. / 16 = 10.4 lb. This is the average or **mean** weight gain for the group.

The **mean** of a set of numbers is equal to the sum of the numbers divided by the number of values.

Most likely, you first learned the importance of the mean as a child concerned with sharing things among your friends. The mean is the formula for fair sharing. If there are 12 pieces of candy to share among Sue, Tim, and Joan, each child should get the mean: 12 pieces divided by 3 kids = 4 pieces per kid. The sum of the number of pieces of candy is 4 + 4 + 4 = 12. Sue gets 4 pieces, Tim gets 4 pieces, and Joan gets 4 pieces. Each gets exactly 4, and, with justice for all, there is no variability among the numbers.

The mean always shares the sum equally among the people sharing. Sometimes, though, it is hard to do with 5 kids and 12 pieces of candy, each kid would get 12/5 = 2.4 pieces. How do you divide the candy to get that $4/10^{th}$ of a piece for each child? One answer would be to increase the number sharing -- count yourself among the kids: 12/6 = 2.

In the weight gain experiment, the total gain of 166.6 lb. was not shared equally. Everyone did not gain the same amount. The largest gain was 15.6 lb., while the smallest was only 3.1 lb. Here we say that the **range** was 15.6 lb. - 3.1 lb. = 12.5 lb.

The **range** of a set of numbers is equal to the difference between the highest and the lowest values. The range is a measure of variability.

The range of 12.5 lb. means that in this experiment people who ate the same amount of food did not put on the same number of pounds. We will see what accounted for this later in the lessons.

When you are dividing candy among kids, you want the range of the number of pieces given to the children to equal zero. This only happens when every number is equal to the mean.

Lesson 5. The Standard Deviation

In the overeating study, the mean weight gain for participants was 10.4 lb., with a range of 12.5 lb. The fact that the range was not equal to zero indicates that there is variability in weight gain despite the fact that everyone overate to the same degree. Variability is an important feature of human behavior and describing variability is an important part of statistical analyses. The range is an easy-to-compute measure of variability - just subtract the smallest number from the largest number. However, statisticians have developed a much better measure, the standard deviation, SD. The SD is harder to compute than the range but this difficulty is justified by its usefulness.

To keep the numbers simple, let's pretend the overeating study was done with only 4 subjects whose weight gains were:

S1 7 lb. S2 8 lb. S3 12 lb. S4 13 lb.

The sum of the gains is 40 lb. and the mean is 40 lb. / 4 people = 10 lb. per person.

If the subjects had all gained the mean amount (10 lb.), together they would have gained 40 lb., and there would be no variability in their weight gains. But this did not happen. Subject 1 gained 7 lb., 3 lb. less than the mean. Subject 2 gained 8 lb., 2 lb. less than the mean. Subject 3 gained 2 lb. more than the mean and S4 gained 3 lb. more than the mean.

These weight gains can be written as the mean plus or minus a deviation:

Gain = Mean + or - deviation

S1	7 lb. =	10	-	3
S2	8 lb. =	10	-	2
S3	12 lb. =	10	+	2
S4	13 lb. =	10	+	3

The deviations here are 2 lb. or 3 lb. above or below the mean. To get one number to represent the deviations for all the subjects considered together, statisticians use a complicated average called the **root mean square** , or **r.m.s.**

To compute the r.m.s. of a set of numbers, follow these steps:

1. square each deviation (the square part of r.m.**s**)
2. add up the squared values
3. divide by the number of values (the mean part of r.**m.** s)
4. take the square root (the root part of **r** .m.s)

To compute the r.m.s. of the deviations from the mean in the example, the numbers 3, 2, 2, 3:

1. square each deviation: 9, 4, 4, 9
2. add them up: 9+4+4+9 = 26
3. divide by 4,(the number of values) 26/4 = 6.5
4. take the square root of 6.5: square root = 2.55 lb.

This value, 2.55 lb., is the r.m.s. of the deviations from the mean. Another more widely used name for this r.m.s. is the **standard deviation, SD** .

