

Chapters 1

What is Statistics?

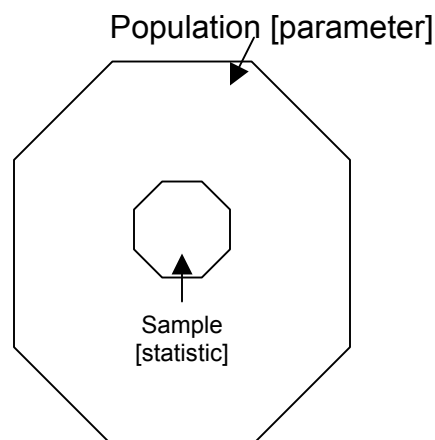
The field of statistics focuses on determining the characteristics of the things around us. It is a field of mathematics, in that its characteristics are expressed in quantitative (numerical) terms.

We are mostly interested in determining characteristics of specific groups. For example, Midlands Tech students, retired people, voters in U.S. elections, and people taking a specific medicine would be such groups. The set of **ALL** individuals or items in such a group is known as the **population** about which we want information. The characteristic about which we want to gather information from the population is known as the **parameter** of the population. For example, if the population is defined as all the people taking the drug Azane, we may be interested in what proportion of Azane users develop the side-effect of high blood pressure. This proportion would be known as a parameter of the population of Azane users. There may be other parameters as well, if there is more than one characteristic that we want to determine.

However, it may not be possible to check the blood pressure of ALL Azane users to determine what proportion of users actually experience high blood pressure. In addition, many Azane users might have high blood pressure even if they never took Azane. Factors such as these greatly complicate the process of determining the link between taking Azane and developing high blood pressure.

So, how do we go about making such a determination? The field of statistics has developed methods to deal with these issues. Since they are not perfect, statistics also allows us to determine how “imperfect” our results may be.

Since it is almost never possible to test an entire population, statistical methods make it possible to make **inferences** about the **parameters** of a population by taking a **sample** (i.e., a subset of the population) randomly selected from the population, measuring the characteristics of the sample, and then generalizing these characteristics to the population. The process of generalization of the sample characteristic to the population is known as **inferential statistics**. The actual characteristic of the sample is known as a **statistic** about the sample. Visually, this process may be represented as follows:



Two major questions arise in taking samples from populations, a process called **sampling**. The first question is, how many of the elements in the population must be sampled to be able to use the measured statistics to make inferences about the population parameter? We've all seen polls that sample 1,000 voters to determine the outcome of an election whose population (all voters) may total in the tens of millions!

The second question is, if a sample is drawn from a population and a statistic about the sample is calculated, how accurately will it represent the true population parameter of the measured characteristic? In other words, will our inference be accurate?

In order to answer these and other questions, the field of statistics has developed a process to collect and manipulate data mathematically. When followed, the process will assure that we sampled the right number of elements and tell us how accurate our results are. The process proceeds in the following steps:

- 1) Plan the experiment – what is the population, characteristic to measured, what results need to be measured, size of sample necessary, etc., **remembering that our goal is to provide meaningful information about a population.**
- 2) Develop methods to collect the data – survey, direct measurement, public record, previous studies or studies done by other groups. Important in this step is assuring that our data is collected in a **random** manner. That is, that it will reflect all parts of the population and not just a part that has something in common.
- 3) Obtain the data in accordance with 1) and 2)
- 4) Analyze the data – apply statistical tools to the collected information.
- 5) Interpret the data – what does it really mean? Its meaning is derived from the way in which the data was collected and calculated. **We may try to impart a different meaning to the data, but its meaning can never be separated from the mathematical basis of its calculation.**
- 6) Draw conclusions and make inferences about the data – the ultimate goal of the exercise.

Consider this experiment from a recent pharmaceutical study: The purpose of the study is to assess the impact of drug x on surgical pain following knee surgery as compared to the narcotic pain relievers currently used. 1) The plan: Patients who agree to participate in the study receive two capsules of drug x or a placebo immediately before the surgery in place of the intravenous narcotic usually administered. Neither the patients nor the surgeons know if the patient is receiving the real drug x or the placebo. After surgery, the patient is instructed to take another tablet of drug “x” (or placebo) only if needed for pain. If pain relief is still needed, another narcotic medication is provided.

