

Section 2: Methods for Describing Sets of Data

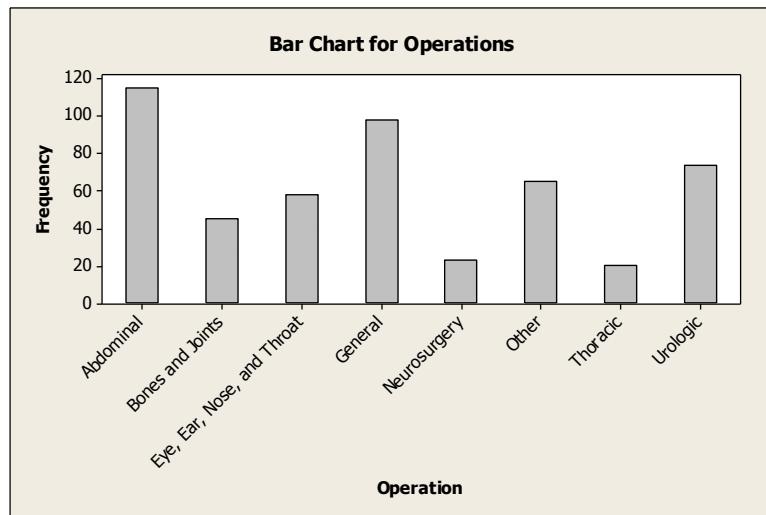
We are now to the point in the class where we will begin looking at descriptive statistics. Remember that descriptive statistics are methods of summarizing a set of data. One way to do that is graphically. The main reason statisticians use graphs is to look at the distribution of data.

Distribution – all outcomes for a variable along with how often each outcome occurs

Graphing Qualitative Data

The following qualitative data set will be utilized in order to illustrate a bar chart. The variable of interest in this illustration is type of operation (data is non-numerical and was collected for a one year period from a single hospital). Qualitative data is commonly summarized with either frequencies or percentages which are both identified in the following table.

Type of operation	Frequency	Percentage
Abdominal	115	23.1
Bones and joints	45	9.0
Eye, ear, nose, and throat	58	11.6
General	98	19.7
Neurosurgery	23	4.6
Other	65	13.1
Thoracic	20	4.0
Urologic	74	14.9
Total	498	100.0



Important characteristics of a **Bar Graph**

- Used only for qualitative data (This is a common mistake that I see with this graph. These are very easy to make in Excel and sometimes they are made with quantitative data. That should never be done because the x-axis is not numerical in a bar graph. It does not matter in what order the bars are displayed. All horizontal dimensions in the graph are based on the idea that there is no numerical meaning to this axis.)
- The length of a bar represents the quantity we wish to compare
- Users tend to notice the tallest bars
- The bars should be of uniform width and uniformly spaced (This goes back to the x-axis being non-numerical. The width of the bars and spacing has no numerical meaning)

Graphing Quantitative Data

The goal with these graphs is to look at the distribution for a single quantitative variable. When looking at the graphs presented, it is important to identify if there are extreme values in the data and the shape of the distribution.

Extreme value (or Outlier) – observations that are separated from the rest of the data set by some margin

Example

Identify the extreme value in the following data.

14, 15, 17, 21, 60, 23, 25

Answer

The extreme value is 60.

The impact of an extreme value can be great. Suppose we want to estimate the average income for this class. If a billionaire was in the class, then their income would be an extreme value. That would make the average appear very high simply because of one observation.

Shape – the pattern displayed when the graph is created (The most common shapes of distributions will be introduced after the graphs.)

Important characteristics of a **Stem-and-Leaf**

- Separates data entries into “leading digits” or “stems” and “trailing digits” or “leaves”
- Organizes and groups data but allows us to recover the original data if desired
- Good for spotting extreme values and identifying shape

Example

14 male weights in pounds: 139,153,179,201,163,168,157,170,172,165,145,155,161,151

In this case the first two digits of each number represent the stems and the last digit represents the leaves. The stem-and-leaf follows.