The standard deviation, SD, of a set of numbers is the r.m.s. value of the deviations from the mean.

The standard deviation can be interpreted as a typical deviation from the mean. In this example, the mean is 10 lbs. and the SD is 2.55 lbs. Think of the typical score being 2.55 lbs. from the mean of 10 lbs.

The standard deviation is tedious to calculate, especially if there are many numbers in the set. Using the computer for this calculation is recommended.

The website http://muststudy.com/LearnStat/descriptiveStatHist.html calculates the standard deviation and other descriptive statistics.

Lesson 6. Changing Units

The 16 participants in the overeating experiment had a mean weight gain of 10.4. lb., with a SD of 3.7 lb. These values were calculated in the last lesson. These numbers describe the weight gain using the English system of weights which uses units like the lb., oz., pint, etc. Scientific journals use the metric system, so to publish these findings we would need to convert from lb. to kilograms (kg.). One kg. is equal to 2.2 lb. A weight gain of 10 lb. is the same as a gain of 10/2.2 = 4.55 kg.

To compute the mean and SD in kg., we could transform each subject's gain into kg. and then compute the mean and SD of these new numbers. But there is an easier way. The mean in kg. is just the mean in lb. divided by 2.2; also, the SD in kg. is the SD in lb. divided by 2.2. So the mean gain was 10.4/2.2 = 4.73 kg. and the SD was 3.7/2.2 = 1.68 kg.

Whenever numbers are changed by multiplying or dividing by a positive number, the mean and SD change by the same factor. For example, if all the numbers are multiplied by 3, the mean and SD also are multiplied by 3.

Multiplying or dividing by a negative value is slightly different.

When numbers are multiplied or divided by a negative value, the mean changes by this factor but the SD changes by the absolute value of the factor. For example, if the mean and SD of a set of numbers are 10 and 5, respectively, and each number in the set is multiplied by 3, then the new mean is 30 and the new SD is 15. The SD cannot be negative.

The rule for adding and subtracting numbers is as follows:

Changing each number in a set by adding or subtracting a value changes the mean but not the SD. For example, subtracting 3 from each number, lowers the mean by 3 but does not change the SD.

Let's practice with a simple example: The numbers 3, 4, 5, and 6 have a mean of 4.5 and a SD of 1.12. First, add 7 to each number giving 10, 11, 12, and 13. What is the new mean and SD? The SD will not change since we just added a number. The mean will increase by 7. So the new mean and SD are 11.5 and 1.12, respectively. Now multiple each number by 2 giving 20, 22, 24, and 26. This changes the mean and the SD by the factor of 2. The mean and SD are 23.0 and 2.24, twice the old values.

Now, for a more realistic example, let's change Fahrenheit temperatures to Celsius temperatures. The high temperature for 5 US cities one summer day is 92, 74, 92, 98, and 101. The mean is 91.4 and the SD is 9.37. To change from degrees Fahrenheit to degrees Celsius subtract 32 from the Fahrenheit temperature and then multiple by 5/9. So the new mean in degrees Celsius is (91.4 - 32)*5/9 = 33.0 The new SD is just the old SD times 5/9. SD Celsius equals 9.37*5/9 = 5.21.

Lesson 7. The Mean and SD as Descriptive Statistics

T he mean and the SD work together to provide a good two-number summary of a set of numbers. To see how they work together, let's look at the weight gains in the overeating experiment. The gains, in order of magnitude, are:

3.1, 5.5, 5.7, 7, 7.9, 8.6, 8.6, 10.6, 11.7, 12.8, 13.2, 13.2, 14.1, 14.3, 14.7, 15.6

The mean is exactly 10.4125 lb. The deviations from the mean are:

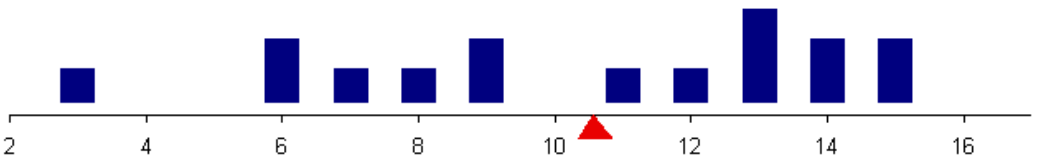
-7.3125, -4.9125, -4.7125, -3.4125, -2.5125, -1.8125, -1.8125 , .1875, 1.2875, 2.3875, 2.7875, 2.7875, 3.6875, 3.8875, 4.2875, 5.1875

The negative deviations, shown in red type, are for gains below the mean; the positive deviations are for gains above the mean. There are 7 negative deviations; their sum is -26.4875. There are 9 positive deviations; their sum is +26.4875. The fact that these sums are equal in value is not a coincidence.

The sum of the deviations below the mean is always equal in value to the sum of the deviations above the mean.

The mean is a balance point or center of gravity for a set of numbers.

Figure 1 shows the weight gains rounded to whole numbers and placed on a imaginary weightless beam at the position in accord with their values. Each value is shown as a box in the figure; there are a total of 16 boxes, one for each person. The beam will balance, like a teeter-totter, at a point equal to the mean, shown by the triangle in the figure. The mean is the only value where this balance occurs.



The mean is located in the center of the numbers at the value where the deviations of all the numbers below the mean add up to equal the sum of these deviations above the mean. The mean is as close to the smaller numbers as to the larger numbers, making the mean an excellent one-number representation of a whole set of numbers.

The mean describes the center of the numbers and the SD measures the variability from the mean. For the weight gains, the SD is 3.72. Now compare this standard deviation with the actual deviations (shown below) and count the number of deviations that are less than or equal to 3.72 in absolute value (forget about the signs). These deviations are shown in green type.

-7.3125, -4.9125, -4.7125, -3.4125, -2.5125, -1.8125, -1.8125 , .1875, 1.2875, 2.3875, 2.7875, 2.7875, 3.6875 , 3.8875, 4.2875, 5.1875

There are 10 deviations less than the standard deviation. This is 10 out of

16 or 62.5%. Now count the number of deviations that are less than two times the SD, or 2*3.71 = 7.42. All the actual deviations, 100%, are less than 2*SDs.

These two percents, 62.5% and 100%, are common values for the percent of numbers in a set that are within plus or minus 1SD (62.5%)and within plus or minus 2SDs (100%) of the mean. Values close to these are found in many, many sets of numbers. Consider the following five very different sets. Pay attention to the percent of the numbers within 1SD (shown in green type) and within 2SDs of the mean, and also within 3 SDs of the mean.

Set 1: The initial weight in lb. for the 16 participants in the weight gain experiment:

117.3 120.8 122.5 123 127.2 131.1 135.7 137.1 139.5 149.8 155.5 161.3 162.1 163.2 166.3 201.7

mean = 144.63, SD = 21.97

Percent of numbers with deviations from the mean within:
+/- 1 SD = 75 %
+/- 2 SDs = 93.8 %.
+/- 3 SDs = 100 %.

Set 2. The weights in lb. of a sample of 20 cars sold in the US:

2885, 2895, 2960, 2980, 3235, 3240, 3345, 3355, 3445, 3470, 3785, 3985, 3995, 4050, 4050, 4440 , 4605, 4710, 4875, 5335

n = 20, mean = 3782, SD = 703.41

Percent of numbers with deviations from the mean within:
+/- 1 SD = 60 %
+/- 2 SDs = 95 %.
+/- 3 SDs = 100 %.

Set 3. The heights in inches of 49 students in a college class:

56 58 59 60 60 60 61 61 62 62 62 62 62 63 63 63 63 63 63 63 63 64 64 64 65 65 65 65 65 65 66 66 66 66 66 66 67 67 67 67 68 68 68 68 68 68 69 69 70

n = 49, mean = 64.31, SD = 3.07

Percent of numbers with deviations from the mean within:
+/- 1 SD = 65.3 %
+/- 2 SDs = 95.9 %.
+/- 3 SDs = 100 %.