180 subjects are to be tested at about 10 doctors' offices around the country. Each patient who agrees to participate will be given a screening visit by the surgeon to determine if they are medically eligible for the study and a post-surgical visit to determine if they maintained their eligibility throughout the study period. Patients who complete the study receive any additional medicines needed free of charge and also receive \$150 cash on completion of their study period.

2 & 3) Data Collection: Following surgery, the patient must complete a log to record pain levels and medications taken. The surveys are given to the patient by the surgeon and collected by the surgeon at the completion of the study period. Instructions as to completion of the survey are also given to the patient by the surgeon. The surgeon then returns the data to the pharmaceutical manufacturer.

4 through 6) Analysis, interpretation, conclusions: The data so collected is used to determine if drug x is a suitable alternative to the normal narcotic treatment. Appropriate statistical tools and tests are applied to the data to determine the effectiveness of the treatment.

After reviewing the survey, what statistical pitfalls do you see in using the data to change approved surgical procedures for knee surgery?

Types of functions used in statistics

Below are two graphs of functions which show the relationship between x and y for a straight-line and a normal distribution, as described above.

Straight – line (linear function)

Normal distribution (non-linear)

Both of these functions are said to be **continuous**. This means that, for every value of “x” over a certain range, a value of “y” can be calculated using the function (formula). Hence, for the straight line $y = 2x - 1$, a value of y can be calculated for $x = 6$, $x = 7$, $x = 6.2251$, $x = 8.013253$, etc. In fact, you cannot find a value for x for which it is impossible to calculate a value for “y”. Since any x selected will work, we say the function is continuous.

By contrast, a function in which only certain values of “x” are possible is said to be **discrete**. For example, we take a sample of 500 persons and administer the drug Azane to each person for a period of six months. Afterwards, we measure their blood pressure. How many people in our sample could possibly have high blood pressure? The number of people who could have high blood pressure in our sample of 500 could be 0 or 1 or 11 or 500 or any integer in between. Note that we cannot have 51.0435 people that have high blood pressure. It must be 50 or 51 or some other **finite, countable number** of people. This is the characteristic of a **discrete function**.

Homework: p. 14, Exercise 4
pp. 19-20, Exercises 5, 6, 12, 13, 14

Chapter 2

Describing Data: Frequency Distributions and Graphic Presentations

Discrete Functions

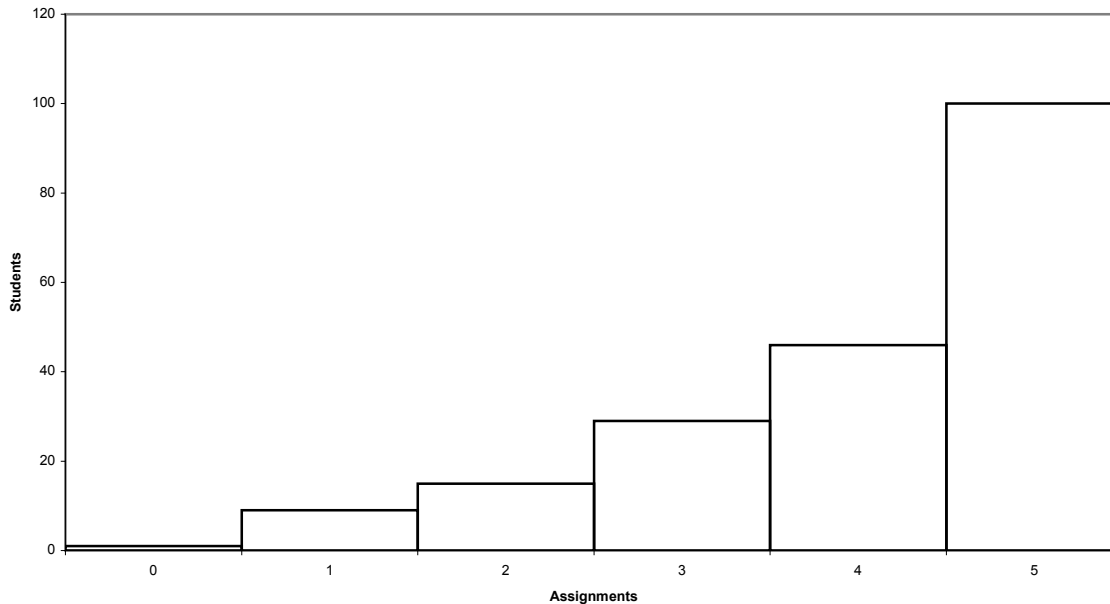
If we were to graph a discrete function, the fact that not every x on the x -axis is used forces our graph to be a bar chart, or **histogram**.

