CHAPTER 4

Summary Statistics

Today we look at how to describe a data set using a ***histogram,*** a type of data display. The data in a histogram show how they group together and how far they spread out from the center. The numbers we shall consider are called 'Numerical Descriptors.' The ones that tell us how the data ***group*** together are called **Measures of Location or Measures of Center,** and the ones that indicate the ***spread*** are known as **Measures of Dispersion or Measures of Spread.** We will consider Measures of Dispersion in the next lesson.

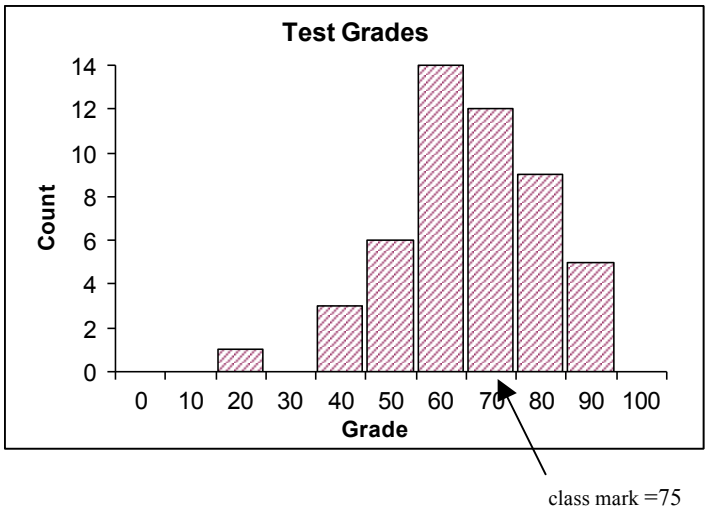
Histograms

Consider the following data set from Chapter 1. We will demonstrate how to display this data using *class intervals*. Notice in the graph below that the horizontal *Axis* is labeled with the category and the vertical *axis* is labeled with the *frequency* (how many).

Look at a list of the grades received by the students in the class, such as the following:

73	47	71	65	85	61	74	80	65	67
62	84	71	99	70	88	69	83	81	71
80	68	26	58	71	90	58	64	70	95
91	67	48	41	62	75	66	77	78	66
50	52	83	58	50	87	91	60	78	67

It is very difficult to get much of an idea of how the class did just by looking at these numbers. However, there are ways of presenting the data that make things much clearer. For Example, the diagram below, called a histogram, shows you at once how the grades are distributed.



This histogram was set up using a **class interval** of size 10. For example, the column between 60 and 70, contains the number of students who received grades between 60 and 70 but not 70. If the class interval was 5, then the intervals would be 20-25, 25-30, 35-40, 45-50, etc.

***NB**: If we are given a histogram like this one WITHOUT knowing the 'raw' data, we use the 'class mark' for all calculations, e.g. mean, median. The 'class mark' is the middle value of each interval even though the last value is not included in that interval. For example, the class marks of the above histogram are 25, 35, 45, 55, etc.