Set 4. The high temperatures (Fahrenheit) of 21 cities in the US on a hot summer day.

74 85 86 87 88 88 88 89 90 92 92 92 92 93 94 94 96 98 98 101 107

n = 21, mean = 91.62, SD = 6.53

Percent of numbers with deviations from the mean within:
+/- 1 SD = 81 %
+/- 2 SDs = 90.5 %
+/- 3 SDs = 100 %

Set 5. 100 pennies were placed in a jar, shaken well and dumped on the floor, the number of heads was recorded. In 20 repetitions, the results were (in rank order):

37 44 46 46 46 46 47 48 48 49 49 49 50 51 51 53 56 56 57 59

n = 20, mean = 49.4, SD = 4.97

Percent of numbers with deviations from the mean within:
+/- 1 SD = 70 %
+/- 2 SDs = 95 %.
+/- 3 SDs = 100 %.

In these examples, the percent of numbers within plus or minus 1SD of the mean varied from 60% to 81%. The percent within plus or minus 2SDs varied from 93.5% to 95.9%. None of the numbers were over 3SD away from the mean.

For many sets of numbers, about 50% to 80% (about **68%** is typical) of the numbers in a set are **plus or minus 1SD from the mean** and about 90% to 100% (**95%** is typical) of the numbers are **plus or minus 2SDs from the mean.** Less than **1/3 of 1%** are **3 or more SD from the mean** .

Because these percents apply to so many data sets, these percents are used to interpret the mean and SD. For example, the mean height for men in the US is about 69 inches with a SD of 3 inches. From this description, we know that about 68% of men have heights in the range from 66 inches to 72 inches (69 +/- 3). Most mens' heights are in the range from 63 inches to 75 inches (from 5 foot 3 to 6 foot 3); and less than 1% of men have heights less than 63 inches or more than 75 inches. The mean and SD work together to give a comprehensive description of this set of heights.

Lesson 8. The Median and Other Percentiles

Not all sets of numbers are described well by the mean and SD. For example, the seven numbers: 1, 2, 3, 4, 5, 6, 50. This sequence has a number, 50, that is much larger than the other numbers. This extreme value causes the mean, 10.1, to be higher than all but one score in the set. Also, there is a higher than expected percent of scores within +/- 1 SD of the mean, 85% vs. the typical 68%, and fewer scores than expected within +/- 2 SDs of the mean, 85%, vs. the expected 95%. When the mean and SD are not good descriptive statistics, as is the case for these numbers, statisticians recommend the **five number summary.**

The **five number summary** uses the low score, the high score, and the 25th, 50th, and 75th percentiles to describe a set of scores.

The **25th percentile** is the value that divides a set of numbers so that 25% of the scores are less than this value and 75% of the scores greater than the value.

The **50th percentile**, also called the **median** , is the value that divides the number set 50-50, i.e., 50% of the scores are less than the median and 50% of the scores are greater than the median.

The **75th percentile** divides the scores 75-25; 75% of the scores are less than the 75th percentile and 25% are greater than the 75th percentile.

These percentiles, the 25th, 50th, and 75th, are called **quartiles** since they divide the number set into four quarters. One quarter of the scores are less than the 25 th percentile, one quarter are between the 25th and 50th percentiles, one quarter are between the 50th percentile and the 75th percentile, and the final quarter is greater than the 75th percentile.

These three percentiles are easy to calculate after the scores in the set are placed in rank order. Let's say there are n scores in the set. First, calculate (n+1)/2. This is the rank order of the median in the number set. For example, for the set of 7 numbers: 1, 4, 7, 10, 13, 16, 19; (n+1)/2 is (7+1)/2 = 4. So the fourth score in rank order is the median. 1 is the first score, 4 is the second score, 7 is the third score, and the median is the fourth score, which is equal to 10.