Suppose we wanted to count the number of students that turned in each of their five homework assignments in a class of 200 students. The data is presented in the **frequency distribution** below:

<u>No. of assignments turned in (x)</u>	<u>No. of students (f)</u>
0	1
1	9
2	15
3	29
4	46
5	<u>100</u>
	200

Note that x is the number of homework assignments turned in. It can ONLY be 0 (none turned in) up to 5 (all of them turned in). The graph of such a function must look like this: (see next page)

Number of Students Turning in Homework



The number of students turning in a certain number of homework assignments is termed the **frequency** associated with each x (no. of assignments). Hence, the table above is referred to as a **frequency table** and the bar chart as a **frequency histogram**.

In statistics, we are frequently interested in knowing the area between the “ y ” values and the x -axis. In the case of a bar chart or histogram, the area of each bar is defined as $\text{Area} = \text{width} \times \text{height}$. Note that for each bar the width is 1 unit (assignment). It increases in units of 1 assignment at a time. The height of each bar is its frequency (number of students). The area of this histogram can be determined by applying the formula $\text{Area} = \text{width} \times \text{height}$ to each bar and then adding the areas as follows:

$$\begin{aligned}
 A_1 &= w_1 \times h_1 = 1 \times 1 = 1 \\
 A_2 &= w_2 \times h_2 = 1 \times 9 = 9 \\
 A_3 &= w_3 \times h_3 = 1 \times 15 = 15 \\
 A_4 &= w_4 \times h_4 = 1 \times 29 = 29 \\
 A_5 &= w_5 \times h_5 = 1 \times 46 = 46 \\
 A_6 &= w_6 \times h_6 = 1 \times 100 = 100
 \end{aligned}$$

Therefore, $\Sigma A_i = 1 + 9 + 15 + 29 + 46 + 100 = 200$, which is the total number of students. **We will call this quantity Σf (sum of the frequencies) or “ n ” (sample size). For our purposes here, these terms are interchangeable.**

If it is of interest to know the percentage of students that turned in 0 assignments, 1 assignment, etc., we calculate the **relative frequency** for each number of homework assignments – that is the percent of total. Note especially that the total of the relative

frequency column is 1.000. **This is because the total represents 100%. All students turned in either 0, or 1, or 2, or 3, or 4, or 5 assignments.**

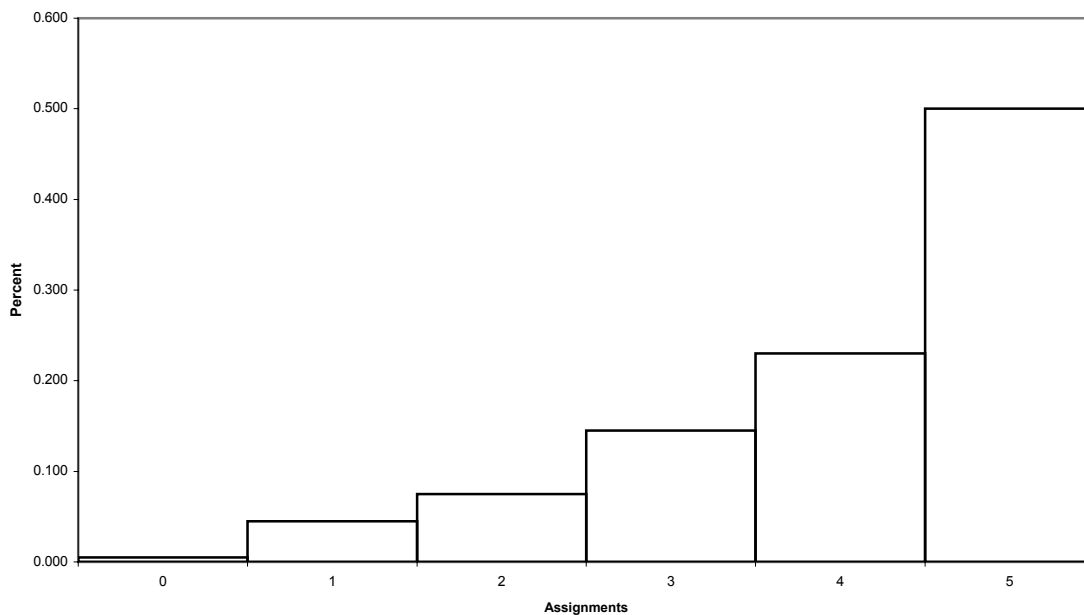
No. of assignments turned in (x)	No. of students (f)	% of students (rel. f)
0	1	0.005
1	9	0.045
2	15	0.075
3	29	0.145
4	46	0.230
5	100	0.500
	200	1.000

