

TOPIC D3: SUMMARIES FOR DATA



SPOTLIGHT: NUTRITION VALUE OF ICE CREAM?

Ice cream is one of the favorite American desserts. Actually ice cream has a long history as it can be traced back to the 4th century B.C. The Roman emperor Nero (A.D. 36-68) had ice brought from the mountains and combined with fruit toppings and King Tang (A. D. 618-97) of Shang, China had a method of creating ice and milk concoctions. Europe was likely introduced to ice cream from China and “milk ices” were served to the Italian and French royal courts.

Ice cream was imported to the United States and was served by some famous Americans such as George Washington and Thomas Jefferson. The first ice cream parlor in America opened in New York City in 1776 and supposedly the term “ice cream” was first used by American colonists. There were great advances in the technology of making ice cream in the 19th century. Nancy Johnson in 1846 received a patent for a hand-cranked freezer that established the basic method for making ice cream still used today. The first large-scale commercial ice cream plant was started by Jacob Fussell in 1851. The ice cream cone was introduced at the 1904 St. Louis World’s Fair. There is some debate about the origin of the ice cream sundae. One possible inventor of the sundae was Chester Platt who served ice cream with cherry syrup and candied cherry at his drug store in 1893 for Reverend John Scott on a Sunday and called the concoction “Cherry Sunday.” Soft ice cream was introduced in the 20th century by a chemical research team in Britain who discovered a method of doubling the amount of air in ice cream.

The United States produces about 900 million gallons of ice cream annually. Almost one-tenth of the nation’s milk supply is used to produce ice cream and other frozen desserts. Americans eat an average of about 20 quarts of ice cream annually. According to the web site <http://www.sendicecream.com/>, ice cream consumption is

highest in July and August, most ice cream is purchased on Sunday, and Portland, Oregon purchases the most ice cream on a per capita basis.

In an article published by the Center of Science in the Public Interest (CSPI) in July/August 2003, there is a general concern about the excessive fat and calories contained in ice cream desserts served by major restaurants. The daily recommended number of calories for a male adult, age 25-50, is 2900 and a day's allowance for saturated fat is 20 grams. The CSPI article shows that many of the desserts served by Ben and Jerry's, TCBY, Baskin-Robbins, and Haagen-Dazs contain over 1000 calories and over 20 grams of saturated fat. One problem is that customers often don't know what they are getting, since these stores don't give nutrition labels to their products. "With ice cream portions like these, it is no wonder that two out of three Americans are overweight, diabetes rates are rising, and heart disease is the leading cause of death", says Marion Nestle, chair of the nutrition and food studies department at New York University.

PREVIEW

In the last topic, we were introduced to the notion of a data distribution and learned about different graphs useful for seeing particular characteristics of the distribution such as shape, average, and spread. In this topic, we first describe useful ways of summarizing categorical data and then we will be introduced to some basic methods for describing the average and spread for a group of quantitative data.

In this topic, the learning objectives are to:

- Understand how to summarize categorical data by computing percentages and a mode.
- Understand how to interpret two popular measures of center, the median and the mean, and understand when these two measures will be the same or different.
- Understand how to measure spread by the quartiles and by a typical size of a deviation from the mean.
- Relate a graph of a data distribution with the measures of center and spread.



NCTM Standards

- ✓ In Grades 6-8, all students should find, use, and interpret measures of center and spread, including mean and interquartile range.
- ✓ In Grades 9-12, all students should recognize how linear transformations of univariate data affect shape, center, and spread.

SUMMARIZING CATEGORICAL DATA

In this topic we focus on some numbers that are useful for summarizing a collection of data. First, let's consider data of the categorical type. In Topic D2, we looked at the country of birth of all professional baseball players who were born in the year 1975. We organized the data by use of the following frequency table.

Country of origin	Frequency
USA	135
Latin America	62
Other	8

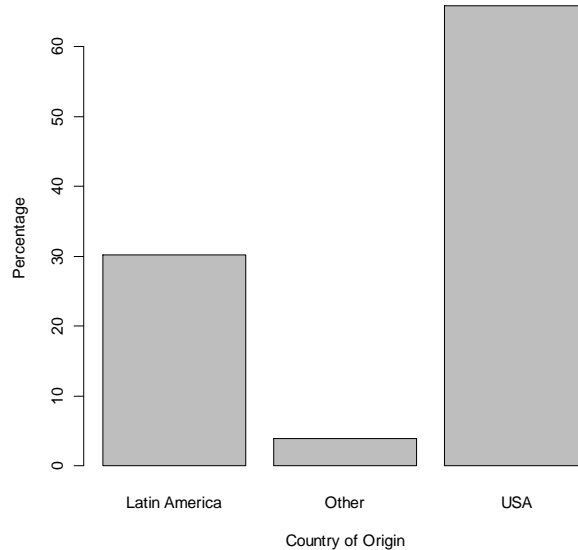
How do we summarize these data? First, to understand the relative sizes of the frequencies, it is helpful to compute the proportions of the categories – we find these by dividing each frequency by the total count (205). So, for example, the proportion of USA ballplayers is $135/205 = .659$ and the proportion of Latin American players is $62/205 = .302$. We convert these proportions to percentages in the table by multiplying by 100. If we multiply this proportion of American players (.659) by 100 and round to the nearest integer, we get the percentage of American ballplayers to be 66.

Country of origin	Frequency	Proportion	Percentage
-------------------	-----------	------------	------------

USA	135	0.659	66
Latin America	62	0.302	30
Other	8	0.039	4
TOTAL	205	1.000	100

For categorical data, a useful summary is the *mode*, the category with the highest frequency or percentage. Here the mode of the country of origin is USA. Also it is helpful to describe other categories that have large percentages. For this example, we could say that approximately 66 percent of these baseball players are American, but a sizeable (30%) percentage of the players are from Latin America. The remaining countries make up only a small percentage of the players.

In topic D2, we introduced a bar chart where we graphed the category frequencies on the vertical scale against the category names on the horizontal scale. One variation of this display is to plot the category proportions or the category percentages on the vertical scale. The figure below shows a bar chart of the category percentages.



The frequency and percentage versions of the bar chart have the same visual appearance. But perhaps the percentage bar chart is more useful since one can read the category percentages directly from the graph.

PRACTICE: SUMMARIZING CATEGORICAL DATA

In the Practice Graphing Categorical Data section of Topic D2, we collected several variables for 50 used sedan cars.

1. We classified the cars into four groups by the manufacturer (General Motors, Ford, Chrysler, and Foreign). Copy the frequencies of the four groups from your work in Topic D2. Compute the proportion and percentage of each group.

Manufacturer	Frequency	Proportion	Percentage
General Motors			
Ford			
Chrysler			
Foreign			

2. Find the mode of the manufacturer for these 50 used cars.
3. Cars in this group were also classified by mileage (low, medium, and high). Construct a frequency table for the mileage below including the proportions and percentages.

Mileage	Frequency	Proportion	Percentage
Low			
Medium			
High			

4. Construct a bar chart for mileage using percentage as the variable on the vertical axis.
5. Find the mode of the mileage for these used cars.

HOW MANY CALORIES ARE IN AN “AVERAGE” SCOOP OF ICE CREAM? (INTRODUCING THE MEDIAN)

On the Ben and Jerry Ice Cream website <http://www.benjerry.com/>, one finds a list of many of their ice cream flavors and the number of calories contained in a single serving of each flavor. Here is a listing for 13 flavors:

Flavor	Calories/serving
Aloha Macadamia	330
Apple Crumble	280
Bovinity Divinity	290
Cherry Garcia™	260
Chocolate Chip Cookie Dough	300
Chocolate Fudge Brownie	280
Chubby Hubby™	350
Chunky Monkey™	310
Coffee Heath™ Bar Crunch	310
Concession Obsession	300
Festivus (limited edition)	300
Island Paradise	240
Kaberry Kaboom	240

A first step in understanding the variation in these calorie measurements is to construct a stemplot.

```
24 | 00
25 |
26 | 0
27 |
28 | 00
29 | 0
30 | 000
31 | 00
32 |
33 | 0
34 |
35 | 0
```

24|0 means 240 calories

We see a center cluster of calorie measurements in the 280 to 310 range and values ranging from 240 to 350.

We would like to summarize this distribution of calorie numbers with a single "average." There are a couple different averages that are commonly reported in the media. The first one we will discuss is the **median** (denoted by M) that is the *middle value* when the data values are arranged in ascending order.

We find the median in two steps:

- We find the position or location of the median in the list when the measurements are arranged in increasing order.
- The median is the data value that has the middle position.

What is the position of the median? We first arrange the calorie measurements in increasing order. We assign position 1 to the smallest value, position 2 to the next smallest value, etc.

Calories	240	240	260	280	280	290	300	300	300	310	310	330	350
Position	1	2	3	4	5	6	7	8	9	10	11	12	13

We can see that the middle position of the 13 measurements is 7. The median is the data value in the 7th position, which we see is $M = 300$.

Suppose, instead, that you have 10 measurements, like the first 10 calorie measurements, 330, 280, 290, 260, 300, 280, 350, 310, 310, 300. We list these measurements in ascending order:

Calories	240	240	260	280	280	290	300	300	300	310
Position	1	2	3	4	5	6	7	8	9	10

In this case (with an even number of measurements), there is not one middle measurement. In this case we define the position to be the average of the middle two positions $(5+6)/2 = 5.5$ and the median is defined to be the average of the measurements with these two middle positions. So $M = (280 + 290)/2 = 280.5$.

Generally if n is the number of values in the dataset, then the position of the median is given by

$$pos(M) = \frac{n+1}{2}.$$

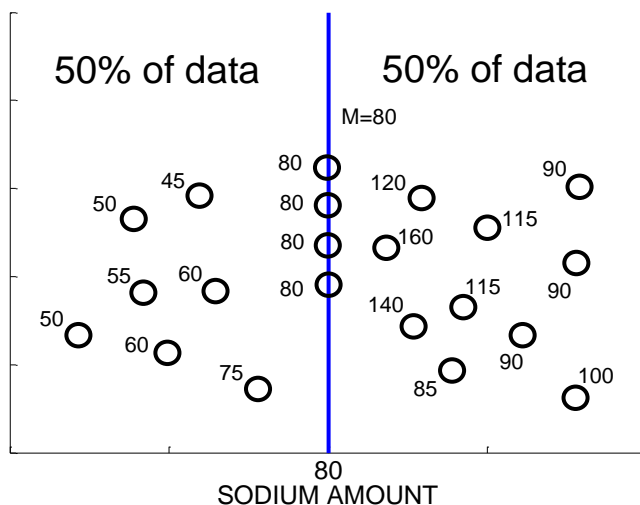
- For our first example of 13 measurements, $n = 13$ and $\text{pos}(M) = (13+1)/2 = 7$.
- For our second example of 10 measurements, $n = 10$ and $\text{pos}(M) = (10+1)/2 = 5.5$

We illustrate the computation of the median for each of the data distributions (calories and sodium amounts for a collection of ice cream flavors) displayed using stemplots. For each distribution, we show the number of measurements n , the position of the median $\text{pos}(M)$, and the value of the median M .