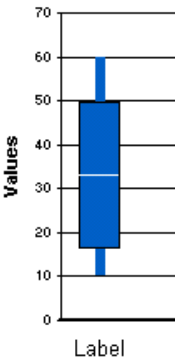
If there is an even number of scores in the set, e.g. 8 or 12, then the expression (n+1)/2 will not be a whole number. If n = 8, then (n+1)/2 = 4.5. Here the median is the number midway between the 4th and the 5th scores in the set when they are in rank order. For example, for the set 1, 4, 7, 10, 13, 16, 19, 21, there are 8 scores; so (n+1)/2 = 4.5, and the median is midway between 10 and 13. The median is 11.5, the average of 10 and 13.

The 25th percentile is computed using the same procedure as the median, but for just the lower half of the numbers. Take all the numbers less than the median (do not include the median), then compute the median for these numbers; this is equal to the 25 th percentile for the whole set of numbers.

To calculate the 75th percentile, take all the numbers in the set greater than the median (again, do not include the median), then compute the median of these scores. This gives the 75th percentile for the whole set. For example, for the 12 numbers 12, 14, 17, 19, 24, 33, 35, 41, 49, 51, 55, 60, (n+1)/2 = 6.5, so the median is midway between 33 and 35, which is 34. There are 6 scores less than

the median, so (n+1)/2 now is 7/2 = 3.5, and the 25 th percentile is midway between 17 and 19, that is 18. There are 6 scores above the median, so (n+1)/2 is again 7/2 = 3.5. The 75th percentile is midway between 49 and 51, which is 50.

The five number summary can be presented in a graph called a **box and whisker plot**. The plot for the number set above is shown below. The box is drawn between the 25th and 75th percentiles, a line is drawn in the box indicating the median, and the "whiskers" are drawn from the box to the lowest and highest scores in the set. The graph is useful for comparing different set of numbers.



Lesson 9. Comparing the Mean and the Median

Both the mean and the median are popular statistics to describe the "central tendency" or mid-point of a set of numbers. For many sets, the two statistics are very close in value. However, the statistics do have different properties and can have different values when computed from the same set of numbers.

The properties of the median are:

- 50% of the scores are less than the median; 50% are greater than the median.
- The median is not sensitive to extreme scores. For example, for the three numbers 10, 15, 20, the median is 15. The median of 10, 15, 1000, is also 15. The third number in the set increased from 20 to 1000 without changing the median.
- The median is the value that is closest to all the scores in the set. That is, if you compute the deviation of each score from the median (consider each deviation a positive value) and add up the deviations, the sum of the deviations from the median is smaller than from any other possible value.

The properties of the mean are:

- The mean, or average, shares the sum equally among the members of the set. So the sum can be computed from the mean by multiplying by the number of values.
- The sum of the deviations less than the mean is equal to the sum of the deviations greater than the mean. So the mean is the "balance point," or center of gravity, of the set of scores. The mean is as close to the small numbers as it is to the large numbers.
- The mean is the value that minimizes the square of the deviations. That is, compute the deviations of each score from the mean, square them, and add them up. This sum has a minimum value when the deviations are computed from the mean rather than any other possible value.
- The mean is sensitive to extreme scores.

In practice, the choice of whether to use the median or the mean as a descriptive statistic depends upon your purpose in describing the set of numbers. For example, say a day trader's daily profits from the stock market for the last week were $20, $-50, $35, $25, and $10,000. The median here is $25 and the mean is $2006. Which is a better descriptive statistic? In describing profits what is important is the bottom line, the sum, and the mean divides the sum equally among the five days in the week. Also, if we know the trader's daily profit is $2006, we can easily compute the trader's weekly total and project the total for any period of time. The median, $25, is completely insensitive to the high profit on the last day of the week; the total cannot be computed from the median.

The lesson of this example is that when the total of the number set is the concern, then the mean is the appropriate statistic. As examples, use the mean to describe the pay of a waiter per day, the number of calories eaten per day in a diet experiment, the amount of crop grown (bushels) in an agricultural experiment, and the rainfall per day during a month in the summer.

