Ch. 9 ELEMENTARY STATISTICS

You've been using statistics most of your life without ever thinking too much about it. Now, we're going to formalize some of that knowledge.

The term "statistics" refers to both a set of data (information) and methods used to analyze the data.

Imagine your little brother or sister running home from school all excited one day telling you he got eight right on a spelling test. You might wonder, what does that mean? Is that good or bad? You might decide you need more information to make that decision. So you ask, how many questions were on the test?

See you received data, then you tried to analyze it. That's statistics. Often times you find you might need more information to analyze the data, so you ask more questions. Don't you just love math? You really get to participate.

The first statistic that many of us learned formally was a percent. You may have taken a quiz in school and got 14 correct out of 17, you wondered what that meant. The teacher may have converted that score to a percentage – 82%, then said you eared a "B". The interpretation given to a grade of "B" in school is above average. Again, you received information, then using a statistical method (finding percentages), analyzed the information.

It's possible to look at a whole set of information and use a single number to describe it. Amazing, don't you think? Have you ever done that before? My guess is the answer to that is yes. You might be a bowler, golfer, basketball player, or you might have received a grade in school. A single number described your performance.

However, when that is done, some information will be lost is such a simple description.

So, let's chat. The first Law of Statistics is : *Ask the next question*.

We are going to define three measures of central tendency, most of us refer to those measures as averages, that are used to describe a lot of data with a single number.

<u>3 Measures of Central Tendency</u>

1. Mean
2. Median
3. Mode

Mean

The mean is the one you are probably most familiar with, it's the one often used in school for grades. To find the mean, you simply add all the scores and divide by the number of scores. In other words, if you had a 70, 80, and 90 on three tests, you'd add those and divide by three. The mean is 80. Your average is 80.

To find the mean, you need to know two pieces of information. The ***total*** and the ***number of pieces of data***.

Mathematically, we write: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ where $\bar{x}$ is the mean, n is the number of pieces of data and x_i represents the data.

Example 1 Find the mean of the following data: 78, 74, 81, 83, and 82.

The sum of the data is 398. There are 5 scores, so the mean is 79.6

Example 2 In Ted's class of thirty students, the mean on the math exam was 80. Andrew's class of twenty students had a mean of 90. What was the mean of the two classes combined?

The sum of the scores in Ted's class, 30 x 80 = 2400
The sum of the scores in Andrew's class, 20x90 = 1800

Total points of two classes combines is 4200. There 30 students in one class and 20 students in the other class, so there are 50 scores.

So, 4200/50 = 84 is the mean of the 2 classes combined.

Example 3 Ted's bowling scores last week were 85, 89, and 101. What score would he have to make on his next game to have a mean of 105?

With a mean of 105, the sum of the 4 scores would be 4x105 = 420. The sum of his scores on the first three games was 275.

To get a total of 420 on his four games, he would need 420 – 275 = 145

Example 4 One of your students was absent on the day of the test. The class average for 24 students was 75%. After the other student took the test, the mean increased to 76%, what did the last student make on the test?

The total points for the students present was 24x75 = 1800

When the absent student too the test, the average was 76 and there are 25 students. 25x76 = 1900.

1900 – 1800 = 100, the student who was absent would have to make 100% to raise the mean to 76%.

A problem with using the mean is that it can be skewed by a very high or low score.

Median

The median, often used in finance, is the middle score when the data is listed in either ascending or descending order. If there is no middle score, then you take the two middle scores, add them and divide by 2. It's also referred to as an average.

Example 5 Find the median of 72, 65, 93, 85, and 55.

Rewriting in order, I have 55, 65, 72, 85 and 93. The middle score is 72, so the median is 72. Piece of cake, right?

Example 6 Find the median of the following data; 81, 55, 92, 64, 78, 81, 76 and 64.

Rewriting the same data in order from least to greatest, we have

55, 64, 64, 76, 78, 81, 81, 92

Since there is an EVEN number of pieces of data, there is no middle. So, I find the mean of the 2 middle numbers; 76 and 78. The mean of those two numbers is 77, so the median is 77.

Mode

If you have ever described the average weight of a particular population, the average height, shoe size, shirt size, the number of points scored in a particular type of game. Those are all examples of you using the mode.