If you randomly select a student from the class, what is the probability that the student turned in all 5 homework assignments? _____. Only 3 homework assignments? _____. 3 or more homework assignments? _____.

Is there a connection between relative frequency and our normal understanding of probability?

The histogram of the above relative frequency table looks just like the frequency histogram except that the y-axis is now in terms of percent, not number of students.

Percent of Students Turning in Homework



Calculate the area under the bars for the relative frequency histogram:

$$A_1 = w_1 \times h_1 = 1 \times .005 = .005$$

$$A_2 = w_2 \times h_2 = 1 \times .045 = .045$$

$$A_3 = w_3 \times h_3 = 1 \times .075 = .075$$

$$A_4 = w_4 \times h_4 = 1 \times .145 = .145$$

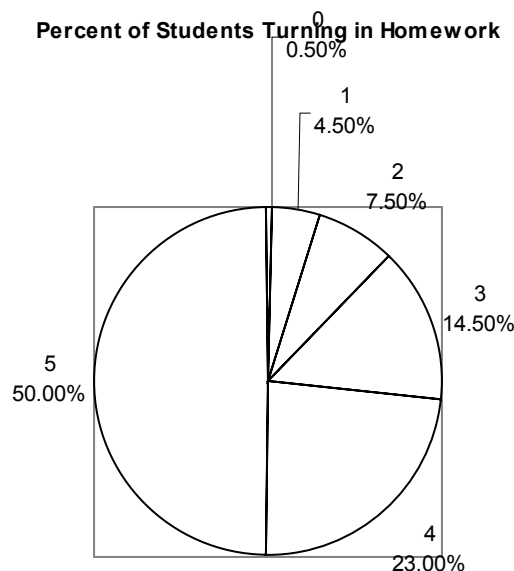
$$A_5 = w_5 \times h_5 = 1 \times .230 = .230$$

$$A_6 = w_6 \times h_6 = 1 \times .500 = .500$$

Notice that $\sum A_i$ (for rel. frequency) = $.005 + .045 + .075 + .145 + .230 + .500 = 1.000$

This is so because it includes 100% of the students when expressed as a percentage. It is important to recognize that, in statistics, any percentage (probability) graph will always have the characteristic that the area between the graph and the x-axis will represent 1.000, i.e., 100% of the total.

It is also sometimes very useful to present percentage data in the form of a pie chart. The pie chart has the advantage that it is readily obvious what represents 100% and the relative size of each A_i is easily visualized.



Frequently, the data that we have is not as clean and neat as the number of assignments turned in. Suppose we are concerned with daily sales totals for a large retail outlet. Since every day has a different amount of sales, we would be forced to list all the days separately, thus making for a very large frequency distribution of up to 365 lines for a single year. Additionally, the data would be very hard to understand because it is raw data and there is a lot of it.

In order to make the data understandable, we will break the daily sales figures into several brackets, or **classes**, each one containing a certain range of sales. Then, we will count each day in that class which includes the sales figure for that day. Such a classified frequency table would look like this:

Daily Sales (Class)	No. of days (frequency f)	Relative Frequency (%)
\$5,000 up to \$10,000	1	0.0625
\$10,000 up to \$15,000	3	0.1875
\$15,000 up to \$20,000	6	0.3750
\$20,000 up to \$25,000	4	0.2500
\$25,000 up to \$30,000	<u>2</u>	<u>0.1250</u>
	16	1.0000

Notice that there are 3 days (18.75% of the days) that have sales between \$10,000 and \$15,000. This is all we know about the sales on these days. We do not know the exact amount of sales on those days, only the range. For calculation purposes, we assume that all 3 days had sales at the midpoint of the range. This is the value that would be used as “x” in any calculations done on daily sales. The midpoint is found by adding two consecutive lower class limits and dividing by 2. Thus, the midpoint of the class “\$10,000 up to \$15,000” is $(10,000 + 15,000) / 2 = \$12,500$. We could assume that all three days in this class had sales of \$12,500. Though this is unlikely to be exactly true, it does enable us to perform calculations on daily sales with only a small amount of error caused by the assumption. The midpoint of each class is called the **class midpoint** and is the “x” in any calculations we do.