Calories of 28 flavors	Sodium amounts for 21 flavors
24 000	4 5
25	5 005
26 00	6 00
27 00000	7 5
28 0000000	8 00005
29 000	9 000
30 000	10 0
31 00	11 55
32 0	12 0
33 0	13
34	14 0
35	15
36	16 0
37	
38 0	
24 0 means 240 calories	4 5 means 45 grams of sodium
$n = 28$	$n = 21$
$\text{pos}(M) = (28+1)/2 = 14.5$	$\text{pos}(M) = (21+1)/2 = 11$
$M = (280+280)/2 = 280$ calories	$M = 80$ grams

The median has a simple interpretation as a center of a dataset. The value of M divides the data into two groups of the same size. So we can say (approximately) that half of the measurements are smaller than M and half are larger than M . For the sodium amounts, approximately half of the values are smaller than 80 grams. We illustrate the interpretation of the median by the graph below. The individual observations are represented by circles; the median $M = 80$ can be thought as a dividing line between the “low” sodium amounts and the “high” sodium amounts. When we calculate a median,

the actual sodium measurements are not important – it only matters if the sodium amount is below or above the median M .



PRACTICE: COMPUTING A MEDIAN

Consider again the daily high temperatures (in degrees Fahrenheit) for Atlanta for the month of March 2005. A stemplot of these daily temperatures is displayed below.

```

4 | 01
4 | 78
5 | 04
5 | 555889
6 | 01124
6 | 5566777
7 | 04
7 | 5679
8 | 0

4|0 means 40 degrees F
    
```

1. Find the sample size and the position of the median.
2. Find the median M .
3. Based on our work, we can say that (approximately) half of the daily high temperatures are larger than _____.

4. Suppose by mistake the largest temperature 80 degrees really should have been 90 degrees. Calculate the median of the new dataset. Did the value of the median change? If so, by how much?

DEVIATIONS AND THE MEAN

The second measure of center, the **mean** (usually denoted by \bar{x}) is the value you get when you sum all of the values and divide by the number of values. For our collection of 13 calories of ice cream flavors, the mean is equal to

$$\bar{x} = \frac{330 + 280 + 290 + 260 + 300 + 280 + 350 + 310 + 310 + 300 + 300 + 240 + 240}{13} = 291.54$$

We saw that the median had a simple interpretation -- what's the interpretation of the mean?

To help understand the mean, we introduce a new idea -- a *deviation*. The deviation of a data value from a given number, say c , is the difference of that data value from c :

$$\text{deviation} = \text{data value} - c.$$

For our original 13 ice cream calorie numbers, suppose that $c = 300$; this number can represent our guess at a typical calorie number. To obtain the deviations, we subtract 300 from each data value.

Calories	Deviation = Calories - 300
330	$330 - 300 = 30$
280	$280 - 300 = -20$
290	$290 - 300 = -10$
260	$260 - 300 = -40$
300	$300 - 300 = 0$

280	$280 - 300 = -20$
350	$350 - 300 = 50$
310	$310 - 300 = 10$
310	$310 - 300 = 10$
300	$300 - 300 = 0$
300	$300 - 300 = 0$
240	$240 - 300 = -60$
240	$240 - 300 = -60$

How good is our guess of 300 as a typical number of calories? We can judge the goodness of this guess by adding up all of the deviations – we call this the *sum of the deviations*.

$$\text{SUM OF DEVIATIONS} = 30 + (-20) + (-10) + \dots + (-60) + (-60) = -110.$$

Here the sum of the deviations is negative (-110) – this means that our guess of 300 as a typical number of calories is a bit high.

One way of defining a “best” measure of center is to find the value of c such that the sum of deviations about c is equal to zero. By trial and error, we could try different values of c , compute the sum of deviations about that value, and then find the value of c that makes the sum of deviations equal to zero. But fortunately we don’t have to do this work -- one can show mathematically that the sum of deviations about the mean $\bar{x} = 291.54$ is equal to zero.

Another way of stating this property is that the sum of positive deviations about the mean (that is, the deviations of the data values above the mean) will be equal to the sum of negative deviations about the mean (corresponding to the data values below the mean).

PRACTICE: DEVIATIONS AND THE MEAN

1. The below table contains the high temperatures of Atlanta, Georgia in the first thirteen days in May 2005. In the below table, suppose that your guess at the “average” is 50

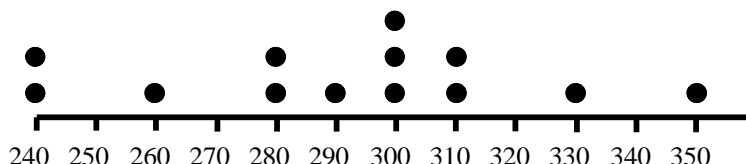
degrees. Find the deviations (in the first empty column of the table) and the sum of the deviations.


Temperature	Deviation = Temperature – 50	Deviation = Temperature - 55
41		
48		
55		
61		
66		
61		
67		
55		
47		
54		
41		
48		
55		

- Now suppose that your guess at the average is 55 degrees. Find the deviations (in the last column of the table) and sum of deviations for this guess.
- Based on your work from parts 1 and 2, do you think that 50 or 55 is closer to the mean \bar{x} ?
- Compute the mean \bar{x} . Was your choice in part 3 correct?

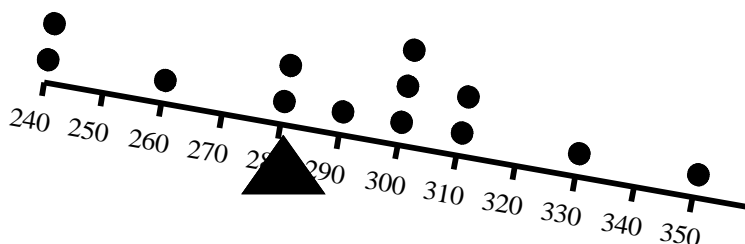
GEOMETRICAL INTERPRETATION OF THE MEAN

There is a geometrical interpretation of this result about deviations and the mean. First we draw a dotplot graph of the data.

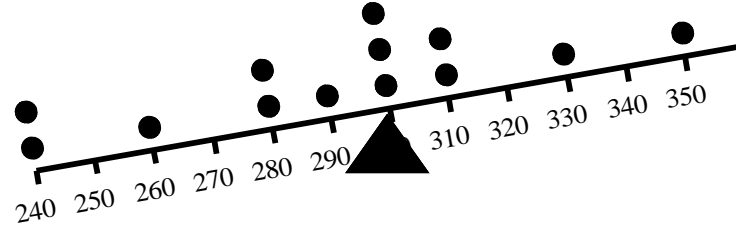


Suppose that we placed a stiff board under the graph and each dot on the graph represents a weight of a given amount. We place a fulcrum  under the board and try to balance the weights.

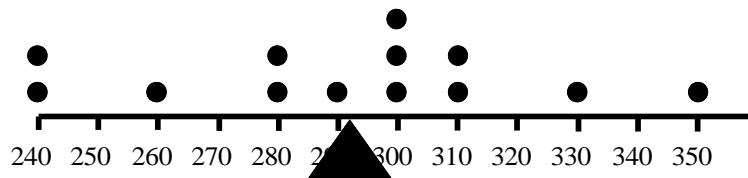
If we try to place the fulcrum at the value 280, the dotplot is unbalanced -- there is too much weight on the right. Here the positive deviations about 280 outweigh the negative deviations about 280.



We slide the fulcrum to the right at 300 -- now the dotplot is unbalanced with too much weight on the left. Here the negative deviations are greater than the positive deviations.



After some more moving of the fulcrum, we find a spot so that the board is balanced. The sum of the positive deviations is equal to the sum of the negative deviations.



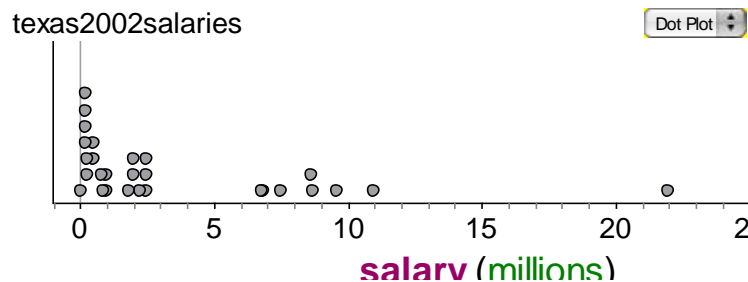
The location of the fulcrum will be at the mean value $\bar{x} = 291.54$.

COMPARING THE MEDIAN AND THE MEAN

In the first example, the median and mean of the group of 13 calorie measurements were approximately the same -- will that usually be the case?

Actually, no. In many situations, the mean and median will be different and one should understand why.

Suppose we look at the salaries of the players on the 2002 Texas Rangers baseball team. The salaries have been graphed below using a dotplot.



Note that there is a cluster of low values, a second cluster of salaries about 10,000,000 (10 million dollars) and a single large value (that corresponds to Alex Rodriguez who was making 22 million dollars that particular season).

Here the median and mean salaries are very different. The median salary is $M = \$2,000,000$ and the mean is $\bar{x} = \$3,634,728$, so these two measures are over 1.5 million dollars apart.

Here are some comments about the interpretation of these measures and why the mean and median are different.

1. First, the median is easy to interpret. We can say that approximately half of the Rangers have salaries larger than 2 million dollars and half have salaries smaller than 2 million.
2. The mean is larger than the median since there are a number of large salaries that make the mean larger. When the dataset is skewed to the right, the mean will be larger than the median.
3. What is a better measure of average -- the mean or the median? The answer depends on what type of average you are interested in. If you are interested in a typical or representative salary, then the median is the better average, since many players have salaries close to 2 million dollars. The median is a reasonably typical salary. On the other hand, suppose the owner is concerned about the total amount of money that he or she is spending on payroll. The mean is a better measure for this owner since it includes all of the players' salaries and one can compute the total payroll from the mean. Here the mean is equal to $\bar{x} = \$3,634,728$ and there are 29 players – so the total payroll is

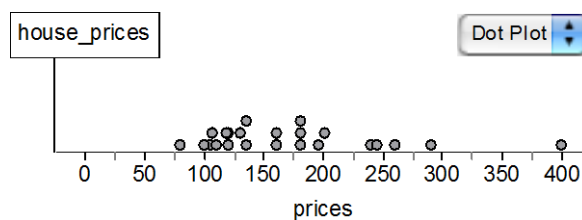
$$\text{PAYROLL} = 3,634,728 \times 29 = \$105,407,112.$$

The owner is paying over 105 million dollars for the salaries of his players.

PRACTICE: COMPARING THE MEDIAN AND THE MEAN

Recently the author collected the sale prices for some houses that were open for viewing on a particular weekend. Here are the prices (in thousands of dollars) and a graph of the data.

195 135 104 399 107 120 240 130 180 180 180 100 160
245 200 259 160 80 110 120 290 118 135



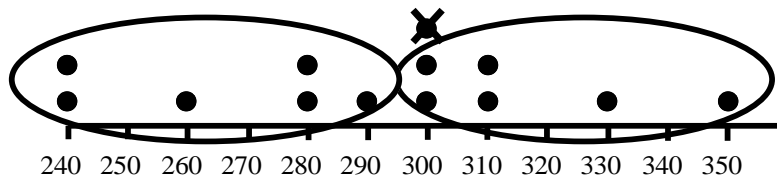
1. Write a short paragraph about these data.
2. Based on your description in part (a), do you believe that the mean and median would be approximately equal, or do you think that one measure would be larger than the other? Why?
3. Compute the mean and median for these data.
4. Would it be possible to change the value of a single observation, so that the mean and median for the new data would be approximately equal? Explain.