Counting and Organizing the Data

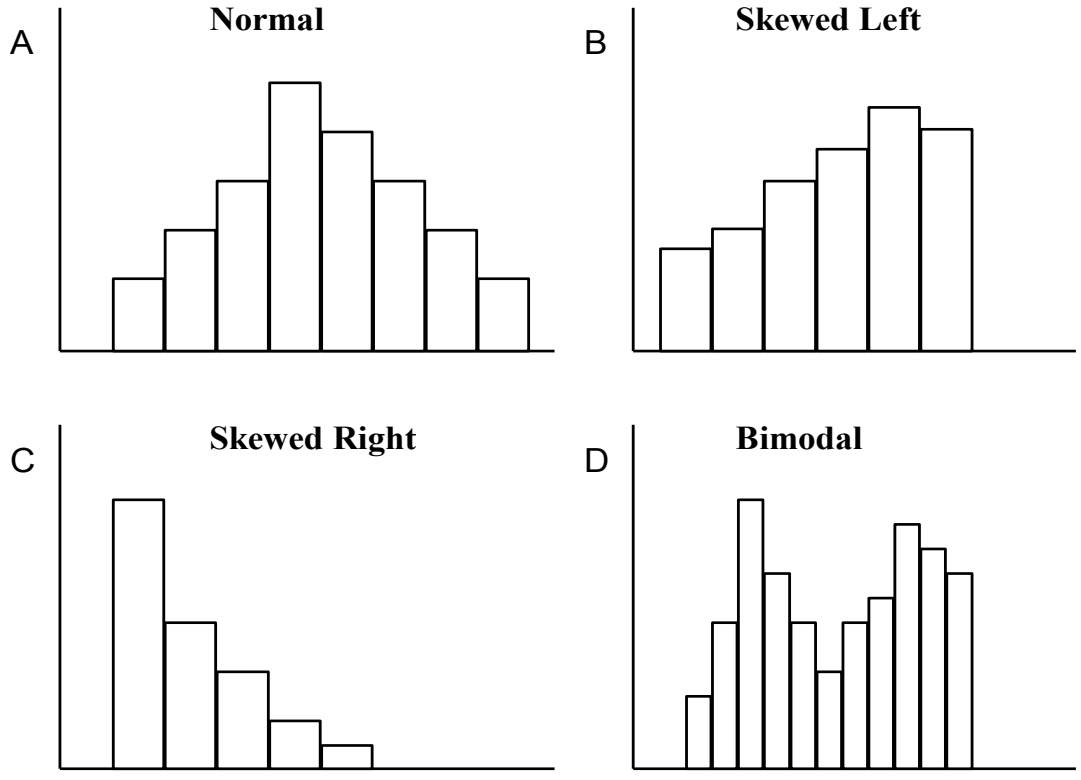
The histogram on the previous page was constructed by first organizing the data. There are several ways to *count* up the number of student grades in each interval, but one easy and fun way is by using a Stem and Leaf display. The stem is the first digit of our grades. The second digit is put into the row next to the stem. We will demonstrate this problem in class.

Stem	Leaf (1st Pass)
2	6
3	--
4	7 8 1
5	0 2 8 8 0 8
6	2 8 7 5 2 1 9 6 4 0 5 7 6 7
7	3 1 1 0 1 5 4 7 0 8 8 1
8	0 4 3 5 8 7 0 3 1
9	1 9 0 1 5

Count	Stem	Leaf (2nd Pass – numerical Order)
1	2	6
0	3	--
3	4	1 7 8
6	5	0 0 2 8 8 8
14	6	0 1 2 2 4 5 5 6 6 7 7 7 8 9
12	7	0 0 1 1 1 1 3 4 5 7 8 8
9	8	0 0 1 3 3 4 5 7 8
5	9	0 1 1 5 9
Total 50		

Contours - (Shapes)

The following four histograms demonstrate the general ***shapes*** of histograms. Notice that some histograms are *leaning* in one direction. These are called ***skewed*** histograms and they are defined as *skewed* in the direction of the '*tail*.'



Definitions: **The *Measures of Location* - These '*location descriptors*' are often known as '*measures of center*' or '*measures of central tendency.*' They are *mean*, *median*, and *mode*.**

Mean – The mean is the ***Average*** of all data values.

i.e. add data and divide by how many there are. (we donate 'how many' or frequency as 'n')

$$\textbf{Mean} = \frac{\text{Sum of all data values}}{\text{total number of data}} = \frac{\sum x}{n} = \bar{x}$$ We denote the mean by an x with a bar, $\bar{x}$

Example 1: A class of 15 students in English 101 received the following grades on a test. Find the class average.

100, 60, 90, 90, 70, 100, 100, 90, 90, 80, 70, 100, 80, 90, 80

$$\bar{x} = \frac{\sum x}{n} = \frac{100+60+90+90+70+100+100+90+90+80+70+100+80+90+80}{15} = \frac{1290}{15} = \mathbf{86}$$

$\bar{x} = \mathbf{86}$ Note: $\mathbf{n = 15}$

A more efficient way of calculating the mean, especially with large data sets, is to **GROUP** the data into a ***frequency table.*** That is, count up 'how many' we have of each value, etc. Our Data Value is the Grade.