13	9
14	5
15	1 3 5 7
16	1 3 5 8
17	0 2 9
18	
19	
20	1

Notice in the stem-and-leaf plot it is easy to identify the one extreme value of 201. If a data point is separated from the rest of the data by at least one empty class (in this case two) then it is considered an extreme value. The shape is identified by looking at the pattern created with the numbers. This will be discussed later in this section.

Important characteristics of a **Frequency Distribution**

- A summary table in which quantitative data are arranged into conveniently established class groupings (the groups must satisfy the following criteria)
 - Should have between 5 and 15 classes
 - Each class grouping should be of equal width
 - Overlapping the classes must be avoided
- Useful when dealing with very large data sets
- Through the grouping process the original data is lost
- **Class midpoint** – the point halfway between the boundaries of each class

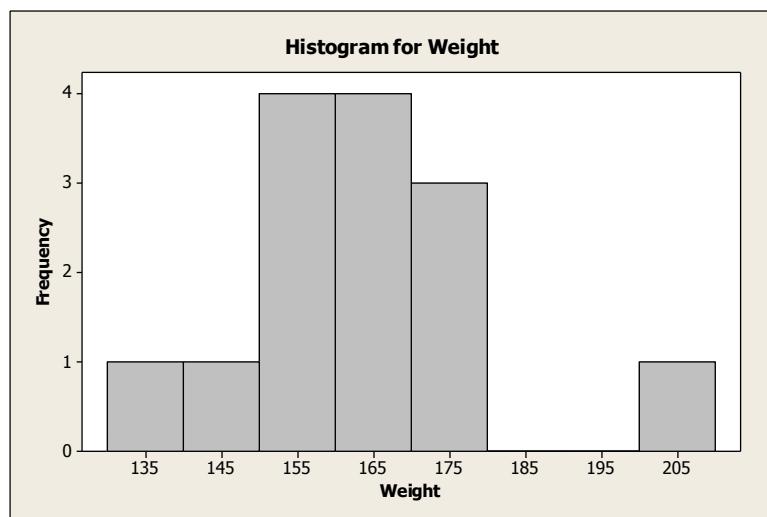
In order to illustrate a frequency distribution, we will look at the same data presented above in the stem-and-leaf (the 14 male weights).

	Weight	Frequency
13	9	1
14	5	1
15	1 3 5 7	4
16	1 3 5 8	4
17	0 2 9	3
18		0
19		0
20	1	1
	Total	14

In the frequency distribution for weight, the same groupings were used as was naturally created with the stem-and-leaf plot. Notice with a big data set, the stem-and-leaf would become very messy, but the frequency distribution would simply have bigger frequencies. The down side is that the actual data values cannot be identified in the frequency distribution.

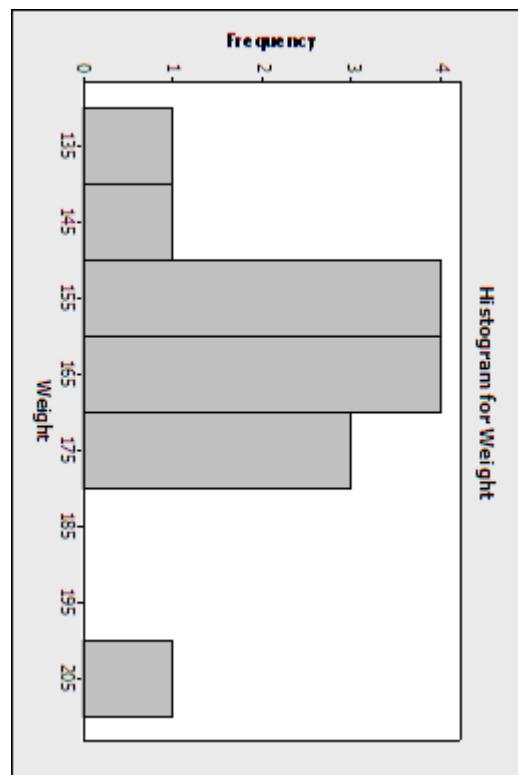
Important characteristics of a **Histogram**