Adding the class midpoint to the table:

Daily Sales (Class)	Class Midpoint "x"	No. of days (frequency f)	Relative Frequency (%)
\$5,000 up to \$10,000	\$7,500	1	0.0625
\$10,000 up to \$15,000	\$12,500	3	0.1875
\$15,000 up to \$20,000	\$17,500	6	0.3750
\$20,000 up to \$25,000	\$22,500	4	0.2500
\$25,000 up to \$30,000	\$27,500	<u>2</u>	<u>0.1250</u>
		16	1.0000

Other important definitions for classified frequency tables:

Lower Class Limits – the **smallest** numbers that can actually belong to the different classes. In the example above, the lower class limits are \$5,000, \$10,000, \$15,000, \$20,000, and \$25,000.

Upper Class Limits – the **largest** numbers that can actually belong to the different classes. In the example above, the upper class limits are \$10,000, \$15,000, \$20,000, \$25,000, and \$30,000.

Class Midpoint (“x”) – the midpoints of the classes. In the example above, the class midpoints are $(5,000 + 10,000)/2 = \$7,500$; $(10,000 + 15,000)/2 = \$12,500$; $(15,000 + 20,000)/2 = \$17,500$; $(20,000 + 25,000)/2 = \$22,500$; and $(25,000 + 30,000)/2 = \$27,500$.

Class Interval – the difference between two consecutive lower class limits. In the example above, the class intervals are $(10,000 - 5,000) = \$5,000$; $(15,000 - 10,000) = \$5,000$; $(20,000 - 15,000) = \$5,000$; $(25,000 - 20,000) = \$5,000$; and $(30,000 - 25,000) = \$5,000$. Note that the class intervals are all equal. For reasons of visual presentation and intuitive understanding of the frequency table, each class interval should be chosen to be equal when possible.

Relative Frequency – the relative frequency represents the percent or proportion of all frequencies that fall within that class. That is

Relative frequency of a class = $(\text{class frequency})/(\text{sum of all frequencies})$
The relative frequencies of the classes are: $(1/16) = .0625$; $(3/16) = .1875$; $(6/16) = .3750$; $(4/16) = .2500$; and $(2/16) = .1250$. Note that the relative frequencies must add to 1.000, since their total represents 100% of the days.

There are certain rules that should be followed when constructing classified frequency tables. These rules are designed to make the tables as useful and accurate as possible.

1. All classes should be mutually exclusive. That is, there must be no overlap among the classes. For example, we could have used the classes \$5,000 - \$10,000; \$10,000 - \$15,000; etc. However, where would we tally a sale price of exactly \$10,000? We wouldn't know which class to use or use it consistently. Therefore, we expressly label the classes as "\$5,000 up to \$10,000" and "\$10,000 up to \$15,000". With these labels, we know that a value of exactly \$10,000 would be placed in the class "\$10,000 up to \$15,000".
2. Include **ALL** classes, even if the frequency is zero.
3. Use the same class width for all classes when possible.
4. All classes should have convenient numbers for the class limits.
5. The number of classes should be between 5 and 20. Less than five provides limited information, and more than 20 provides too much detail for effective presentation.

Frequency Histograms for Classified Frequency Tables

Histograms for classified frequency tables are constructed the same as other histograms except the labels on the x-axis are changed to reflect the nature of the data, that is the classes. **The x-axis labels used are the class limits.** The “bar” then extends from lower class limit to upper class limit.

Data: 728 774 859 882 791 731 838 862 880 831
759 774 832 816 860 856 787 715 752 778
829 792 908 714 839 752 834 818 835 751
837.

Construct a frequency table of 6 classes and identify all pertinent class parameters (interval, limits, etc.)