MEASURES OF SPREAD: QUARTILES AND IQR

In this first half of this topic, we talked about ways of measuring an average value in a dataset. Here we focus on ways of describing the spread or variation in a dataset of measurement data. To define our first measure of spread, the IQR, we need to define additional division points in our data.

Let's return to our Ben and Jerry's ice cream example, where the number of calories in a single serving of 13 flavors of ice cream was recorded. We earlier found the median and the mean for these data. How can we measure the spread in the calorie values that we saw in the dotplot?

In this example we have 13 data values. To define a measure, we wish to divide the dataset into two halves. In a case such as this where the number of observations is odd, we discard the middle measurement (the median) indicated below with an X. Then the measurements can be divided equally into a lower half and an upper half that are circled below.



We find the median of the lower half and the median of the upper half of measurements. The median of the lower half of measurements is the median of {240, 240, 260, 280, 280, 290} which is equal to 270 calories. The median of the upper half is 310 calories.

These new dividing points are called *quartiles* -- the lower quartile Q_L is the value such that (approximately) one quarter of the data is smaller than Q_L , and the upper quartile Q_U is the value such that one quarter of the data is larger than Q_U . Here $Q_L = 270$ and $Q_U = 310$. By reporting the quartiles Q_L and Q_U , we get some idea about the spread in the data.

Unfortunately, there is not a universally accepted definition of a quartile -- you will find different definitions in different statistics books. Here is our definition of a quartile that is easy to learn.

1. After arranging the data in order, divide into two halves. If we have an even number of data values, we can make a clean break into halves. If there is an odd number of data values (as in the above example), we discard the median and divide the remaining observations into halves.
2. The lower and upper quartiles, Q_L and Q_U , are the medians of the lower half and the upper half of the data, respectively.

A useful summary of a group of measurement data are the five numbers

$$(LO, Q_L, M, Q_U, HI),$$

where LO and HI are respectively the lowest and highest values in the dataset -- we call this the *five-number summary*.

To illustrate, we find the five-number summary for the following prices of houses (in thousands of dollars)

```

195   135   104   399   107   120   240   130   180   180   180   100   160
245   200   259   160    80   110   120   290   118   135

```

By graphing these data using a stemplot with ascending leaves, we have an ordered arrangement of the data. (Note that we have placed the unusually high price of 399 thousand dollars on a separate line so that the display isn't too long.) We show the calculation of the five-number summary below.

```

 8 | 0          n = 23
 9 |           pos(M) = (23+1)/2 = 12
10 | 047       M = 160
11 | 08
12 | 00
13 | 055       Since there are 23 observations, we discard the median and
14 |           break data into two groups of 11.
15 |
16 | 00       QL is median of lower half
17 |         pos(QL) = (11+1)/2 = 6
18 | 000       QL = 118
19 | 5
20 | 0       QU is median of upper half
21 |         pos(QU) = (11+1)/2 = 6
22 |         QU = 200
23 |
24 | 05
25 | 9
26 |         five-number summary is
27 |         (80, 118, 160, 200, 399)
28 |
29 | 0

HI 399

8|0 means 80 thousand dollars

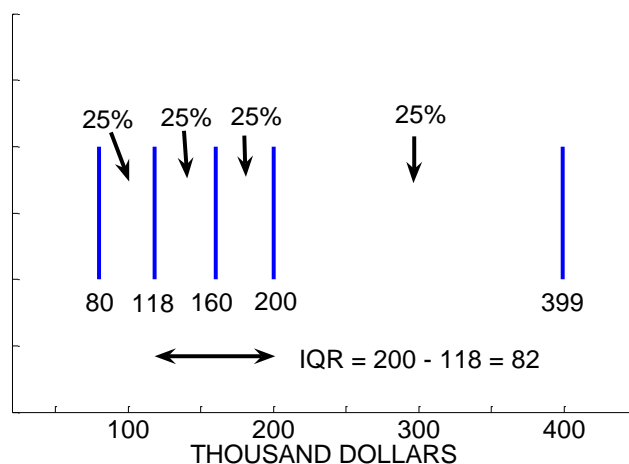
```

A simple measure of spread in a dataset is the difference between the upper and lower quartiles -- we call this the *interquartile range* (IQR):

$$IQR = Q_U - Q_L.$$

The IQR is the spread of the middle 50% of the data.

We graphically show the location of the five-number summary below. We see that the numbers divide the house prices into four parts where 25% of the data is between the low value and Q_L (80 and 118 hundred thousand dollars), 25% falls between Q_L and M (118 and 160), and so on. Here the interquartile range is equal to $IQR = Q_U - Q_L = 200 - 118 = 82$. This means that the spread of the middle half of the house prices is 82 thousand dollars.



PRACTICE: COMPUTING A FIVE-NUMBER SUMMARY

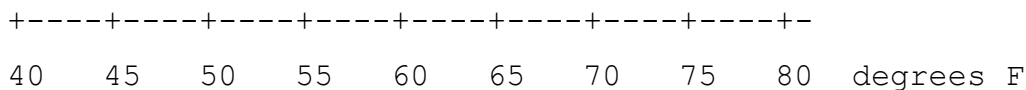
Consider again the daily high temperatures for Atlanta for March of 2005. A stemplot of the temperatures is shown below.

```

4 | 01
4 | 78
5 | 04
5 | 555889
6 | 01124
6 | 5566777
7 | 04
7 | 5679
8 | 0
    
```

4|0 means 40 degrees

1. Compute the five-number summary.
2. Compute the interquartile range IQR.
3. On the number line below, mark using X's the locations of LO, Q_L , M, Q_U , and HI.



4. In words, describe what the IQR is telling you about the spread in the daily temperatures.
5. Find an interval that contains the lowest 25% of the temperatures.

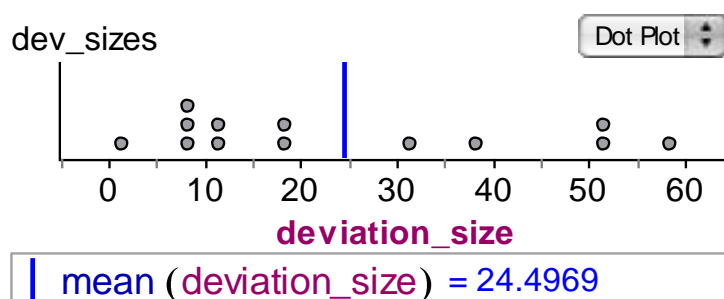
MEASURES OF SPREAD USING DEVIATIONS

Alternative measures of spread can be defined on the basis of the deviations. Recall a deviation is the difference of a data value from a particular number c . Suppose we consider the deviations of each value from the mean \bar{x} . In our ice cream calorie example, we show for each flavor, the calories (in a single scoop), the mean $\bar{x} = 291.54$ and the deviation = calories $- \bar{x}$. The deviation for the first value, 330, is deviation = $330 - 291.54 = 38.46$, the deviation for the second value, 280, is deviation = $280 - 291.54 = -11.54$, and so on.

Flavor	Calories	Mean	Deviation	Absolute Deviation
Aloha Macadamia	330	291.54	38.46	38.46
Apple Crumble	280	291.54	-11.54	11.54
Bovinity Divinity	290	291.54	-1.54	1.54
Cherry Garcia™	260	291.54	-31.54	31.54
Chocolate Chip Cookie Dough	300	291.54	8.46	8.46
Chocolate Fudge Brownie	280	291.54	-11.54	11.54

Chubby Hubby™	350	291.54	58.46	58.46
Chunky Monkey™	310	291.54	18.46	18.46
Coffee Heath™ Bar Crunch	310	291.54	18.46	18.46
Concession Obsession	300	291.54	8.46	8.46
Festivus (limited edition)	300	291.54	8.46	8.46
Island Paradise	240	291.54	-51.54	51.54
Kaberry Kaboom	240	291.54	-51.54	51.54

A natural measure of spread of a dataset is the average or mean *size* of these deviations. A *size* of a deviation is simply its absolute value. We graph these deviation sizes using a dotplot:



Note that deviation sizes close to zero correspond to data values close to the mean and large deviation sizes (such as the three deviation sizes larger than 50 calories) correspond to calorie numbers that are far from the mean.

What is a typical deviation of a calorie number from its mean? A natural summary of deviation size is the Mean Absolute Deviation (or MAD for short) that is the mean of the absolute values of the deviations:

$$MAD = \frac{|330 - 291.54| + |280 - 291.54| + \cdots + |240 - 291.54|}{13} = 24.4969.$$

In the table, we showed the absolute deviations in the rightmost column and found the mean of these values to be 24.50. So $MAD = 24.50$ – we can say that 24.5 calories is a

typical or representative size of a deviation from the mean. In other words, on average, calories are 24.5 units from the mean. The above graph shows that the MAD is a reasonable measure of center of the deviation sizes.

A second measure of spread based on the deviations is the well-known *standard deviation*, which we abbreviate by s . This measure is based on the *squared deviations* instead of the deviation sizes. If we *square* each of the deviations, find the *sum* of the squared deviations, then the standard deviation, s , is defined to be

$$s = \sqrt{\frac{\text{sum of squared deviations}}{n-1}},$$

where n is the number of items in the dataset.

The table below illustrates the work in computing the standard deviation. Remembering that there are $n = 13$ observations, we find that

$$s = \sqrt{\frac{1479.17 + 133.17 + \cdots + 2556.37}{13-1}} = \sqrt{\frac{12369.21}{12}} = 32.11.$$

Flavor	Calories	Mean	Deviation	Squared Deviation
Aloha Macadamia	330	291.54	38.46	1479.17
Apple Crumble	280	291.54	-11.54	133.17
Bovinity Divinity	290	291.54	-1.54	2.37
Cherry Garcia™	260	291.54	-31.54	994.77
Chocolate Chip Cookie Dough	300	291.54	8.46	71.57
Chocolate Fudge Brownie	280	291.54	-11.54	133.17
Chubby Hubby™	350	291.54	58.46	3417.57
Chunky Monkey™	310	291.54	18.46	340.77
Coffee Heath™ Bar Crunch	310	291.54	18.46	340.77
Concession Obsession	300	291.54	8.46	71.57

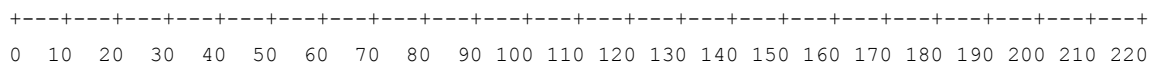
Festivus (limited edition)	300	291.54	8.46	71.57
Island Paradise	240	291.54	-51.54	2656.37
Kaberry Kaboom	240	291.54	-51.54	2656.37
				Sum of Squared Deviations = 12369.21

For these calorie numbers, we have computed two measures of spread based on the deviations $MAD = 24.50$ and $s = 32.11$. Both measures represent typical distances of the data from the mean $\bar{x} = 291.54$ calories. Which is a better measure? The MAD is probably easier to interpret since it is a simple function of the sizes of the deviations. We will see next that the standard deviation has a nice interpretation when we have a set of data whose distribution is approximately bell shaped.