The mode is a piece of information that appears most frequently. Can't get much easier than that.

Example 7 Find the model of the following data; 55, 64, 64, 64, 81, 76, 78, 81

What scores appears most often? If you said 64, you just named the mode.

Dispersion – Spread

These three measures of central tendency, most often referred to as averages, describe a set of data using a single statistic (number). By condensing that information like that, we might not see the whole picture. So just like when the little kid ran home and indicated he had 8 right, we had to get more information to determine what that meant. In math, we call that analyzing data.

Let's say we have three students, Abe, Ben and Carl. In math, we love to abbreviate, so we'd call them A, B, and C. They all went bowling, three games later, they all found they had a mean (average) of 80. Neato!

Abe's scores	80, 80, 80
Ben's scores	70, 80, 90
Carl's scores	65, 75, 100

Sure enough, when I add each one's scores and divide by three, they all have a mean (average) of 80.

Do you think one of those averages might be a better descriptor than the other two? Looking at Carl's scores, it appears he's a little erratic. It might be difficult to predict what he might score on the next game.

Abe, on the other hand, looks pretty stable. I might guess he'll score an 80 on the next game.

Now, knowing they both have the same average, I'm analyzing them and have determined that one mean is a pretty good descriptor, which would allow me to predict more comfortably what might happen next. In other words, Abe's mean is doing a better job of describing the data.

Carl's mean is not as good as a descriptor as Abe's. Although his average is 80 too, like Abe's, I'm not sure how good a predictor that statistic will be.

The point being, Abe's mean better describes what is occurring than Carl's mean. But, if I didn't see those scores, I would not know that. I wouldn't realize how consistent Abe is and how Carl is erratic because their averages are both 80. Then first rule of statistic should be; a*sk the next question*.

So, we look for more information, as we have done. One way to do this is to look at all the scores and try to determine consistency. In math, rather than looking at the whole set of data, we might want to examine his high and low scores. It might be someone just had one super high or low score that really affected the mean.

When we do this, we are trying to determine the spread of the scores. In statistics, that's referred to as dispersion.

There are four ways to measure this spread or dispersion.

<u>4 Measures of Dispersion</u>

1. Range
2. Mean Absolute Deviation
3. Variance
4. Standard Deviation

Range

The range is just the difference between the top score and the bottom score. The larger the range, the less likely the mean can be depended upon as a good descriptor or predictor.

Example 8 Find the range of the following:

Abe's scores 80, 80, 80
Ben's scores 70, 80, 90
Carl's scores 65, 75, 100

While the mean for each of the three people was 80, the range of Abe's scores was zero. The range of Carl's scores was 35 and Ben's range was 20. It would appear, the smaller the range, the mean is a more accurate descriptor.

Mean Absolute Deviation (MAD)

Another measure of dispersion is the Mean Absolute Deviation. This formula looks very much like the formulas for Variance and Standard Deviation.

The mean absolute deviation is the arithmetic average of the absolute value of the differences between each number and the mean of the collection of numbers. The MAD allows us to find the average distance the scores are from the mean. The reason we take the absolute value is if we did not, then scores above the mean would cancel out the scores below the mean.

To find the MAD

1. Find the mean of the data.
2. Subtract the mean from each data point and take absolute value
3. Add those absolute values of those difference
4. Divide that total by the number of data points.

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \overline{x}\right|$$

Let's use the data from **Example 9** and find the Mean Absolute Deviation.

Example In Carl's case, his scores were 65, 75, and 100. We've already determined the mean was 80. Now I subtract the mean from each of those scores and take the absolute value, find the sum of the absolute values and divide by ***n***, which in our case is 3.

$|65 - 80| = |-15| = 15; \quad |75 - 80| = |-5| = 5; \quad |100 - 80| = |20| = 20$

$$MAD = \frac{15+5+20}{3} = \frac{40}{3} = 13\frac{1}{3}$$

Just like other measures of dispersion, the smaller the measure, the more compact the scores are. Measures of dispersion give us an indicator how good a descriptor and predictor the mean actually is.