Homework: p. 31, Exercises 1, 2, 4, 6, 7
pp. 37-38, Exercises 9, 10, 12
pp. 41-42, Exercises 14, 16
pp. 46-52, Exercises 18, 20, 26, 30, 33, 34, 38, 42

Chapter 3

Numerical Measures

Mean or “average”

To determine the mean of a population or a data sample, add all the scores in the population or sample and divide it by the number of scores. The population mean is denoted by the symbol μ (Greek letter "mu") and the sample mean by \bar{X} , pronounced x-bar.

If the number of values in the population is N , then the mean of the population is given by:

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

If a sample of size n is drawn from a population of size N , then the mean of the sample is given by:

$$\bar{x} = \frac{\sum x}{n}$$

For the sample data given by the numbers 5, 7, 10, 12, and 14, (5 numbers) the mean is calculated as follows:

$$\bar{x} = (5+7+10+12+14)/5 = 48/5 = 9.6$$

The mean is usually the best measure of central tendency, especially when combined with the standard deviation (discussed later).

Example: The following are the times (in minutes) required to resolve customer complaints: 8, 1, 2, 4, 12, 7, 8, 10, 10, 11, 10. Find the mean time.

Homework: p. 62, Exercises 1, 3, 5, 7, 9

Median

The median is the middle score when the scores are arranged in order of increasing magnitude.

For the sample data 5,7,10,12,14, the median is 10 – as many numbers appear above 10 as appear below ten. This works when the number of scores is odd, because there is a definite middle number in an odd number of scores.

If there is an even number of scores, like 6, 5, 7, 10, 12, 14, 16, there is no definite middle number. Therefore, we **average** the **two middle numbers** to determine the median. In this case, the median is

$$(10 + 12) / 2 = 22/2 = 11.$$

The median is often a good choice to measure central tendency, especially when there are some extreme scores or outcomes in the sample that could affect the mean.

Example: The following are the times (in minutes) required to resolve customer complaints: 8, 1, 2, 4, 12, 7, 8, 10, 10, 11, 10. Find the median time.

Mode

The mode is simply the number that appears most frequently in a list of sample outcomes. For example, in the sample data 1, 2, 2, 2, 3, 4, 4, 5, 6, 7, 9, the number 2 is the mode as it appears the most number of times. If another number appeared the same number of times as the number 2 (i.e. 3 times), they would both be considered to be modes and the sample would be called **bimodal**. If there are more than two modes, the sample is said to be **multimodal**.

Example: The following are the times (in minutes) required to resolve customer complaints: 8, 1, 2, 4, 12, 7, 8, 10, 10, 11, 10. Find the mode.

Range

The range of a sample is the difference between the highest and lowest numbers. For example, in the sample data 1, 2, 2, 2, 3, 4, 4, 5, 6, 7, 9, the range is found by taking 9 – 1. Therefore, the range is 8.

Example: The following are the times (in minutes) required to resolve customer complaints: 8, 1, 2, 4, 12, 7, 8, 10, 10, 11, 10. Find the range.

Midrange

The midrange is found by adding the highest and lowest scores and dividing by 2. In other words, it is the average of the highest and lowest numbers. Using the sample data above, the midrange is $(9+1)/2 = 5$.

In terms of a formula: $\text{Midrange} = (\text{highest score} + \text{lowest score})/2$.

The range and midrange are very sensitive to extreme values and hence are rarely used as measures of central tendency.

Example: The following are the times (in minutes) required to resolve customer complaints: 8, 1, 2, 4, 12, 7, 8, 10, 10, 11, 10. Find the midrange.

Weighted Mean

A characteristic of the mean is that each score or outcome counts the same as every other score. This may not always be appropriate. For example, in class the instructor may want to give a greater weight to the final exam than to the other exams given during the course. In this case, we would have to determine the “weighted” mean because the straight mean would not be able to reflect the extra importance of the final exam in determining the final grade.

Assume that in Statistics there are four exams. The final is to count 40% of the grade, and each of the other three tests count 20% each toward the final grade. The weighted average grade will then become the final grade.

<u>x</u>	x	w	=	Contribution to total
Score		<u>weight</u>		<u>grade</u>
75		0.20		15
80		0.20		16
75		0.20		15
90		<u>0.40</u>		<u>36</u>
Final grade		1.00		82

$$\text{Weighted mean} = \sum(wx)$$

Note that a straight mean of the four scores (equal weights) would have resulted in a final score of $(75 + 80 + 75 + 90)/4 = 80.0$, a different result.