The median is the better statistic if you want to avoid the impact of extreme scores. If you wished to describe the day trader's profits for a "typical" day, the median would be better than the mean. As another example, to describe the cost of houses in a particular neighborhood to a prospective buyer, the median would be better than the mean. When the buyer knows the median price, they know that 50% of the houses cost less, and 50% cost more than this price. The mean may be inflated by a few very expensive homes. The tax collector, however, would be more interested in the mean price of the houses, since his calculation of taxes is based on the total value of the real estate in the neighborhood.

You are not forced to choose between the median and the mean; both statistics can be used to describe a set of numbers. For some number sets, like the day trader's profits, knowing both the mean and median gives you a better understanding of the set of numbers than just knowing either the mean or median by itself. In fact, there is a standard terminology for describing number sets depending on the relationship between the mean and the median. If the mean is markedly greater than the median, the set of scores is said to be **skewed to the right** (toward higher values), and if the mean is markedly less than the median the set is **skewed to the left** (toward lower values). If the mean and median agree in value, the set is **not skewed**.

Lesson 10. Percents

The mean and standard deviation appear frequently in research reports published in scientific journals to describe the outcome of experiments. However, in the popular media (newspapers, TV, radio, and magazines) a different statistic often is used to describe the results of the same study. This is because the mean and standard deviation are difficult to understand without a course in statistics. However, the statistic used in the popular media is straightforward and can be understood without special training. Here is a quote illustrating the use of the statistic from the science section of the New York Times, July 27, 1999:

> Psychiatrists estimate that 60 to 70 percent of people who can tolerate the side effects of antidepressants get better with the first drug they take. But 10 percent do not respond, even after trials on many different drugs.

The statistic is the **percent** . With a single number, the percent can describe both the central tendency or typical response in a set of results and indicate the variability. Using the mean and SD requires two numbers and the 5 number summary requires, of course, 5 numbers. The percent is a sensational statistic. If you could take only one statistic to use on a desert island, this would be the one to take.

In the above quote, we learn that about 70% of depressed patients improve with the first drug they try. From this, it follows that 30% of patients do not improve. The single percent, 70, describes the number who are cured and also informs us that there is variability in the response, since 30% are not cured.

> The percent is easy to compute: Count up the number of cases of interest, then divide by the total number of cases and multiply by 100. For example, if in a group of 40 patients who received an experimental drug, 30 improved, then 30/40 = .75 or 75% improved. The proportion .75 could be used, but multiplying by 100 gets rid of the decimals, which are a nuisance.

The percent is by far the most popular statistic used in media accounts of research. Read the science section of the newspaper and look for references to the percent, mean, median, and standard deviation. Percents are everywhere. We see the mean and median less often, and we have seen the standard deviation mentioned only once in years of reading newspapers.

Statisticians use a special method of coding to compute percents on a computer. Most computer programs require entering results individually for each subject in a research study or survey. For example, age in years would be entered as a three-digit number, 027 or 104. Cholesterol level also would be a three digit number, say 160. But how would you enter whether the patient was improved or not improved so the percent improved could be calculated? The answer is to code "improved" with the number 1 and "not improved" with the number 0. Then calculate the mean of this coded variable and multiply by 100. The sum for this coded variable is equal to the number of improved patients, because each subject who improved is coded 1. The mean, then, is the proportion of patients who improved. Finally, multiply by 100 to get the percent, the number who improved per 100 patients.

This coding, which treats the percent as a special mean, is useful for computer calculations as well as for probability calculations.

Percents are displayed in a graph called a **pie chart**. The **pie chart** for the results quoted above on the treatment of depression is shown below. The circle is divided into regions so that the percent of the area of the circle corresponds to the percent of cases in the research in the different categories. Twenty-five percent is indicated by 1/4 of the area of the circle or pie, 50% by half of the circle, and so on.