PRACTICE: MEASURES OF SPREAD USING DEVIATIONS

The following table lists the house prices (in thousands of dollars) for homes that were on sale and available for viewing in a recent weekend in the hometown of the author. The mean price of these homes is $\bar{x} = 171.6$. The table also lists the deviations from the mean and the squared deviations.

1. Compute the absolute deviations in the table. Construct a dotplot of these absolute deviations on the following grid.



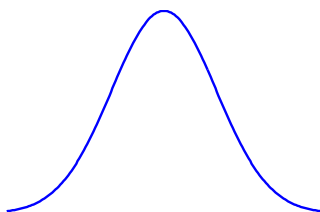
2. Find a house price that is close to the mean. Will this price have a small or large deviation size?
3. If a house price has a large deviation size, is this price close or far away from the mean?
4. Compute the mean absolute deviation. Interpret what this number tells you about the spread of the house prices.

5. Complete the empty cells of the table and compute the standard deviation of the house prices s . Is the value of s close to the value of the MAD? Explain why they should be similar in size.

Price in thousands of \$	Deviation	Absolute Deviation	Squared Deviation	Price	Deviation	Absolute Deviation	Squared Deviation
195	23.4		547.56	160	-11.6		134.56
135	-36.6		1339.56	245			
104	-67.6		4569.76	200			
399	227.4		51710.76	259	87.4		7638.76
107				160			
120	-51.6		2662.56	80	-91.6		8390.56
240				110	-61.6		3794.56
130	-41.6		1730.56	120	-51.6		2662.56
180	8.4		70.56	290	118.4		14018.56
180	8.4		70.56	118	-53.6		2872.96
180	8.4		70.56	135	-36.6		1339.56
100	-71.6		5126.56				

INTERPRETING S: THE 68/95/99.7 RULE FOR BELL-SHAPED DATA

Unlike the IQR, there is no general simple interpretation for the standard deviation s . However, if the data have an approximate symmetric bell or bell shape, shown here,



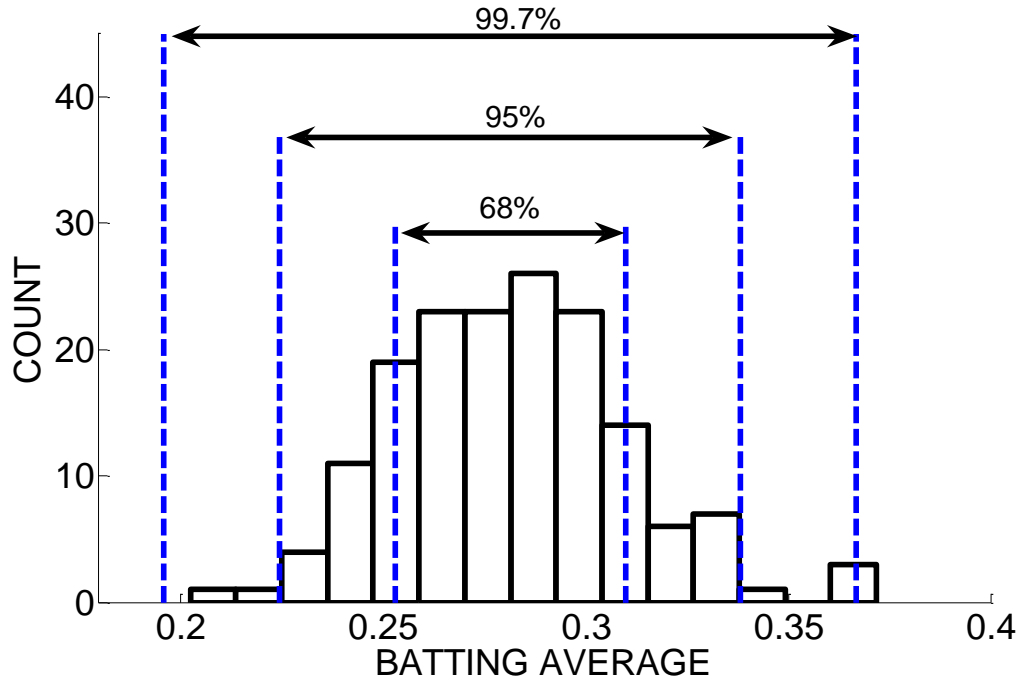
then there is a nice interpretation for s . If the data are bell-shaped, then we expect

- about 68% of the data will fall in the interval $(\bar{x} - s, \bar{x} + s)$
- about 95% of the data will fall in the interval $(\bar{x} - 2s, \bar{x} + 2s)$
- about 99.7% of the data will fall in the interval $(\bar{x} - 3s, \bar{x} + 3s)$

An example of a dataset that is bell-shaped is the collection of batting averages for all baseball players who are “regular” players during a particular baseball season. The figure below displays a histogram of the batting averages for the 162 players. We can compute the mean and standard deviation of these batting averages to be .281 and .028, respectively. Then we expect

- 68% of these batting averages to fall between $.281 - .028$ and $.281 + .028 = .253$ and .309.
- 95% of these batting averages to fall between $.281 - 2 \times .028$ and $.281 + 2 \times .028 = .225$ and .337.
- 99.7% of these batting averages to fall between $.281 - 3 \times .028$ and $.281 + 3 \times .028 = .197$ and .365.

These intervals are shown on the figure together with the expected percentages.



We can check the validity of the rule in this example by actually counting how many batting averages fall in the above intervals. Looking at the data, the table gives the number and percentage of values in each interval:

Interval	Number of batting intervals in interval	Percentage in interval
(.253, .309)	109	$109/162 \times = 67.3$
(.225, .337)	156	$156/162 \times = 96.3$
(.197, .365)	160	$160/162 \times = 98.8$

We see that the interval percentages match up pretty well with the expected percentages 68, 95, and 99.7. This is expected since the batting averages have a distribution that is close to a bell-shape.

PRACTICE: THE 68/95/99.7 RULE

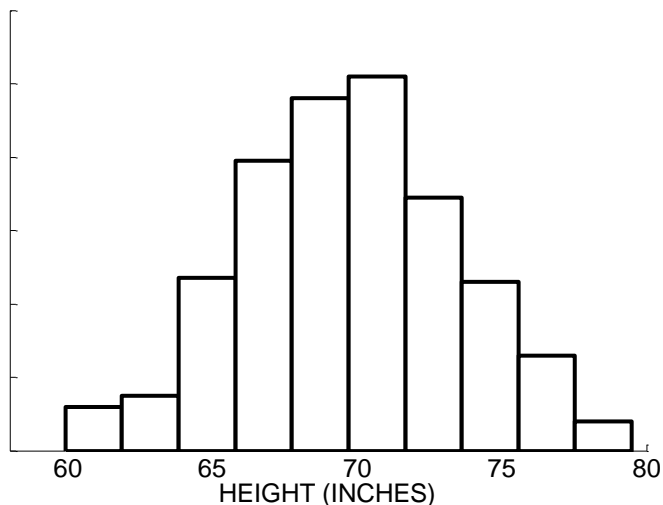
Below we apply this rule to our house prices dataset.

1. Does the dataset have a bell-shaped distribution? If not, describe the shape of its distribution.
2. The mean and standard deviation of these 23 house prices are 171.6 and 75.1, respectively. Find the interval $(\bar{x} - s, \bar{x} + s)$.
3. Of the 23 house prices, how many prices fall in the interval you found in part 2?
4. What percentage of the house prices fall in the interval? Is this percentage close to the expected percentage 68?
5. Find the interval $(\bar{x} - 2s, \bar{x} + 2s)$ and find the percentage of prices that fall in this interval. Is this percentage close to 95?
6. Find the interval $(\bar{x} - 3s, \bar{x} + 3s)$ and find the percentage of prices that fall in this interval. Is this percentage close to 99.7?
7. You may find that the percentages you found in parts 4, 5, and 6 aren't close to the values 68, 95, and 99.7 in our rule. Relating back to your answer to part 1, can you suggest a reason why the rule may not work well in this particular situation?

SPECIAL NOTE: When the data is approximately bell-shaped, then one can estimate the value of the standard deviation directly from a graph such as a histogram. To illustrate, consider the following graph of the heights (in inches) of a group of 500 men that appears to be bell-shaped. To estimate the standard deviation, we find from the graph an interval that contains approximately 95% of the heights. From the 68/95/99.7 rule, we expect 95% of the data to fall in the interval $(\bar{x} - 2s, \bar{x} + 2s)$. By equating the width of this interval ($4s$) to our estimate, we can obtain an approximate value of s .

Here the interval (61, 78) seems to contain most of the heights; this interval has a width of $78 - 61 = 17$. So $4s = 17$ and solving for s , we get $s = 17/4 = 4.25$. To see if

this is a reasonable answer, we compute the actual value of s to be 3.8 which is close to our estimate.



ACTIVITY: COLLECTING SOME DATA ON CITIES

1. Go to www.cityrating.com (click on City Guides or Weather History links).
2. For each of 20 cities of your choice (choose cities across a broad range of the U.S.), collect two quantitative variables of your choice (some possibilities might be population, percentage of women, median age, average temperature during a particular month, precipitation, etc.). Put your data in the table below.

My Variable 1 was _____; my Variable 2 was _____

Units of Variable 1 _____; units of Variable 2 _____

Number	CITY	Variable 1	Variable 2
1			
2			

3			
4			
5			
6			
7			
8			
9			
10			
11			
12			
13			
14			
15			
16			
17			
18			
19			
20			

3. For each variable ...

- draw a good stemplot of the data
- find a five-number summary
- find a city that represents the “average” for the variable
- find a measure of spread
- write a descriptive paragraph about the data

ACTIVITY: V IS FOR VARIATION

DESCRIPTION: In this activity, we look at the different measures of the average deviation considering the deviations of the data values from the mean. We use a “V” measurement, specifically the distance when you make a “V” with your fingers, to emphasize the focus on measures of variation. In comparing groups, we will see that comparing measures of variation tells us something very different than comparing measures of average.

This activity works best with groups of 3-5 students.

1. Measure your “V-span” as follows: With the palm of your writing hand on a flat surface, make a “V” between the index and middle fingers. Measure the distance (in centimeters) from the outside of your index finger tip to the outside of your middle finger tip when spread as far as possible.



Record the V-span (in cm) for all members of your group below. (We'll fill in the Deviation Size and Square of Deviation columns later.)

Student	V-span (cm)	Deviation Size	Square of Deviation

2. Construct a dotplot of the V-spans of your group. Label each V-span with the student's initials. Compute the mean and show the value of the mean on the graph.
3. For each V-span measurement, compute the deviation size, defined to be the absolute value of the difference of the measurement from the mean. Place these absolute deviations in the table.
4. Construct a dotplot of the deviation sizes. This graph shows you how far on average each member's V-span falls from the mean. By looking at your graph, what is a typical deviation size?

5. There are several ways of computing an “average” deviation size. One way is to compute the mean absolute deviation, called MAD, equal to the mean of the deviation sizes. Another way is to compute the standard deviation that is found using the formula

$$s = \sqrt{\frac{\text{sum of squared deviations}}{(\text{number of measurements}) - 1}}.$$

Compute the MAD and standard deviation for your group’s measurements. (The squared deviation column in the above table can be helpful in computing the standard deviation.)