When we write the above “table” in the form of a formula, we have the following generalized formula:

$$\bar{x} = \frac{\sum(w x)}{\sum w}$$

In the example above, note that $\sum w = 1.0$ because each weight is expressed as a percent of the total grade. Note how similar the table above is to a frequency table. This is because in a frequency table the values of x are “weighted” by the frequencies “ f ”, as discussed below.

Mean from a Frequency Table

To get the mean of sample data presented in a frequency table, it is necessary to take the weighted mean of the data where the class midpoints are considered to be the scores or outcomes (x) and the frequencies (or relative frequencies, as appropriate) for each class are considered to be the weights (w). This effectively means that we assign the class midpoint as the value for each outcome or frequency in the class. We can construct a table (from daily sales data):

<u>x</u> <u>Class Midpoint</u>	<u>f</u> <u>Weight (frequency)</u>	<u>fx</u> <u>Contribution to</u> <u>Total</u>
7,500	1	7,500
12,500	3	37,500
17,500	6	105,000
22,500	4	90,000
27,500	2	55,000
Totals	16	295,000
	($\sum f$)	($\sum fx$)

$$\text{Weighted mean} = \bar{x} = (295,000)/16 = \$18,437.50$$

Or, using the weighted mean formula $\sum(w x) / \sum w$ and substituting f (frequency) for w (weight) and x (class midpoint) for x (score) gives the formula

$$\bar{x} = \sum(f x) / \sum f$$

Example:

Class <u>Distance (miles)</u>	Class Midpoint <u>x</u>	Frequency <u>f</u>	Contribution <u>fx</u>
0 up to 40		17	
40 up to 80		41	
80 up to 120		80	
120 up to 160		49	
160 up to 200		4	

Weighted mean = mean of the distribution =

Homework: p. 64, Exercises 12,13
 p. 67, Exercises 15, 17, 19
 p. 70, Exercise 22

Standard Deviation and Other Measures of Sample Dispersion

The use of the mean alone can be misleading. Consider that you are planning to have family come into Columbia from all over the country. You want to select a date when the temperature would be good for a picnic, say 70°. You find two dates with a mean or average temperature of 70. The first date is June 15 and the second date is September 21. Since they both have the same average temperature, you might be indifferent as to which date to choose. However, you look at the last three years temperature data for these three dates and find the following:

	<u>June 15</u>	<u>Sept. 21</u>
2000	69	57
2001	70	73
2002	<u>71</u>	<u>80</u>
Mean	70	70

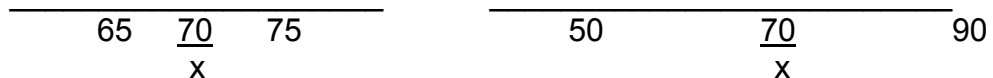
Now which date would you choose?

The deficiencies of knowing only the mean are seen in the above example. Wouldn't it be good to also have a measure which would let us compare the two dates as to variability of temperature also. The **standard deviation** is just such a measure.

Look at the above data in a graphical format:

June 15

Sept. 21



An obvious solution seems to be just to add up the differences of each score (x) from the mean of the sample ($x - \bar{x}$) and take the mean of these values. Then, wouldn't the sample with the greatest dispersion have the highest value?

$69-70 = -1$	$57-70 = -13$
$70-70 = 0$	$73-70 = 3$
$71-70 = 1$	$80-70 = 10$
Total 0 deg.	0 deg.

Mean	$0/3 = 0 \text{ deg.}$	$0/3 = 0 \text{ deg.}$
------	------------------------	------------------------

Every time we try this, the result is zero, so it isn't of much use. Suppose that all the numbers were made positive, then added. Mathematically, the best way is to square the numbers.

$(69-70)^2 = -1^2 = 1$	$(57-70)^2 = -13^2 = 169$
$(70-70)^2 = 0^2 = 0$	$(73-70)^2 = 3^2 = 9$
$(71-70)^2 = 1^2 = 1$	$(80-70)^2 = 10^2 = 100$
Total 2 deg. ²	278 deg. ²

Mean	$2/3 = .67 \text{ deg}^2$	$278/3 = 92.67 \text{ deg}^2$
------	---------------------------	-------------------------------