Variance

To find the variance,

1. find the mean of the group.
2. subtract the mean from every score.
3. square each of those differences.
4. add all those and divide by the number of scores.

Mathematically, that's said this way. The variance is the arithmetic average of the squared differences between each number and the mean of the collection of numbers. Sounds impressive, doesn't it? Remember, it's just arithmetic!

Example 9 Find the variance of Carl's scores

In Carl's case, his scores were 65, 75, and 100. We've already determined the mean was 80. Now I subtract the mean from each of those scores.

$$\text{Variance} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2$$

$65 - 80 = -15; \quad 75 - 80 = -5; \quad 100 - 80 = 20$

Squaring each, I have 225, 25, and 400. Now add and divide by three.

$$\frac{225+25+400}{3} = \frac{650}{3}$$

The variance is 216.6

If we found the variance for Ben's scores, we'd see the variance is smaller. Abe's variance is zero. In other words, Abe's scores did not vary. The smaller the variance indicates the scores are grouped closer or the mean is a better descriptor.

Standard Deviation

The most common measure statisticians use to describe the spread is called the standard deviation. To find the standard deviation, all you do is take the square root of the variance.

So, using Carl's scores, we found the mean was 80, his variance was 216.6. Now take the square root of that number.

$$\text{Standard Deviation} = \sqrt{216.6} = 14.7$$

I know what you are saying, you love this!

The standard deviation tells you how great the spread is in your data. The greater the standard deviation, the greater the spread. So, the next time you go bowling and need someone to fill in, ask their average. If you want to know if that average is a number you can depend on for your team's score, go ahead, ask what his standard deviation is. That way you'll know what to expect – or – what to hope for.

Since the standard deviation makes use of the mean in calculation, it includes all scores, it can be significantly affected by extremely high or low scores.

The standard deviation is based upon the deviation of each score from the mean.

Special note – sometimes statisticians divide by $\left(\overline{N}-1\right)$ instead of *N*, like we did. *N* is the number of scores.

Normal Distribution

Suppose you took a box of toothpicks, held them above a spot marked on a desk, then emptied them out all at once aiming for that spot.

What do you think might happen to the toothpicks? They'll land! Well if we looked closer, we might notice that most of the toothpicks stayed pretty close to where I dropped them. The further I got from the drop point, the fewer the toothpicks in any direction.

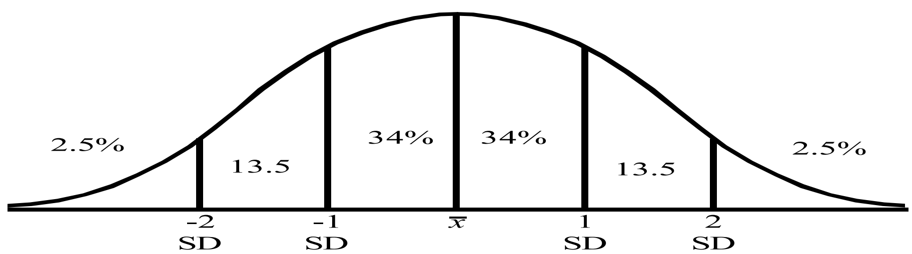
While that's interesting, that can also be applied to other samples. For instance, we looked at the height of a large sample of 35-year old males, we might see a similar situation develop.

Most of the men would be around the same height, and there would be fewer and fewer men that are either real tall or real short. In other words, the farther away from the middle height, the fewer the number of people.

That observation brings us to a discussion of the Normal Distribution. We might see the number of men shorter than the average seems to approximate the number of men taller. In math, we might describe that by saying if we graphed the result of our observations, it would be symmetric around the mean. We might also notice, if we used a standard deviation in calculations, that about 68% of all the heights fall within one standard deviation of the mean. In other words, from one standard deviation below the mean to one standard deviation above the mean, we'd expect to see 68% of the population. 34% on each side of the mean.

We could better see this through the use of a graph. It's called the Standard Normal Curve, lovingly referred to as the Bell Curve. Let me draw it, throw some numbers in there as we talk and play.