Table 1: *A Frequency Table for English 101*

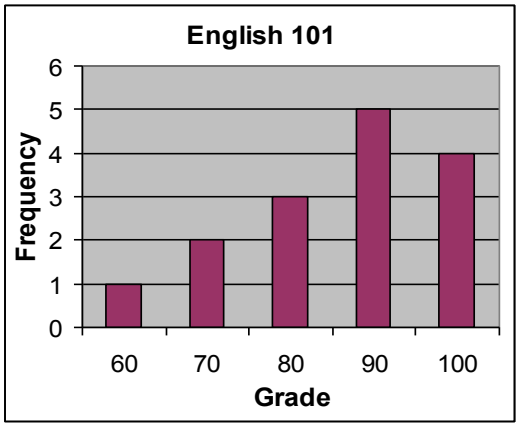
Data Value	How Many
60	1
70	2
80	3
90	5
100	4

Now that the data are grouped, simply multiply the value by the frequency and divide by total frequency. That is, multiply across the table and sum:

$$\textbf{Mean} = \frac{60(1)+70(2)+80(3)+90(5)+100(4)}{15} = \frac{60+140+240+450+400}{15} = \frac{1290}{15} = \mathbf{86}$$

Drawing the Histogram:

In this case, use a class interval of one. There really is no class interval, simply a bar over each data value. Single numbers represent the class.



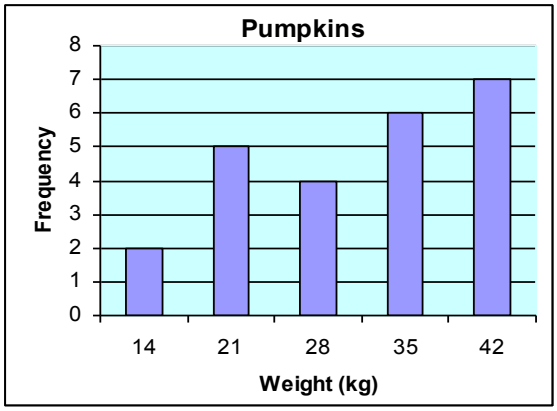
Example 2.- The following is list showing the weights (in kg) of **24** prize pumpkins. (The weights have been rounded to the nearest whole pound.) Find the mean (average) weight and draw the histogram associated with this data. The raw data has been put into a frequency table, which facilitates our calculations and drawing.

28	35	42	14	28	42
42	14	21	35	42	35
21	42	28	42	21	35
35	21	42	28	35	21

Note: **n = 24** (total number of pumpkins.)

Table 2: *Prized Pumpkin Weights (kg)*

Weight	**Frequency**
14	2
21	5
28	4
35	6
42	7



$$\textbf{Mean} = \frac{2(14)+5(21)+4(28)+6(35)+7(42)}{24} = \frac{28+105+112+210+294}{24} = \frac{749}{24} = 31.2$$

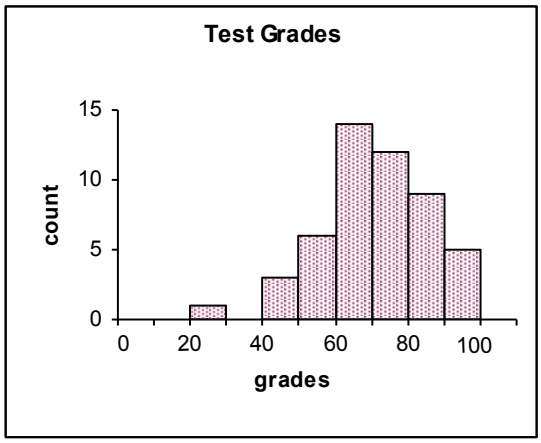
$\bar{x}$ **= 31.2 kg** The mean weight of this group of pumpkins is 31.2 kg.

Example 3. –Let's find the mean of the very first histogram with 50 data values *assuming we do not know the raw data*. Assume we are given only the histogram. In this case, we need to use the 'class mark' or the value that sits in the middle of each 'interval.'

$$\textbf{Mean} = \frac{1(25)+3(45)+6(55)+14(65)+12(75)+9(85)+5(95)}{50}$$

$$= \frac{25+135+330+910+900+765+475}{50} = \frac{3540}{50} = 70.8$$

(The actual mean from the raw data is 69.86)



Median - *Middle* data value - data values are in numerical order

The median is that value such that half the data values are to the left of it and half to the right of it. It need not be exactly one of the data values, but it must be between them. The only criteria for finding the median is that the data values must be put in numerical order.