This looks much better, except it is in degrees squared while the original data is in degrees. To get back to degrees, we take the square root. The units of the square root of degrees squared is degrees, so now the measure looks reasonable. In addition, from the mathematical derivation of this concept, it is necessary to calculate the mean of the differences by dividing by n-1, not n. Making these changes leads to:

$$\begin{aligned}\text{Std deviation} &= \sqrt{(2/2)} \\ &= 1 \text{ degree}\end{aligned}$$

$$\begin{aligned}&\sqrt{(278/2)} \\ &11.8 \text{ degrees}\end{aligned}$$

Hence, the standard deviation is the average (mean) of the difference between each x and the mean of the x 's. Each day has a mean of 70° , but June 15, on average, varies from 70° by only 1° , while Sept. 21, on average, varies from 70° by 11.8°

Writing as a general formula,

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

The standard deviation of the population, σ , is exactly the same concept, but it applies to the population parameter, not the statistic, and the population symbols are used:

$$\sigma = \sqrt{\sum(x - \mu)^2/N}$$

The **mean deviation** is another such measure. It measures the average distance that each x in a sample is from the sample mean, \bar{x} . However, since the mean deviation uses absolute value to make the sum of $x - \bar{x}$ a value other than zero, it is not often used. The absolute value is difficult to use in mathematical calculation due to the fact that one must manually adjust the value of a negative number to a positive number:

$$MD = \sum|x - \bar{x}|/n$$

We will use the following notation to denote standard deviation:

- s for the standard deviation of a sample
- σ for the standard deviation of a population
- s^2 for the variance of a sample (this is the std dev. without taking the square root).
- σ^2 for the variance of a population (the std dev. without taking the square root).

Note that variance is the standard deviation multiplied by itself (squared). Therefore, the variance is the number that appears beneath the square root sign in the standard deviation calculation. It is used some in statistics, but it must be kept in mind that its units (in this case degrees²), are also squared.

Example: The following are the waiting times (in minutes) required to resolve customer complaints: 8, 1, 2, 4, 12, 7, 8, 10, 10, 11, 10. Find the standard deviation (s) and variance (s^2) of this sample.

\underline{x}	$\underline{x-\bar{x}}$	$\underline{(x-\bar{x})^2}$	n =
			$\bar{x} =$

Homework: pp. 76-77, Exercises 31, 33, 35
 p. 79, Exercises 37, 39, 42
 p. 81, Exercises 43, 45, 47

Standard Deviation Shortcuts

A couple of shortcut formulas exist to calculate the standard deviation of a sample. These will always produce the same results as the derived formula above.

For data that is not grouped into classes the formula is

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

Example: For the waiting time example, calculate the standard deviation using the shortcut formula.

xx²

For data that is grouped into classes, remember that each “x” represents the class midpoint of one of the classes and the frequency “f” represents the number of times we are to count that “x”. Therefore, each x must appear f times in the formula:

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

Example: For the distance example under “weighted mean”, calculate the standard deviation using the shortcut formula.

Class	Class Midpoint	Frequency		
<u>Distance (miles)</u>	<u>x</u>	<u>f</u>	<u>fx</u>	<u>fx²</u>
0 up to 40		17		
40 up to 80		41		
80 up to 120		80		
120 up to 160		49		
160 up to 200		4		
Totals				

Homework: Work some of the problems of the previous section using the shortcut formulas.

Estimations of Standard Deviation

The Range Rule of Thumb

For many samples and populations that are close to “normal” (defined later), the standard deviation can be approximated by determining the range of the data and dividing by four.

$$s = \text{range}/4$$

That is, the range of the data is generally 4 standard deviations wide. This technique is good for “testing” the standard deviation calculation, although for some populations and samples, the range rule could be off by a wide margin.

Example: For the standard deviation calculated in the waiting time example above, what is the standard deviation using the range rule of thumb? How does it compare to the real standard deviation?

The Empirical Rule

The empirical rule states that, for normal distributions (“bell-shaped”), 68% of all outcomes or scores will fall within one standard deviation of the mean, while 95% will fall within two standard deviations, and 99% within three standard deviations. As noted in the range rule of thumb, 4 standard deviations would encompass the whole sample.

Mathematically, **68%** of all outcomes lie within the range determined by

$$(\bar{x} - s) \text{ and } (\bar{x} + s)$$

Example: In the waiting time example above, what range of waiting times includes approximately 68% of all waiting times? Can this be verified in this example?

Homework: p. 84, Exercise 51
pp. 86-88, Exercises 53, 55, 58, 65, 68,69