6. Suppose one group has a MAD value that is very close to zero. What does this say about the V-spans in this group?
7. Suppose Group A has four women and Group B consists of two men and two women. Which group would have a larger value of MAD -- Group A or Group B? Explain.
8. Compare the values of MAD across groups in your class. Explain why some groups have large deviations and other groups have small deviations.
9. Suppose you went to a large family gathering and measured the V-spans for all people. Now that you have the V-spans of people at this family gathering, suppose you want to compare the V-spans for the children with the V-spans for the adults. Which group (adults or children) will have the larger median V-span? Why?
10. For the same data of V-spans introduced in question 9, which group (adults or children) will have the larger MAD of the V-spans? Explain.


TECHNOLOGY LAB – DEVIATIONS, THE MEAN, AND MEASURES OF SPREAD

PART A: Deviations and the Mean

Open up a new Fathom document.

(a) The dataset baseball_ages.txt contains the ages of 20 randomly selected professional baseball players. Import this into Fathom.



(b) Define a variable m. Drag down a Slider . Double-click on this Slider to change its properties.

- Change the name of the Slider from V1 to m.
- Make the Lower value 20 and the Upper value 40.

(c) Construct a dotplot of the age variable. With the graph selected, choose the menu item Graph -> Plot Value. In the Expression for Value, type in “m.” You should see the value of m displayed on the graph.

(d) Now we will define a new Attribute called “deviation.”

- Click in the first empty box next to “age” and type in “deviation.”
- Select the deviation Attribute and choose Edit Formula from the Edit Menu.

In the Formula box, type

$$\text{age} - m$$

(For each age, the deviation will be the difference between the age and m.)

(e) We will compute the sum of the deviations by use of the Summary Table.

Drag a Summary Table from the Fathom shelf

Drag the variable “deviation” to the Summary Table – you will see the mean displayed.

Double-click on mean() – change the formula from mean() to sum().

The Summary Table shows the sum of the deviations of the ages from the value m.

(f) By moving the value of m on the Slider, find the value of m such that the sum of the deviations is equal to 0. (This is the value of m that balances the positive and negative deviations.)

Questions:

- 1.: If m is equal to 25, find the deviation for Andy Phillips.
2. If m is equal to 25, find the deviation for Fred McGriff.
3. What was the value of m such that the sum of deviations is equal to 0?
4. The value that you found in Q3 – is this a popular measure of “average”, such as the mean or median?
5. Suppose that you wish to add three players to our group of players from the list below 2004

Jim Thome, 34	Eric Milton, 29	Doug Glanville, 34
David Bell, 32	Randy Wolf, 28	Marlon Byrd, 27

so that the sum of deviations about m remains 0. Which three players can you add?

PART B: Using Deviations to Construct Two Measures of Spread

We can define different measures of spread by using the deviations about the mean.

Make sure that m is equal to the mean of the player ages.

(a) Define a new Attribute called `size_deviation`. Select this new Attribute and choose Edit Formula from the Edit Menu. In the Formula box, type `abs(deviation)`. This will give you the size of each deviation from the mean.

(b) Graph the sizes of the deviations using a dotplot. Write a short paragraph about these sizes, including comments about the smallest and largest values and a “typical” value.

Different measures of spread can be defined based on these deviation sizes.

(c) One possibility is to compute the mean of these sizes – this is called the MAD (for mean absolute deviation). Find this by dragging the Attribute size_deviation to a Summary Table. The MAD is equal to _____ .

(d) Another way of summarizing these sizes is by means of the “standard” deviation. To compute this, we square each deviation size, find the sum of these squared deviations, divide the result by the sample size minus one, and take the square root of the answer. In the Summary Table you just used, select Summary -> Add Formula and type the following formula.

$$\text{sqrt} \left(\frac{\text{sum} (\text{size_deviation}^2)}{\text{count} (\text{age}) - 1} \right)$$

The standard deviation here is equal to _____.

(e) Compare the values of MAD and the standard deviation – which is larger? Looking at your dotplot of the deviation sizes, which seems to be a better “average” of the sizes? Why is one a better average of deviation size?

ACTIVITY: MEASUREMENT BIAS

DESCRIPTION: In this activity, we are introduced to the notion of measurement bias. We get experience in making measurements where a bias is present. That means that there is a known tendency to take measurements that are too small, or too large, on average. By exploring the distribution of measurements and knowing the true value, we can measure the size of the bias.

MATERIALS NEEDED: Two strings of different lengths, where the exact length of each string is known. A set of cardboard measuring instruments.

PART A: Bias in measuring the length of strings

1. Collecting the data

Look carefully at the string (marked A) your instructor is holding out straight. Without using any measuring instrument (except your eyes), estimate the length of the string to the nearest whole inch.

LENGTH OF STRING A = _____

Your instructor will collect the estimated string lengths and give you the data for your class. You will use the data in the next part of the activity.

2. Describing the data graphically

- (a) Make at least two different plots of the data on string length.
- (b) Describe the plots of the data in terms of symmetry or skewness of the distribution; clusters and gaps that might be present and outliers that might be present, including a possible reason for the outliers.

3. Describing the data numerically

- (a) Compute the following numerical summaries of the data: mean, median, standard deviation, and interquartile range.
- (b) Which of these measures seems to provide the best description of center? Why?
- (c) Which of these measures seems to provide the best description of variability? Why?

4. Collecting and summarizing another set of data

- (a) Your instructor is now holding another string, string B. As before, estimate the length of the string to the nearest whole inch.

LENGTH OF STRING B = _____

Your instructor will collect the estimated string lengths and give you the data for your class.

- (b) Describe the data for string B using the graphical and numerical techniques you found most useful in the analysis of the data from string A.

5. Making comparisons

From your analysis of the two sets of data, decide which is the longer string. How much longer do you estimate it is?

6. Determining the bias

(a) Your instructor will provide you with the correct lengths for each string. Plot the correct values on the plots of the data made previously. What do you see?

(b) It is likely that the true value is not at the center of the data display. This discrepancy between the center of the measurements and the true value is called bias. Bias is a property of the measurement system, not of an individual person making an estimate.

Does the “system” of estimating string lengths appear to be biased? What factors might be causing the bias?

(c) What was the effect of the bias on your answer to step 5? That is, does the bias affect the accuracy of your estimate of which string is longer and by how much? Why is this the case?

PART B: Optical illusions

Optical illusions are related to bias. There are many demonstrations of optical illusions that help prove this point, and some of them, such as the one described below, are well-suited to studies of bias in a measurement process or device.

1. You will be given a cardboard measuring instrument. The goal is to make the line with arrowtails at the end equal in length to the line with arrowheads at the end. To make this measurement, you slide the line with arrowtails out until you think the lengths of the two lines match.

When you are done, turn the card over and record the length of your line to the nearest tenth of a centimeter.

LENGTH OF LINE = _____

2. Your instructor will collect the lengths of the lines from the class. Using a similar analysis as you did in the first activity, graph and summarize the batch of line lengths. Your instructor will give you the true length of the line. Determine the bias of this measurement.

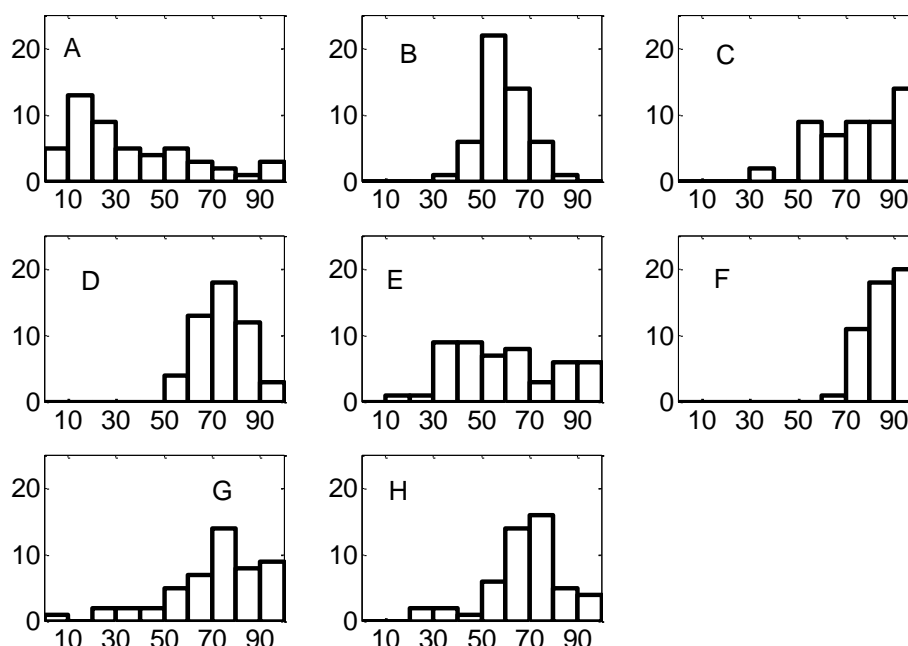
Extension

Find a printed article using data that are subject to a bias that could have a dramatic effect on the conclusions reached in the article. Summarize the conclusions in the article that are based on data, discuss possible biasing factors in the way the data were collected or analyzed, and explain how the bias in the data might affect the conclusion. (Be sure to separate bias in the data from biased reporting of the conclusions for other reasons.)

ACTIVITY: MATCHING STATISTICS WITH HISTOGRAMS

DESCRIPTION: In this activity, we will match up histograms and their corresponding summary statistics. The relative locations of the mean and median are informative about the shape of the distribution of the data. In addition, the value of the standard deviation is useful in understanding the spread of the distribution. For bell-shaped data, by the 68/95/99.7 rule, the standard deviation is helpful in understanding the proportion of data in particular intervals. Also we described how to use the 68/95/99.7 rule to estimate the standard deviation directly from a histogram.

Below are the histograms for eight datasets and following are summary statistics (the mean, median, and standard deviation) for eight datasets.



In the “Histgm” row of the below table, write down the letter of the matching histogram.

Dataset	1	2	3	4	5	6	7	8
mean	59.44	60.58	75.00	69.02	86.68	77.12	36.48	71.14
median	60.00	56.00	76.00	70.50	86.00	79.50	30.00	75.00
st dev	9.46	20.79	10.58	16.64	8.42	17.47	26.46	20.74

Histgm



Classroom Capsule: Summarizing Risky Behavior

Overview: This activity introduces the main ways of summarizing a dataset by an “average value” and a measure of spread. This particular dataset illustrates a situation where the measures of center can be different. Two measures of spread are described; one that is based on the quartiles and a second that is based on deviations from the mean.

Objectives: The students will learn the basic principles in the computation of the mean, median, IQR, and the standard deviation. The mean and median can be different in value, and the students will understand why the two measures can be different. The

students will see how one can interpret a five number summary. Also the students will gain some intuition on the standard deviation s by looking at the distribution of the sizes of the deviations.

Description:

Part 1: Measures of Center

The following table gives the percentage of young people, aged 11, 13, 15, who have been drunk two or more times for a group of 29 countries in the UNICEF study. This variable that we will call “risk” is one measure of risk of children who are living in a particular country.