- The variable (x-axis) in a histogram is quantitative (different than a bar chart). Since the variable is quantitative, the width of the bars has numerical meaning and the bars always touch. The bars represent the classes identified in the frequency distribution.
- A picture of a frequency distribution
- The actual values identified on the x-axis are the class midpoints
- The y-axis represents the frequency for each class



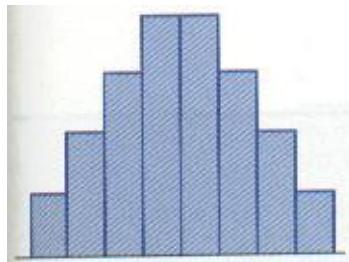
It is important to understand how the above histogram was created based on the frequency distribution. It is also important to see how it is similar to the stem-and-leaf. Notice the same extreme and shape would be identified if the graph is turned as can be seen below.

13	9
14	5
15	1 3 5 7
16	1 3 5 8
17	0 2 9
18	
19	
20	1

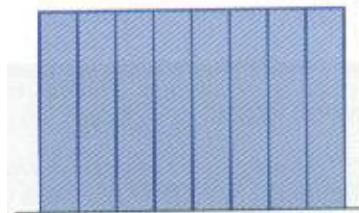


Shapes of Distributions

Symmetrical – both sides are the same when the graph is folded vertically in the center

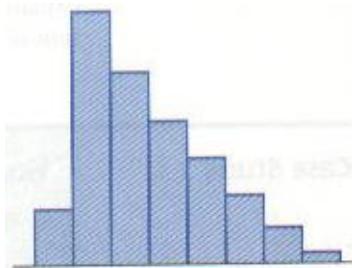


Uniform – every class has equal frequency (bars are the same height)

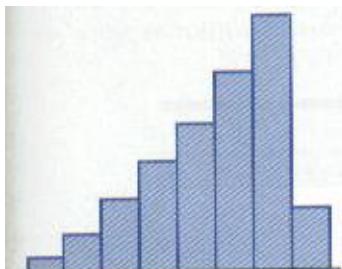


Skewed – one tail is stretched longer than the other (not symmetric). The direction of the skewness is on the side of the longer tail. We start at the highest frequency class (highest bar) and from there the graph tails to the right and to the left.

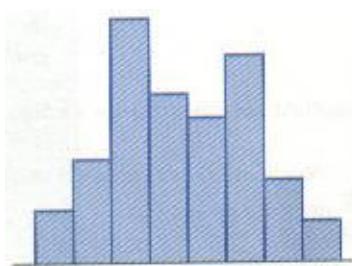
Skewed Right – tail to the right is stretched longer than the tail to the left



Skewed Left – tail to the left is stretched longer than the tail to the right



Bimodal – there are two distinct peaks (the peaks are often the two classes with largest frequencies, but not always)



Summation Notation

The sum of values, $x_1 + x_2 + \dots + x_n$, can be denoted as $\sum_{i=1}^n x_i$.

- Read ‘sum of x sub i for i equals 1 to n’
- x represents the data
- n represents the number of items being summed (sample size)
- i is an index that represents the observation number (x_1 is the first data point, x_2 is the second data point, and so on)

Example

Select 4 students and ask “how many brothers and sisters do you have?”

Data: 2,3,1,3

The observations (data values) would be denoted as follows:

$$x_1 = 2$$

$$x_2 = 3$$

$$x_3 = 1$$

$$x_4 = 3$$

The sum of the observations would be written:

$$\sum_{i=1}^4 x_i = x_1 + x_2 + x_3 + x_4 = 2 + 3 + 1 + 3 = 9$$

In statistics many times we leave off the index in the summation notation. This is because the sum is almost always taken over the entire sample (1 to n). Therefore, when you see a sum written without the index, you should assume it is 1 to n. For this class all sums will be over the sample size. This means that the above sum will commonly be written as:

$$\sum x = 9$$

Properties of sums (c is a constant)

- $\sum cx = c \sum x$

A constant can be pulled out of a sum. If there is a c in every term of the sum then we know that it can be pulled out as in: $cx + cy + cz = c(x + y + z)$.