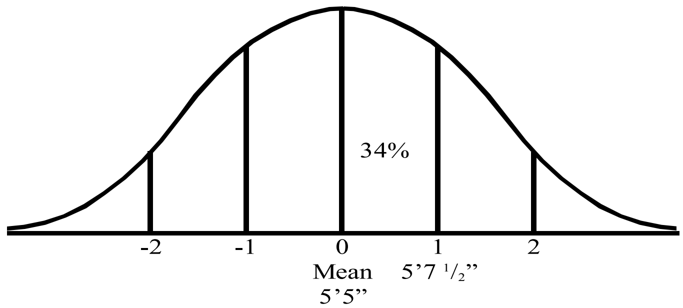
Standard Normal Curve = Bell Curve



I broke the percentages up between each standard deviation so you could readily see them. Notice scores falling within one standard deviation of the mean is 68%, 34% on each side of the mean; (34 + 34), and scores falling within two standard deviations is 95% $(13.5 + 34 + 34 + 13.5)$.

Let's look at a problem. Because of Title IX considerations, I'll use women in this example.

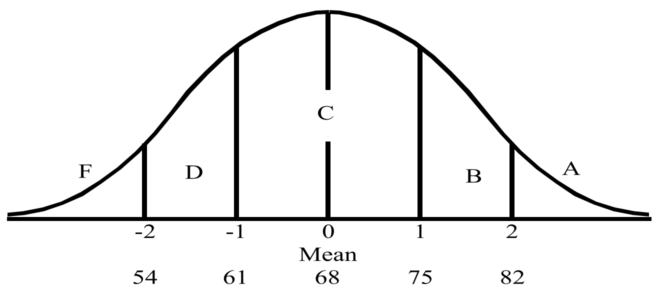
Example 10 The mean height of 500 women, normally distributed, in Las Vegas is 5 ft 5 inches with a standard deviation of 2.5 inches. How many women would you estimate to be between 5 ft, 5 in, and 5 ft 7.5 in? (Between the mean and one standard deviation above the mean?)



Looking at the Normal Curve, you'd think 34%. So, taking 34% of the 500 women, we'd predict 170 women.

Example 11 Let's look at a hypothetical class with a mean of 68 and a standard deviation of 7.

We'll also agree students falling within one standard deviation of the mean are average or C students. Students between one and two standard deviations above the mean are B students, two standard deviations and above are A students. Using the reasoning, what grade would you assign students whose grade is more than two standard deviations below the mean? I hope you said "F", otherwise I'll give you one.



Looking at that on the Bell Curve, we can see how this all falls out.

Notice that a student scoring 75 would be one standard deviation above the mean. Since his grade is on the dividing point, we'll naturally give him the benefit and assign a letter grade of "B". That is, if I like him.

There are many times we might want to compare information that are normally distributed. We might have two groups, A & B, Group A has a mean of 68 with a standard deviation of 4 on their test; Group B has a mean of 74 with a standard deviation of 8 on a similar test. Abe from Group A scored 70 while Ben form Group b also scored a 70, who performed better?

Well, since the tests were supposed to measure the same things, it might be difficult to compare those results. So, what do we do, we convert those distributions (scores) to what we refer to as ***z-scores***. That is, we convert both the groups distributions by making both means zero, then we can measure performance of test-takers by comparing on the same scale where their scores fall with respect to standard deviations. The nice thing about converting to z-scores is that now we can use a table to more accurately describe their performance.

Z- Scores

A z-score (also called a *standard score*) gives you an idea of how far from the mean a data point is. But more technically it's a measure of how many standard deviations below or above the population mean a raw score is by converting the mean to zero.

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

μ – is the mean
σ – standard deviation

The z-score changes the mean to zero.

Example: Let's say you have a test score of 190. The test has a mean (μ) of 150 and a standard deviation (σ) of 25. Assuming a normal distribution, your z score would be:

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

$$= \frac{190 - 150}{25}$$

$$= \frac{40}{25} = 1.6$$

That is 1.6 standard deviations above the mean using Bell Curve

Since z-scores are standardized, we can look up 1.6 on a chart to find a percentile rank that will be described below.