Country	Pct.	Country	Pct.
Austria	15.1	Portugal	12.6
Belgium	14.5	Spain	10.2
Canada	19.8	Sweden	16.1
Czech_Republic	14.7	Switzerland	13.6
Denmark	20.1	United_Kingdom	30.8
Finland	24.7	United_States	11.6
France	8	Croatia	13.6
Germany	17.7	Estonia	23.9
Greece	10	Israel	9.3
Hungary	16.4	Latvia	16.5
Ireland	13.8	Lithuania	24.7
Italy	9.7	Malta	10.7
Netherlands	12.9	Russian_Federation	19.4
Norway	15.6	Slovenia	18.2
Poland	15.2		

1. Construct a dotplot of risk. Describe the basic shape of this data. Are there any particular countries that stand out with high or low values of risk?
2. Order the countries from smallest to largest value of risk; write below the Country and the corresponding risk percentage. (The first two countries with the smallest two values are written to start your work.)

Country	Pct.
France	8

Israel 9.3

3. Find the middle value of Pct M. This is the median that divides the risk values into a lower half and an upper half. What country has this median value?
4. Find the sum of the risk values $\sum x$ and use this sum to find the mean Pct \bar{x} .
5. Compare the values of the median and the mean. Which value is larger? Looking at the shape of the data distribution, what characteristics of the distribution would cause the two measures of center to be different? Explain.
6. There is one notable outlier in this distribution – what country has this outlier?
7. Suppose we remove this one outlier from the distribution. Recompute the values of the median and the mean. (To recompute the mean quickly, first compute the value of the new sum $\sum x$.)
8. Which measure of center, the median or the mean, is most changed with the removal of this outlier?
9. We say that a summary measure is resistant if it is not greatly changed with the inclusion of an outlier. Which measure of center, the median or the mean, is a resistant measure?

Part II: Measures of spread

All of the countries are listed in order from the smallest value of risk to the largest value of risk. The value of the median is marked with an M.

Country	Pct.	Deviation
France	8	
Israel	9.3	
Italy	9.7	
Greece	10	
Spain	10.2	
Malta	10.7	
United_States	11.6	
Portugal	12.6	
Netherlands	12.9	

Switzerland	13.6	
Croatia	13.6	
Ireland	13.8	
Belgium	14.5	
Czech_Republic	14.7	
Austria	15.1	M
Poland	15.2	
Norway	15.6	
Sweden	16.1	
Hungary	16.4	
Latvia	16.5	
Germany	17.7	
Slovenia	18.2	
Russian_Federation	19.4	
Canada	19.8	
Denmark	20.1	
Estonia	23.9	
Finland	24.7	
Lithuania	24.7	
United_Kingdom	30.8	

1. To start thinking about measuring the spread of this dataset, suppose you divide the risk values into a lower half (the values smaller than the median M) and an upper half (the values larger than the median). Here there are an odd number of values. If we remove the median M (corresponding to Austria), then we will have an even number of values and we can evenly divide the data into two halves. Circle these two halves of values.
2. Find the median of the lower half of values and the median of the upper half of values. Write them below. These are respectively the lower quartile and the upper quartile.
3. The five number summary consists of the smallest value, the lower quartile, the median, the upper quartile, and the largest value. Write these five numbers below.
4. On a number line, mark the locations of the five numbers.

5. These five numbers divide the dataset into quarters. Fill in the blanks in these statements.

One half of the countries have a risk value smaller than _____.

One quarter of the countries have risk values larger than _____.

Three quarters of the countries have risk values larger than _____.

Half of the countries have risk values between _____ and _____.

(There are several possible answers to the last question.)

6. The interquartile range, IQR, is the difference between the upper and lower quartiles.

For these data, $IQR =$ _____.

7. Another way of thinking about spread in a collection is based on the notion of deviation that is defined to be a value minus the mean.

Deviation = value - \bar{x} ,

For the risk data, $\bar{x} = 15.84$. The deviation for Ireland would be

Deviation = Ireland's risk - $\bar{x} = 13.8 - 15.84 = -2.04$.

This means that Ireland's risk percentage is about 2 points below the mean.

Compute the deviations for all countries and put your results in the table.

8. Suppose we are interested in a typical size of a deviation. (The size of a deviation is simply the absolute value of a deviation. For example, the size of the deviation for Ireland is +2.04.)

Construct a dotplot of the deviation sizes below.

9. Looking at the graph, find a typical deviation size.

10. One way of finding a typical deviation is to find the mean of these absolute deviations or MAD for short. Find the MAD for these data.

11. There is a second measure of computing a typical deviation size, called a standard deviation, or s for short. Use your calculator to compute s . Compare your answer with your answers in parts 9 and 10.

Share and Summarize: Here are some important points to mention when you discuss the answers to the activity.

1. It is important to focus not on the interpretation of a particular statistic, but rather the interpretation or use of this statistic.

2. The median has a simple interpretation – essentially it is the value that divides the data into halves. The interpretation of the mean isn't quite so obvious. The mean can be thought as the value that balances the distribution. (Think of data values as weights on a number line and the mean is the value on a fulcrum that balances the weight.)

Alternately, you can say the mean is the value that balances the deviations – the sum of the deviations about the mean is equal to zero.

3. Once you're talked about dividing the data into two halves to compute the median, then further division into four parts motivates the definition of the quartiles. The IQR is simply the distance between the two quartiles.

4. The idea of a deviation is important and can be described through simple examples. If one looks at a graph of the sizes of all deviations, you can talk about a typical deviation size, and that motivates the definition of the MAD and the standard deviation.

Application or Extension: Find one dataset that is symmetric and a second dataset that is strongly skewed. In each case, compute the median and the mean and compare these measures of center. When should you expect the median and mean to be of similar size, and when should you expect them to be different?

WRAP-UP

In this topic, we discussed various ways of summarizing a single dataset. For categorical data, it is helpful to find the percentages of data values in each category and the mode is the category with the highest percentage. For quantitative data, there are two primary measures of center or “average”, the median and the mean. The *median* has a clear interpretation – it is the value that separates the data into halves. The *mean* is the value such that the sum of the positive deviations from that value is equal to the sum of the negative deviations. The best choice of average depends on the dataset; we saw that the mean can be influenced by a few extreme values. The *interquartile range (IQR)* is one measure of spread that has a clear interpretation – it is the width of the middle half of the data. A second type of measure is based on the deviations about the mean. The MAD is the average size of a deviation, the mean of all absolute deviations, and the *standard deviation* is based on the sum of squared deviations. The standard deviation is

especially useful for bell-shaped data and we can use it to predict the proportion of data within one, two, and three standard deviations about the mean.

EXERCISES

1. Bird Watching

The Great Backyard Bird Count is an annual four-day event where bird watchers all over the country count birds of all species. The results of this count, published at <http://www.birdsource.org/gbbc/> give a snapshot of the numbers, kinds, and distribution of birds all over the country. The following table gives the frequencies of all types of owls spotted by bird watchers in Ohio who participated in the 2006 survey.

Owl	Frequency	Percentage
Barn Owl	6	
Eastern Screech Owl	13	
Great Horned Owl	16	
Snowy Owl	9	
Barred Owl	27	
Long-eared Owl	3	
Short-eared Owl	14	

- Find the percentage of each type of owl spotted.
- Construct a bar graph of owl type where percentage is graphed on the vertical axis.
- Find the mode.
- What percentage of owls were either of the Short-eared or Long-eared varieties?

2. Position and State Affiliation or Nationality of Professional Basketball Players

The following table gives the position and state affiliation or nationality of all players from four teams (Cleveland, New Jersey, Detroit, and Miami) who made to the second-round of the playoffs from the Eastern Conference in 2005-2006.

DAP 2011 Jim Albert -- Topic D3: Summaries for Data

Player	Position	FROM	Player	Position	FROM
Shandon Anderson	G-F	Georgia	Anderson Varejao	F	Brazil
Michael Doleac	C	Utah	Chauncey Billups	G	Colorado
Udonis Haslem	F	Florida	Kelvin Cato	C	Iowa State
Jason Kapono	F	UCLA	Dale Davis	C-F	Clemson
Alonzo Mourning	C	Georgetown	Carlos Delfino	G	Argentina
Shaquille O'Neal	C	Louisiana State	Tony Delk	G	Kentucky
Gary Payton	G	Oregon State	Maurice Evans	G	Texas
James Posey	G-F	Xavier (Ohio)	Richard Hamilton	G	Connecticut
Wayne Simien	F	Kansas	Lindsey Hunter	G	Jackson State
Dwyane Wade	G	Marquette	Jason Maxiell	F	Cincinnati
Antoine Walker	F	Kentucky	Antonio McDyess	F	Alabama
Jason Williams	G	Florida	Tayshaun Prince	F	Kentucky
Earl Barron *	C	Memphis	Ben Wallace – C	C	Virginia Union
		South Kent Prep HS			
Dorell Wright *	F	(Lawndale, CA)	Rasheed Wallace	F	North Carolina
Drew Gooden	F	Kansas	Vince Carter	G	North Carolina
Stephen Graham	G-F	Oklahoma State	Jason Collins	F-C	Stanford
Alan Henderson	F-C	Indiana	Richard Jefferson	G-F	Arizona
Larry Hughes	G	St. Louis	Jason Kidd – C	G	California
					Serbia &
Zydrunas Ilgauskas	C	Lithuania	Nenad Krstic	C	Montenegro
		St. Vincent-St. Mary			
LeBron James	F	HS (OH)	Lamond Murray	F	California
Damon Jones	G	Houston	Bostjan Nachbar	F	Slovenia
Donyell Marshall	F	Connecticut	Scott Padgett	F	Kentucky
Ronald Murray	G	Shaw	Zoran Planinic	G-F	Croatia
Ira Newble	G-F	Miami (Ohio)	Clifford Robinson	F-C	Connecticut
Aleksandar Pavlovic	G-F	Serbia & Montenegro	John Thomas	C	Minnesota
Eric Snow	G	Michigan State	Jacque Vaughn	G	Kansas
			Antoine Wright	G-F	Texas A&M

a. Construct a frequency table of the positions (G = guard, G-F = guard or forward, F = forward, F-C = forward or center, C = center) of the players. What is the mode? What proportion of players are guards?

- b. Construct a frequency table of nationality categorized into America, South America, and Europe. What proportion of players are non-American?
- c. For the players that are not American, are there differences in the nationality of professional baseball players and professional basketball players? Explain.

3. Gross Sales of Julia Roberts Movies (continued)

Here are the gross sales (in millions of dollars) for 27 movies starring Julia Roberts.

94, 67, 34, 76, 41, 34, 81, 126, 61, 3, 120, 31, 6, 67, 11
64, 127, 116, 183, 126, 101, 178, 152, 102, 51, 91, 111

- a. Compute the median and mean gross sales.
- b. Compare the median and mean and comment what these say about the shape of the distribution of gross sales.
- c. Compute a measure of spread of these gross sales.

4. Salaries of Basketball Players (continued)

The table below gives the salaries (in millions of dollars) for the players on the 2003-2004 Los Angeles Lakers basketball team.