- $\sum c = nc$

If you sum a constant n times then you get the product of n and the constant. Keep in mind that multiplication is nothing more than repeated addition. If you sum 5 values of 10, you get: $10 + 10 + 10 + 10 + 10 = 5(10) = 50$.

- $\sum(x + c) = \sum x + \sum c = \sum x + nc$

When multiple terms are included in the sum, you can separate those out and sum each part. This is because of the commutative property of addition ($a + b = b + a$). It does not matter what order you add numbers, the sum will be the same. Therefore, it does not matter if you add the terms first and then do the sum or if you do the sum for each term first and then add.

In this section, all sum problems will include data which means there are two ways to do the problems. You can use a table to identify the outcome for each x and then sum or you can algebraically simplify the sum. It is important to get comfortable with both techniques because we will continue to use them throughout the semester. In addition, there will be simplifications in later sections where no data is given. This means that you must simplify the expression algebraically. Make sure you understand how to do this now while we are dealing with simple sums.

Example

Solve the following

- $\sum 4x$
- $\sum(x + 3)$
- $\sum(4x + 3)$
- $\sum(4x + 3)^2$

Answers using a Table

x	$4x$	$x + 3$	$4x + 3$	$(4x + 3)^2$	x^2
2	8	5	11	121	4
3	12	6	15	225	9
1	4	4	7	49	1
3	12	6	15	225	9
$\sum x = 9$	$\sum 4x = 36$	$\sum(x + 3) = 21$	$\sum(4x + 3) = 48$	$\sum(4x + 3)^2 = 620$	$\sum x^2 = 23$

Answers using Algebraic Simplification

a.
$$\begin{aligned} & \sum 4x \\ &= 4 \sum x \\ &= 4(9) \\ &= 36 \end{aligned}$$

b.
$$\begin{aligned} & \sum(x+3) \\ &= \sum x + \sum 3 \\ &= \sum x + 4(3) \\ &= 9 + 12 \\ &= 21 \end{aligned}$$

c.
$$\begin{aligned} & \sum(4x+3) \\ &= \sum 4x + \sum 3 \\ &= 4 \sum x + 4(3) \\ &= 4(9) + 12 \\ &= 48 \end{aligned}$$

d.
$$\begin{aligned} & \sum(4x+3)^2 \\ &= \sum(16x^2 + 24x + 9) \\ &= \sum 16x^2 + \sum 24x + \sum 9 \\ &= 16 \sum x^2 + 24 \sum x + 4(9) \\ &= 16(23) + 24(9) + 36 \\ &= 368 + 216 + 36 \\ &= 620 \end{aligned}$$

Numerical Measures of Central Tendency

We will now begin calculating statistics. Remember a statistic is a numerical characteristic of a sample. These are descriptive statistics since they will summarize the data in the sample.

A measure of central tendency is a measure of average or typical value. The three most common measures of central tendency are the mean, median, and mode which you have probably seen before. We will discuss these values along with the midrange.

Sample Mean – when most people use the word average, they are talking about the mean

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

where \bar{x} is the sample mean

Σ is the sum

x is the data values

n is the sample size

The mean is the sum of the data values divided by the sample size.

Example

Back to number of siblings Data: 2,3,1,3

Calculate the mean.

Do you think the mean is a good measure of center for this data?

Answer

$$\bar{x} = \frac{\sum x}{n} = \frac{2 + 3 + 1 + 3}{4} = \frac{9}{4} = 2.25$$

Seems like a reasonable measure of center in this case. With a value of 1, a value of 2, and two values of 3, we would expect the center to be between 1 and 3. Since there are two values of 3 we would expect it to be a little closer to 3 than 1. Also, there are two data values below the mean and two above the mean.

Example

Suppose we had selected a 5th person for our sample which had 10 siblings.

New Data: 2,3,1,3,10

Calculate the mean.

Do you think the mean is a good measure of center for this data?