Percentiles

Percentiles are often used with standardized testing. The most common definition of a percentile is a number where a certain percentage of scores fall below that number. You may have scored 59 out of 80 on a test, but that might not have much meaning unless you know what percentile you fall into. If you know your score is in the 90th percentile, you know you scored better than 90% of the students taking the test. Rather than describing how well many problems were correct based on the total number of problems, percentiles describe how well you did compared to other people taking the test. Little kids use this concept when their parents ask them how they did on a particular test and they answer by saying how poorly everyone else did.

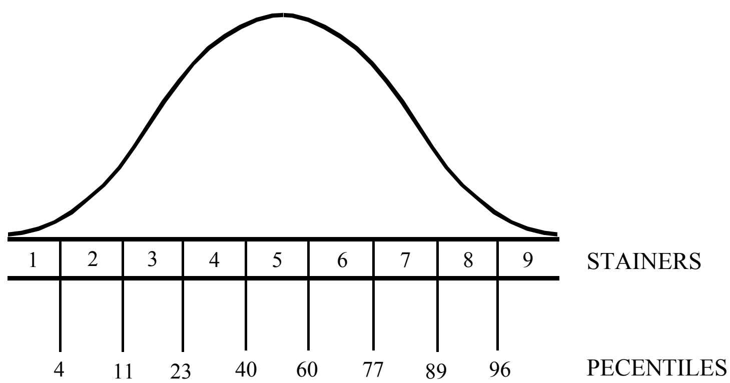
A simple example might be using the scores of the 5 people taking a test; the results were

86, 73, 54, 42, and 21. If you scored 73, there are 3 scores out of the total of 5 that scored below you, that would be the 60th percentile.

We can also find percentiles using the percentages in the Normal Curve. In the previous example, a student scored above 84% of the other students taking the test (Add the percentages of the groups below his score). We would say he's in the 84th percentile. A student scoring at the average would be at the 50th percentile. A person scoring 61 would be one standard deviation below the mean, he would be at the 16th percentile. In other words, he scored above 16% $(13.5 + 2.5)$ of the population.

Stanines are often used in education, as are quartiles. Both are merely percentile rankings. Quartiles divide the percentiles in four groups, stanines into nine groups.

I know, you want to visualize this. To do this I will again sketch the Standard Normal Curve. But rather than putting the percentages that would fall between standard deviations as I did earlier, I'll list the stanine and percentile scores.



Looking at this, we can see a student that has a stanine of 6 is in the 60th to 76th percentile range (77 starts the 7th stanine).

So, what does all this mean? It means we can look at information in different ways. While in school, teachers normally convert your raw score to a percentage, then assign a letter grade. Statisticians might look at that same information and do the very same thing. Or they might find the mean, determine the standard deviation, then see how you performed when compared to others using percentiles. Then, based on the range of those percentiles, determine what stanine you are in.

I know this is so interesting you are hungering and thirsting for more. But I'm hungry too, so I'm going to lunch while you try some problems.

Measures of Central Tendency

Find the mean, mode, median, range

1. 1, 2, 3, 3, 4, 7, 9, 11

2. 12, 14, 16, 13, 20

3. 1, 2, 2, 3, 4, 5, 6, 7, 5, 10

4. 1, 4, 4, 5, 8, 10, 12, 11, 8

5. 1, 3, 8, 12, 10, 8

6. 3, 14, 6, 2, 5, 7, 13, 14

7. 11, 12, 3, 5, 7, 16, 13, 6, 7, 10

8. The following table contains the number of traffic accidents in Nevada for the years 1960 – 1969. Find the mean, mode, median, and midrange for the number of accidents.

YEAR	NUMBER	YEAR	NUMBER
1960	436	1965	820
1961	833	1966	532
1962	714	1967	648
1963	1,201	1968	872
1964	749	1969	648

9. Employees working at the Akron plant of the United Bug company have complained that they are discriminated against in the company's pay scale when compared to the pay scale at the Canton plant. The employees and their salaries are listed below.