PLAYER	SALARY	PLAYER	SALARY
Shaquille O'Neal	26.5	Gary Payton	4.9
Kobe Bryant	13.5	Bryon Russell	1.1
Brian Cook	0.8	Kareem Rush	1.1
Rick Fox	4.5	Ime Udoka	0.4
Derek Fisher	3.0	Luke Walton	0.4
Horace Grant	1.1	Jamal Sampson	0.6
Devean George	4.5	Stanislav Medvedenko	1.5
Karl Malone	1.5		

- a. Compute the median and mean salaries.
- b. Comparing the median and mean, what do these say about the shape of the distribution of salaries?
- c. If you were a basketball fan and wanted a “representative” salary of a Los Angeles Lakers player, would you be interested in the median or mean salary? Explain.

- d. If you were the owner of the Lakers and concerned about the costs of running the team, would you be interested in the median or mean salary? Explain.

5. How Long Does It Take to Score in Basketball?

At college basketball games in the United States, it is common for the home fans for a team to remain standing until their team scores its first points. That raises the interesting question: how long does it take for a college basketball team to score its first points? The espn.com website gives game logs for games played in the 2004 NCAA men's basketball tournament. Looking at the game logs for 17 games, the author recorded the time for each team to score its first points in the game. Here are the times that were recorded (in seconds):

39	66	44	58	338	195	88	23	24	39	11
44	39	62	8	15	107	136	170	66	24	198
90	114	74	122	12	84	53	20	25	37	21
25										

- Construct a dotplot of these times.
- Describe the basic shape of these times.
- Compute the median and mean time.
- If you were asked to report a typical time until the first point was scored, would you report the mean or the median? Why?
- Suppose you are watching a game and it takes 300 seconds for a team to score its first points. Based on your work above, do you think this observation is unusual? Why?

6. Weights of Newborns.

The below stemplot graphs the weights in ounces for fifty babies where the number of gestation weeks exceeds 35.

7		5
8		3
9		466
10		4589
11		011123333455567789

12 | 1233336778

13 | 23444

14 | 0155588

15 |

16 | 0

7|5 means 75 ounces

- Find the position of the median.
- Find the median.
- Suppose that the two largest baby weights of 148 and 160 ounces were recorded incorrectly – these weight values should have been 190 and 200 ounces, respectively. Without performing any computation, find the median of the new baby weights.

7. Fares of Airplane Flights to Different Cities (continued)

The round-trip air fares (in dollars) from Detroit to a number of different cities are shown in the below table.

CITY	FARE
San Francisco	310
Chicago	92
Miami	252
Denver	198
Las Vegas	242
Philadelphia	258
Boston	327
New York	170
Fargo	369
San Diego	312

- Suppose you guess that an “average” air fare for these cities is \$200. In the table below, compute the deviation of each air fare from 200, and compute the sum of the deviations. (The first two deviations have been computed for you.)

CITY	FARE	DEVIATION From 200
San Francisco	310	310-200 = 110
Chicago	92	92-200= -108
Miami	252	
Denver	198	
Las Vegas	242	

Philadelphia	258
Boston	327
New York	170
Fargo	369
San Diego	312
SUM	

b. Next, suppose that you guess that the “average” air fare is \$253. Compute the deviation of each air fare from 253, and compute the sum of the deviations.

CITY	FARE	DEVIATION FROM 253
San Francisco	310	$310 - 253 = 57$
Chicago	92	
Miami	252	
Denver	198	
Las Vegas	242	
Philadelphia	258	
Boston	327	
New York	170	
Fargo	369	
San Diego	312	
SUM		

c. Based on your work in parts a and b, which “average”, 200 or 253, is the mean of the air fares? Why?

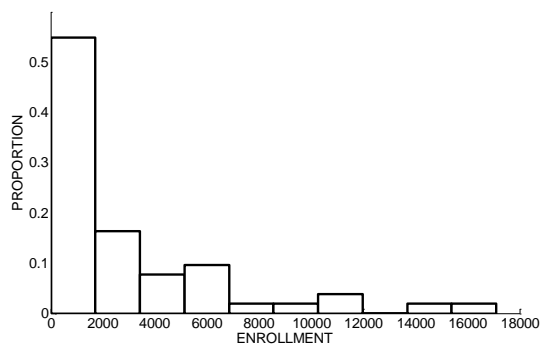
8. Test Scores

Suppose a test is given to eight students and the scores are given by 70, 60, 79, 80, 75, 100, 56, 80. The mean is given by $\bar{x} = 75$.

- Find the deviation from the mean for each data value.
- Verify that the sum of deviations from the mean is equal to 0.
- Suppose four additional students take the test. Two of the students score 80 and 90. If the mean score for all 12 students remains at $\bar{x} = 75$, find the scores of the remaining two students. (There is more than one possible answer.)

9. College Enrollments

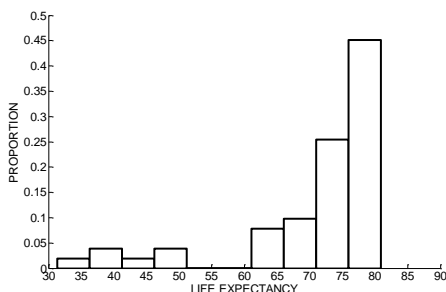
The enrollments for 104 colleges are graphed in the following histogram.



- From the histogram, make a reasonable guess at the median enrollment. (Recall the median is the value M such that half of the enrollments are smaller than M and half of the enrollments are larger than M .)
- Using your guess at the median M and the shape of the data, estimate the value of the mean \bar{x} .

10. Life Expectancy of Selected Countries.

The average life expectancy was recorded for a selection of countries in the *Time Almanac 2004*. A histogram of these life expectancies is displayed below.



Estimate the median and mean life expectancies for these countries.

11. Ages of Women Participating in a Marathon

A stemplot of the ages for 50 women participating in the 2003 Grandma's Marathon is shown below.

```

1 | 8
2 | 1111224
2 | 5566666677999
3 | 001344
3 | 5777799
4 | 02222234
    
```

4 | 677

5 | 024

5 | 8

1|8 means 18 years old

- Find a five-number summary of the ages.
- Using the five-number summary, find an interval that contains the middle 50% of the ages.
- The mean and standard deviation of the ages are given by $\bar{x} = 33.6$ and $s = 10.0$. Using these values, find an interval that contains approximately the middle 68% of the ages.

12. Braking Distances of Cars

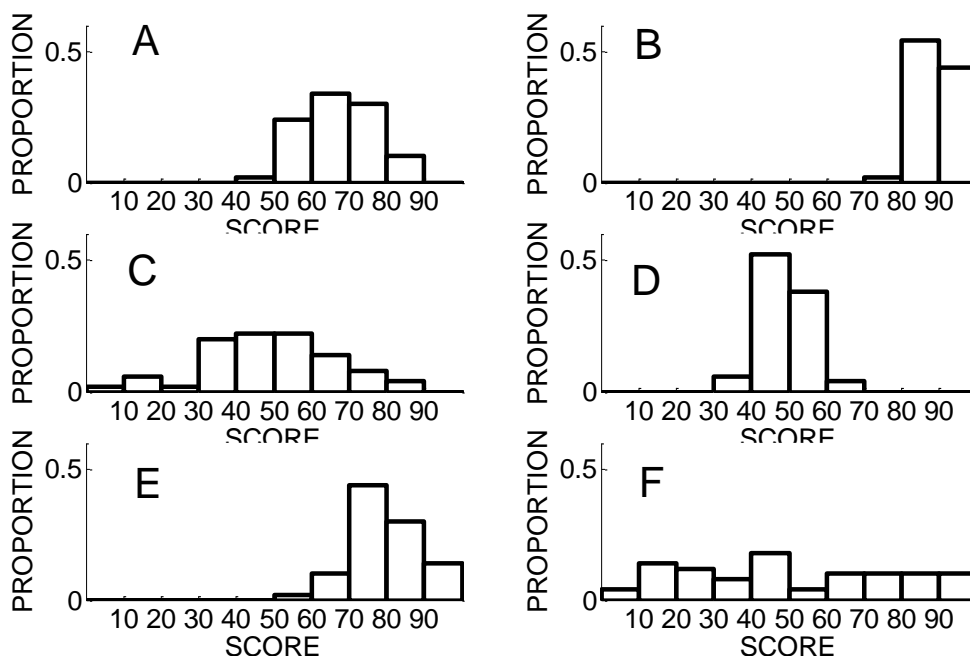
The table below gives the braking distance (feet) of 18 cars listed in *Consumer Reports New Car Preview 2004*.

Model	Braking distance	Model	Braking distance
Acura MDX	151	Nissan 350Z	116
Buick Park Avenue	137	Oldsmobile Silhouette	145
Chevrolet TrailBlazer	154	Subaru Baja	138
Dodge Neon	131	Toyota Corolla	140
Ford Taurus	151	Toyota Tundra	142
Honda Odyssey	147	Volvo S60	133
Infiniti G35	133	Cadillac DeVille	147
Lexus IS300	128	BMW 7-Series	135
Mercedes-Benz S-Class	135	Hyundai Elantra	139

- Construct a stemplot of the braking distances.
- Find the five-number summary of the distances.
- Find a car that has a “typical” braking distance.
- Suppose you are interested in a car that is tested to have a braking distance of 150 feet. Using your work in parts a and b, explain why you might not be interested in buying this car.

13. Matching Graphs and Statistics

Histograms of the test scores for six classes and six sets of statistics are shown below. Write down the letter of the histogram next to the corresponding statistics.



(mean, standard deviation) Histogram

stats 1 (67.7, 9.0) _____

stats 2 (49.1, 6.3) _____

stats 3 (51.0, 28.4) _____

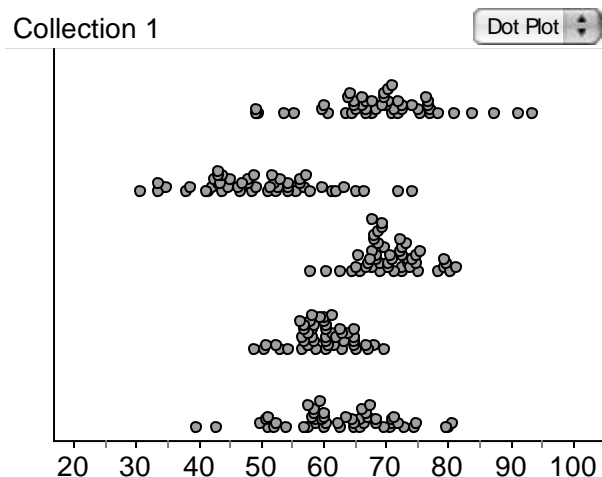
stats 4 (79.5, 9.5) _____

stats 5 (49.5, 17.4) _____

stats 6 (89.9, 4.9) _____

14. Matching Graphs and Statistics

The below figure shows dotplots of five collections of test scores, labeled Test1, Test2, Test3, Test4, Test5 followed by five sets of statistics. Write down the name of the dataset next to the corresponding statistics.



(mean, standard deviation) Dataset

stats 1 (50.2, 9.6) _____

stats 2 (63.1, 8.9) _____

stats 3 (60.0, 4.5) _____

stats 4 (69.7, 9.4) _____

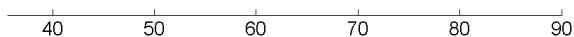
stats 5 (71.0, 4.9) _____

15. Curving a Test

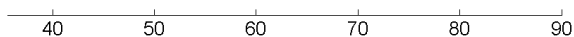
Suppose that the grades on an English test for a class of ten students are given by 45, 65, 50, 44, 66, 70, 58, 40, 60, 52.