Answer

$$\bar{x} = \frac{\sum x}{n} = \frac{2 + 3 + 1 + 3 + 10}{5} = \frac{19}{5} = 3.8$$

In this case it does not seem like a reasonable measure. Notice the mean is above every data point except for the value of 10. This value would be identified as extreme since it is not typical to have 10 sibling. The mean is not a good measure of center when extreme values are present in the data set.

Important characteristics of the sample mean are:

- \bar{x} is sensitive to extreme scores
- \bar{x} is not necessarily a possible outcome

An applied example of the mean not being used when extreme values are present is income. If you hear anyone talk about average income, they should say median income. The median is a much better measure for center in this case than the mean. Consider if I wanted to calculate average income for this class. If a billionaire was in the class, what effect would that have on the mean? It would make it very high and not a good measure for a typical value.

Sample Median - the middle score

Procedure for calculating \tilde{x} (denotes the sample median):

- rank data from smallest to largest
- if n is odd, median is the middle score
- if n is even, median is the mean of the two middle scores

Example

Back to number of siblings Data: 2,3,1,3

Solve for the median.

Answer

1,2,3,3

$$\tilde{x} = \frac{2 + 3}{2} = 2.5$$

Example

New Data: 2,3,1,3,10

Solve for the median.

Answer

1,2,3,3,10

$$\tilde{x} = 3$$

Important characteristics of the median are:

- \tilde{x} is not sensitive to extreme scores
- exactly half of the data is below \tilde{x} and exactly half of the data is above \tilde{x}

Because of the characteristics of the mean and median, if extreme scores exist in a data set the median is a better measure of central tendency.

If extreme scores are unlikely, the mean varies less from sample to sample than the median and is a better measure of center.

It can be determined if a distribution is symmetric or skewed based on the mean and median because of their characteristics. The mean is sensitive to extreme values and is generally more affected by values further away from center. Therefore, if a distribution is skewed, the mean is pulled in the direction of the skewness. For example, if a distribution is right skewed then the data to the right tends to be further away from center so the mean is pulled in that direction. If there is no skewness then the center is obvious and the mean and median will both be at the point of symmetry.

- If $\bar{x} > \tilde{x}$ then the distribution is right skewed
- If $\bar{x} < \tilde{x}$ then the distribution is left skewed
- If $\bar{x} = \tilde{x}$ then the distribution is symmetric

Sample Mode - the most frequent score

Example

Data: 2,3,1,3

New Data: 2,3,1,3,10

Calculate the mode for the above data sets.

Answer

Mode = 3 (for both the Data and New Data)

There are some major weaknesses with the mode. For example suppose that in the New Data the 10 was changed to a 2. Then what is the mode? You can say it has two modes, both 2 and 3 or you can say the mode does not exist. Even worse, suppose one of the values of 3 was instead a 4. Then you can say the mode does not exist or that all the data points are the mode. Another issue is if one of the 3 values changes to a 1 the mode is 1. Notice selecting one value by random chance over another can totally change the mode making it unstable.

The one advantage of using the mode is it can be used for qualitative data and that is when it should be used. For example if I say the average student in this class is female, what does that tell you? It just means that there are more females than males or the mode is female. The mean or median cannot be used for qualitative data.

Important characteristics of the mode are:

- does not always exist / can be more than one
- unstable
- can be used with qualitative data

Sample Midrange – the average of the lowest and highest observation in the data set

$$\frac{\text{Low} + \text{High}}{2}$$

Example

Data: 2,3,1,3

New Data: 2,3,1,3,10

Calculate the midrange for the above data sets.

Answers

$$\text{For Data: Midrange} = \frac{\text{Low} + \text{High}}{2} = \frac{1+3}{2} = 2$$

$$\text{For New Data: Midrange} = \frac{\text{Low} + \text{High}}{2} = \frac{1+10}{2} = 5.5$$

The midrange is utilized to illustrate that there are many statistics available to choose from, but some are much better than others. The midrange is not a very good measure of center for most data sets. Think about it. You may have 1000 data points. You only select two of those data points to calculate the midrange and the two you pick are the most extreme data points. The midrange is totally dependent on extreme scores. The only time it is a good measure is when the data is perfectly symmetric. In practice, perfect symmetry is very rare.