Akron Plant		Canton Plant	
Mr. Jones	$30,000	Mrs. Stein	$20,000
Ms. Arthur	$15,000	Mr. Patrick	$20,000
Mr. Brady	$15,000	Mr. Baron	$20,000

a. What is the mean salary for all employees?
b. What is the mean salary for the Akron workers? For Canton workers?
c. What is the median salary for Akron workers? For Canton workers?
d. What is the midrange salary for Akron workers? For Canton workers?
e. If you were a lawyer acting for United Bug, which measure of central tendency would you use? tendency would you use?
f. If you were a lawyer for the Akron workers which measure of central tendency would you use

10. The mean score of a set of 12 tests is 68. What is the sum of the 12 tests scores?

11. The mean score on a set of 15 college entrance exams is 87. What is the sum of the 15 exam scores?

12. Two sets of data are given: the first set of data has 20 scores with a mean of 50, and the second set of data has 33 scores with a mean of 75. What is the mean if the two sets of data are combined?

13. In a math class Joe takes 10 tests with a mean score of 79. To get a 'B' in the course students need a mean score of 80. What is wrong with the argument that Joe missed getting a 'B' by 1 point?

14. The table below indicates the grades for fifty freshmen registered for MAT 114. Find the mean, mode, median, midrange, and range for the grades of the freshmen.

Grade	# of People
A	4
B	18
C	10
-D	12
F	6

15, The table below indicates the heights for a group of teenagers. Find the mean, mode, median, range, and midrange for the teenagers' heights.

Height (inches)	# of People	Height (inches)	# of People
72	2	64	6
70	1	62	7
68	2	60	5
66	5	58	2

16. On a math test the following scores were made in a class of ten students: 81, 74, 87, 94, 71, 68, 72, 77, 81, 89. Find the mean, mode, median, range, midrange, and standard deviation for the set of data.

17. An experiment consists of tossing 8 coins and recording the number of heads that appear. The coins are tossed 10 times and the number of heads were 2, 3, 4, 5, 5, 6, 3, 2, 7, 3, respectively; find the mean, mode, median, range, midrange, and standard deviation for the number of heads shown.

Standard: Finding Measures of Central Tendency.

Problem Variations in Finding the Mean

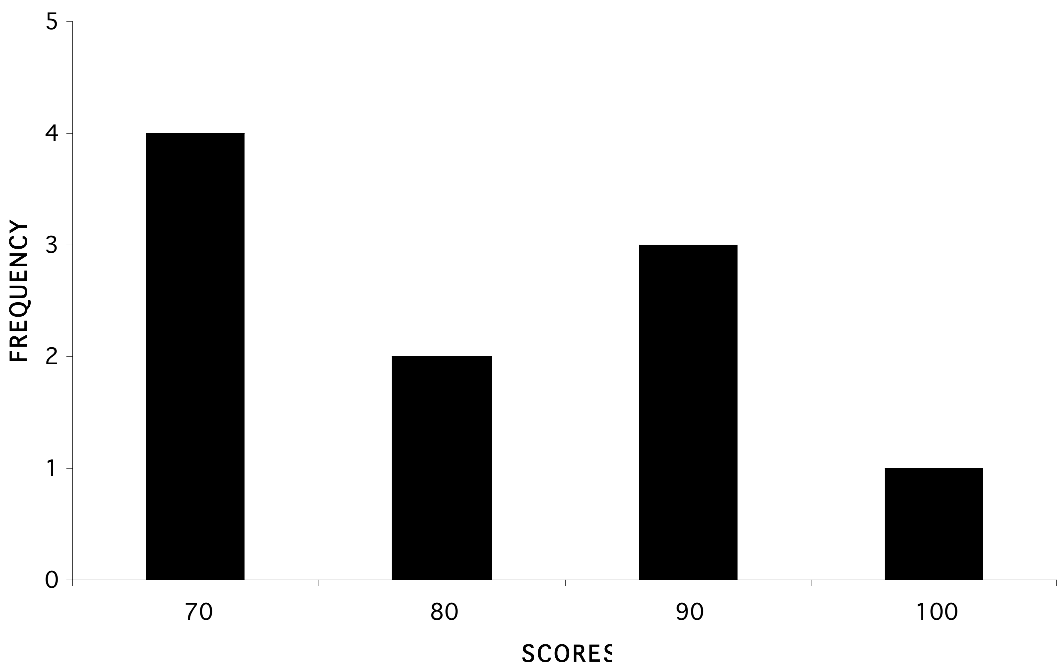
1. Find the mean of the following data: 78, 74, 81, 83, and 82.