- Construct a dotplot of the grades using the scale below.
- Since the grades are low, the teacher decides to curve the scores by adding 15 points to each student's grade. Construct a dotplot of the "curved scores" on the display below.
- The mean and standard deviation of the original test scores are $\bar{x} = 55$ and $s = 10.3$. By comparing the dotplots of the original and new scores (and not by computation), find the mean and standard deviation of the new test scores.

ORIGINAL TEST SCORES



CURVED SCORES



- d. Suppose instead the teacher decides to curve the grades by adding 25 points to everyone's score. The mean of the new grades would be _____ and the standard deviation of the new grades would be _____ .

16. Rescaling a Test

Suppose that a test has a total of 50 points and a class of ten students gets the following grades (out of 50):

30 18 36 38 24 47 47 35 38 37

- a. Find the median and IQR of these new grades.
b. Suppose the teacher decides to rescale these grades by multiplying by two (so that the new grades will be out of 100 points):

60 36 72 76 48 94 94 70 76 74

Make intelligent guesses at the median and IQR of these grades and give some rationale for your guesses.

- c. Find the median and IQR of these new grades.
d. Compare your answers to parts a and b; by multiplying the scores by 2, how has the median changed? How has the IQR changed?

17. Computing a Standard Deviation

The table below gives some of the starting calculations to compute the standard deviation of a collection of test scores. The mean score is $\bar{x} = 35$ and the “Deviation” column contains the deviations (score - \bar{x}). Complete the table and find the standard deviation s . Give an interpretation of s in the context of this problem.

Score	Deviation	Deviation squared
30	-5	
18	-17	
36	1	
38	3	
24	-11	
47	12	
47	12	
35	0	
38	3	
37	2	

18. Computing a Standard Deviation

Suppose you have five people in your family and you measure the V-span (the distance between the middle and index fingers) for all members of your family. The mean value is 6 cm. (This means that the sum of the measurements is 30 cm.) Also the smallest V-span is 3 cm and the largest V-span is 9 cm.

- Find values of the V-spans so that the standard deviation of the measurements is as small as possible.
- Find values of the V-spans so that the standard deviation of the measurements is as large as possible.
- Compute the values of the standard deviations in parts b and c.

19. Church Attendance

The worship attendance at a church in Ohio was recorded for 209 consecutive weeks. The attendance numbers are graphed in the stemplot to the right. The mean and standard deviation of these numbers are given by $\bar{x} = 361.7$ and $s = 58.5$.

```

1 | 8
2 | 0
2 | 2333
2 |
2 | 6666677
2 | 8888889999999999
3 | 00000000001111111
3 | 2222222222222222223333333333
3 | 444444444444444455555555555555
3 | 66666666666677777777777777
3 | 888888888888889999999999999999
4 | 00000000000000001111111111
4 | 2222222222333
4 | 4555
4 | 6777
4 | 899
5 |
5 | 233

```

1|8 corresponds to 180

- Can you apply the 68-95-99.7 rule for this dataset? Why or why not?
- Find an interval that you believe will contain approximately the middle 68% of the attendance numbers.
- From the stemplot, find the proportion of values that fall in the interval you found in part b. Does this proportion agree with the 68% in the 68-95-99.7 rule?
- Find an interval symmetric about the mean that you believe will contain approximately 95% of the numbers.
- Find the actual number of attendance numbers that fall in the interval from part d. Comment if this proportion is close to what you would expect from part d.

20. Snowfall in Tulsa

The yearly snowfall in Tulsa, Oklahoma was recorded for the years 1950-2003. A stemplot of the snowfall amounts (in inches) is displayed to the right. The mean and standard deviation of these snowfall amounts are given by $\bar{x} = 9.5$ and $s = 6.0$ inches respectively.

```

0 | 0011
0 | 233
0 | 4444445555
0 | 6666667777
0 | 88999
1 | 00011
1 | 223
1 | 44445555
1 | 66

```



```

1 |
2 | 00
2 | 2
2 |
2 |
2 |
2 | 9

```

0|1 corresponds to 1 inch

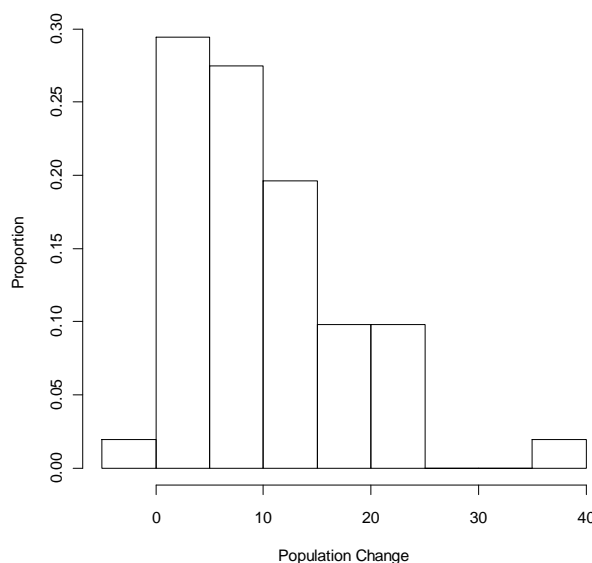
- Is it appropriate to apply the 68-95-99.7 rule to these data? Why or why not?
- Find the interval $(\bar{x} + s, \bar{x} - s)$ and find the proportion of snowfall amounts that fall in this interval.
- Find the interval $(\bar{x} + 2s, \bar{x} - 2s)$ and find the proportion of snowfall amounts that fall in this interval.
- Are the proportions you computed in parts b and c close to what you expected if you use the 68-95-99.7 rule?
- If the answer to part d is no, explain why the 68-95-99.7 rule may not be applicable in this situation.

21. State Population Changes

The table below gives the percentage change in population from 2000 to 2010 for all states in the United States.

State	%change	State	%change	State	%change	State	%change
AL	7.5	IL	3.3	MT	10.8	RI	0.5
AK	13.2	IN	6.6	NE	6.7	SC	15.3
AZ	24.6	IA	4.1	NV	35.2	SD	7.8
AR	9.1	KS	6.1	NH	6.5	TN	11.5
CA	11.0	KY	7.3	NJ	4.5	TX	20.6
CO	16.9	LA	1.4	NM	13.2	UT	23.8
CT	4.9	ME	4.2	NY	2.1	VT	2.8
DE	14.9	MD	9.0	NC	18.5	VA	13.0
DC	5.2	MA	3.1	ND	4.8	WA	14.1
FL	17.6	MI	-0.5	OH	1.6	WV	2.5
GA	18.3	MN	7.8	OK	8.7	WI	6.0
HI	12.2	MS	4.3	OR	12.0	WY	21.5
ID	21.2	MO	7.0	PA	3.4		

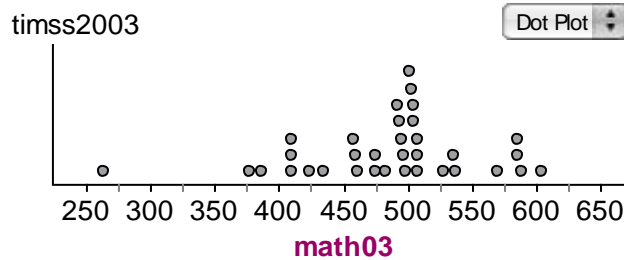
A histogram of these percentage changes in population is shown below.



- Write a short paragraph about this dataset, discussing shape, a typical value, spread, and any unusual characteristics.
- Looking back at the data table, find the location (region of the U.S.) of the states with the large percentage increases.
- The mean and median of the population changes (in percent) are given by $\bar{x} = 9.955$ and $M = 7.8$, respectively. Explain why these two measures of center are different for this dataset.
- Suppose that the largest population change 35.2, corresponding to Nevada, is removed from the dataset. Recalculate the mean and median for this reduced dataset.
- We say that a measure is *resistant* if its value is not influenced by one extremely large or small observation. Are the median or mean resistant measures? Explain why or why not.

22. International Study of Mathematics Achievement

The Trends in International Mathematics and Science Study (TIMSS) collects data on mathematics and science achievement from countries all around the world. The display below shows a dotplot of the mean mathematics score of 8th graders from 34 countries from the 2003 study. The median and mean scores are given by 497 and 485.02, respectively.



- What proportion of countries have scores smaller than 497?
- Suppose the low score of 264 is removed from the dataset. Recompute the median and the mean? Are they more or less similar in value? Is this to be expected? Why?
- Suppose South America's low score of 264 is replaced by Oz's (fictitious country) score of 150. Compute the median and mean of the new dataset.
- Which measure (median or mean) is most affected by the unusually small score?

23. City Temperatures

The below table displays the average temperature (in degrees Fahrenheit) for eight American cities for each month of the year.

Month	San Francisco	Vero Beach	Duluth	Albuquerque	San Diego	Philadelphia	Honolulu	Indianapolis
Jan	48.7	61.6	7.0	34.2	57.4	30.4	72.9	25.5
Feb	52.2	62.7	12.3	40.0	58.6	33.0	73.0	29.6
Mar	53.3	67.2	24.4	46.9	59.6	42.4	74.4	41.4
Apr	55.6	71.3	38.6	55.2	62.0	52.4	75.8	52.4
May	58.1	75.8	50.8	64.2	64.1	62.9	77.5	62.8
Jun	61.5	79.5	59.8	74.2	66.8	71.8	79.4	71.9
July	62.7	81.1	66.1	78.5	71.0	76.9	80.5	75.4
Aug	63.7	81.3	63.7	75.9	72.6	75.5	81.4	73.2
Sep	64.5	80.1	54.2	68.6	71.4	68.2	81.0	66.6
Oct	61.0	75.5	43.7	57.0	67.7	56.4	79.6	54.7
Nov	54.8	69.3	28.4	44.3	62.0	46.4	77.2	43.0
Dec	49.4	63.7	12.8	35.3	57.4	35.8	74.1	30.9

- For each city, find the five-number summary of temperatures for the 12 months. Place your calculations in the table below. Also, for each city, find the interquartile range (IQR) of temperatures.

	Five-number summary					
CITY	LO	Q_L	M	Q_U	HI	IQR
San Francisco						
Vero Beach						
Duluth						
Albuquerque						
San Diego						
Philadelphia						
Honolulu						
Indianapolis						

- b. By use of the medians, order the cities from coldest to warmest.
- c. Suppose you classify a city's temperature as volatile (changing or varying across months) or stable (little change across months). What quantity would you use to measure the volatility of a city's temperature? Using this measure, order the cities from most volatile to most stable.

24. Points Scored for Basketball Games

The number of points scored in the first nine basketball playoff games during the 2004-5 season is recorded below for four Phoenix Suns players (Amare Stoudemire, Shawn Marion, Steve Nash, and Quentin Richardson).

Game	Stoudemire	Marion	Nash	Richardson
1	9	26	11	22
2	34	22	12	15
3	30	14	13	9
4	18	23	24	14
5	40	23	11	12
6	30	23	23	12
7	37	21	27	12
8	15	19	48	13
9	Did not play	16	34	7

- a. On the average, which player scored the most points? Explain what measure you are using to measure the average.

- b. Which player was the most consistent scorer for these basketball games? Explain what measure you are using to measure consistency of scoring.