2. In Ted's class of thirty students, the average on the math exam was 80. Andrew's class of forty students had an average of 90. What was the mean of the two classes combined?

3. Ted's bowling scores last week were 85, 89, and 101. What score would he have to make on his next game to have a mean of 105?

4. One of your students was absent on the day of the test. The class average for 24 students was 75%. After the other student took the test, the mean increased to 76%, what did the last student make on the test?

5. Use the following graph to find the mean.



STATISTICS

- Find the mean, median, mode, and range given raw data
- Find a missing score given other scores and the mean
- Find the mean, median, mode and range using bar, line, and frequency graphs
- Find the measure of the central angle given data for a circle graph.
- Construct bar, line, frequency, and pie charts.

Problem set

1. Bob bowled three games, his scores were 82, 85, and 88. Find his mean average.

2. Ted's scores on his tests were 62, 87, 75, 72, and 62. Find the mean, median, mode, and range.

3. A student has average score of 81 on three tests, if the student scored an 84 on the first two tests, what was the score on the third test?

4. Find the average rainfall per month if it rained 2.18 inches in June, 4.07 inches in July, 5.2 inches in August, and 1.07 inches in September.

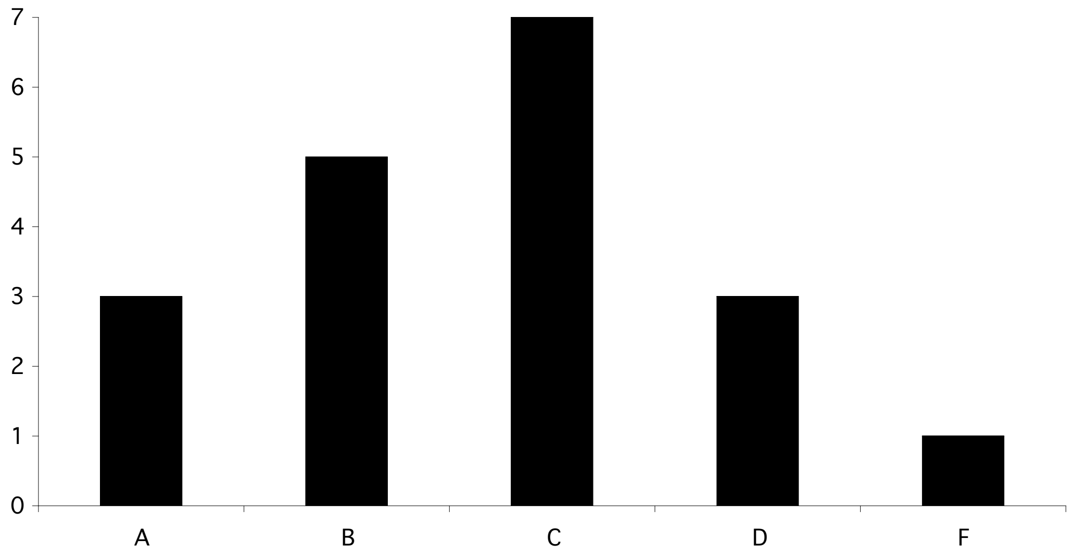
5. Bill scored 7, 12, 15, and 5 points in four basketball games. How many points must he score in the next game to have an average of 12 points per game?

6. The average income for 5 people is $150,000 per year. Four of the people earn $50,000, how much does the fifth person earn?

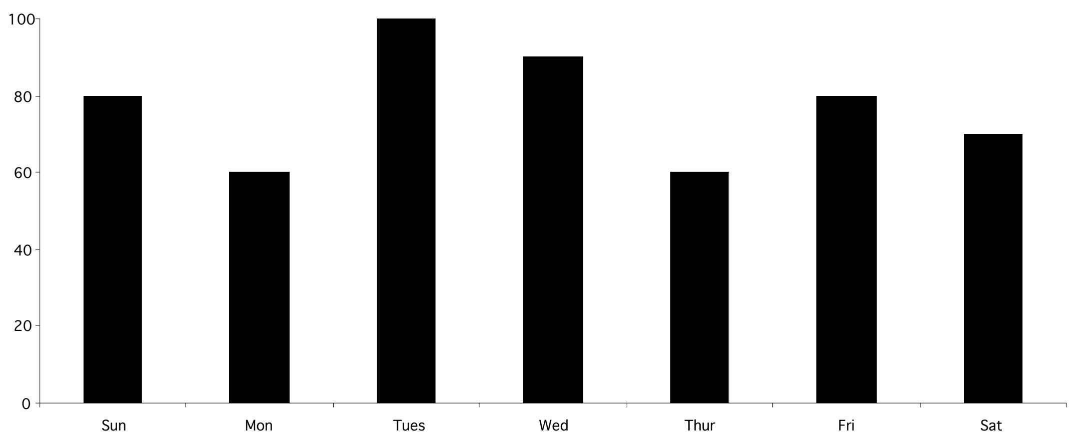
7. Find the median of the following list.
   7, 12, 8, 9, 10, 4, 15, 17, 20

8. John's average on his first four tests is 88. To earn an A, he must have an average of at least 90, what is the lowest grade he can make on his next test to earn an A?

9. Find the median and mode.



10. Find the average temperature (mean). For what days was the temperature below average? What was the range of the temperatures?



11. Each month the Smith family uses its income in the following way: 30% for food, 25% for rent, 20% for transportation, 10% for savings, 5% for entertainment, and 10% for other expenses. Construct a pie chart representing this information.

12. Each dollar the government obtains in taxes is spent in the following manner: 25¢ for defense, 30¢ for social security, 10¢ for subsidies, 15¢ for salaries and 20¢ on social programs. Construct a circle graph representing this data.

13. There are 2000 students attending a certain high school. There are 400 seniors, 300 juniors, 500 sophomores, 600 freshmen, and 200 5^{th} year students. Construct a circle graph showing this information.

14. The heights of 40 students in inches are given as follows:

    62, 65, 54, 55, 50, 73 73, 57, 64,52, 62, 61, 53, 68, 64, 70, 66, 71, 63, 54, 64, 66, 56, 57, 63, 68, 53, 64, 68, 58, 66, 58, 58, 56, 64, 53, 67, 67, 70, 62

    Construct a grouped frequency distribution for the following intervals: 75-72, 71-69, 68-66, etc. Find the median, mode, and range of the heights.

15. The following table contains the number of accidents last week. Find the mean, median, mode, and range for the number of accidents.

Monday	10
Tuesday	12
Wednesday	8
Thursday15	
Friday	8
Saturday	9
Sunday	8

16. Use the following table to find the mean, median, mode, and range for teenagers' heights.

Height	# of people	Height	# of people
72	2	64	6
70	1	62	7
68	2	60	5
66	5	58	2

    How many students have above average height?

17. The grade distribution for the final exam in math is as follows:

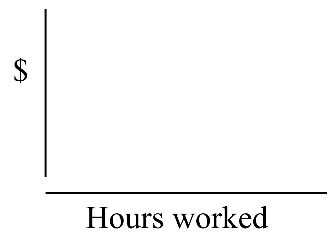
Grade	Frequency
A	4
B	10
C	37
D	8
F	1

    Find the median.

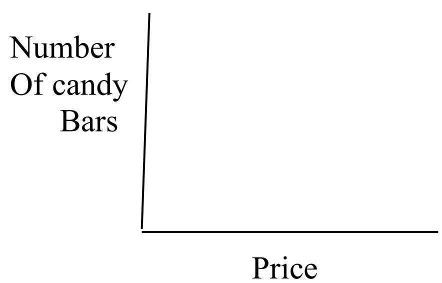
18. In the data in the table were represented in a circle graph, find the measure of the central angle used to describe the tip.

Lunch	Cost
Sandwich	\$5.00
Drink	\$1.00
Dessert	\$3.00
Tip	\$1.00

19. Draw a line graph to show the relationship between the number of hours worked and the amount of money earned.



20. A merchant found that as the price of candy bars increased, the number of sales decreased. Sketch a line graph to show that relationship.



21. You contract with your neighbor to cut and trim his lawn for a fixed fee, construct a line graph to show this relationship.