Finding the Median - Since the median is the 'middle' value we are going to Divide the frequency by 2. That is: $\frac{n}{2}$

However, the middle value will depend on whether this frequency of the data is Odd or Even. When you divide by 2, you are going to find the POSITION of the median, NOT the Median itself. This is the most difficult part of the concept to understand.

ODD: If the frequency (Number of data values) is Odd, then, after dividing by 2, you get a decimal Round up to the next whole number. The value tells you WHERE to find the Median. It is Not the actual Median. It is the POSITION of where the median sits. The MEDIAN is the DATA value that belongs in this Position.

EVEN: If the frequency is even, then after you divide by 2, you get a whole number. The POSITION of the median lies **between** this whole number and the next one.

We will demonstrate.

Look at our **Example 1** (above) Here's the frequency table again:

Table 1: *A Frequency Table for English 101*

Data Value	How Many
60	1
70	2
80	3
90	5
100	4

The total frequency = 15, i.e. n = 15 This is an Odd number

Position of Median = $\frac{15}{2} = 7.5 = 8$ So Median will be in the 8^{th} Position..

Finding the actual MEDIAN data value. In this case, we are finding the MEDIAN score on the English 101 test.

(Notice that we said the MEDIAN can be FOUND in the 8^{th} place. The median is **NOT** = 8. Eight is the **spot** where the median will be found.)

Let's count (in the frequency column, i.e. the 'How Many' column) until we get to the 8th spot. If we start at the top we get

$1 + 2 + 3 = 6$ this is the 6^{th} spot. Thus the 8^{th} spot is in the next row.

This is the row which corresponds to a grade of 90 The Median is therefore **90.**

Note: if we start counting from the bottom of the table, we get

$4 + 5 = 9$ so the 8^{th} spot is still in the row for the data value = 90

In order to SEE the value of the Median, line up all the data in numerical order and count up to the 8^{th} value. (With a large data set, this is obviously inconvenient.) For the above problem, we get:

60, 70, 70, 80, 80, 80, 90, 90, 90, 90, 90, 100, 100, 100, 100

this is the **8^{th} spot.**

Notice the same number of data are to the left and right of the median.
Thus the median is the center of the data.

Let's find the median for the Pumpkin data.

Example 2. Find the median pumpkin weight. There were 24 pumpkins so the total frequency is 24, i.e. **n = 24.**

$\frac{n}{2} = 12$ Since we have an even number, the median is between the 12^{th} and 13^{th} spot.

From the table, count the frequencies until you get to the 12^{th} and look at the 13th spot.

Table 2: *Prized Pumpkin Weights (kg)*

Weight	Frequency	
14	2	
21	5	
28	4	
35	**6**	◄— Median is in this row, since the 12^{th} and 13^{th} spots are here.
42	7	

We see that the weight that matches this depth is **35 kg.** **Median pumpkin weight is 35 kg.**

On many occasions, we may end up with two different data values, in that case we 'average' the two data values.

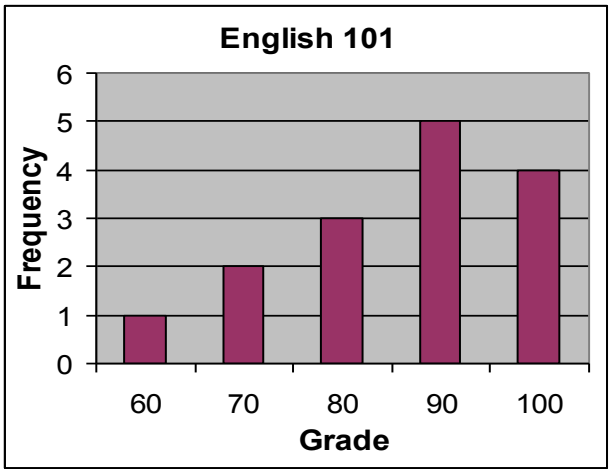
For example: Say the 12^{th} spot was 28 and the 13^{th} spot was 35. Then the median weight Would have been the average between 28 and 35.

$\frac{28 + 35}{2} = 31.5$ So in this hypothetical case, the median pumpkin weight would be **31.5 kg.**

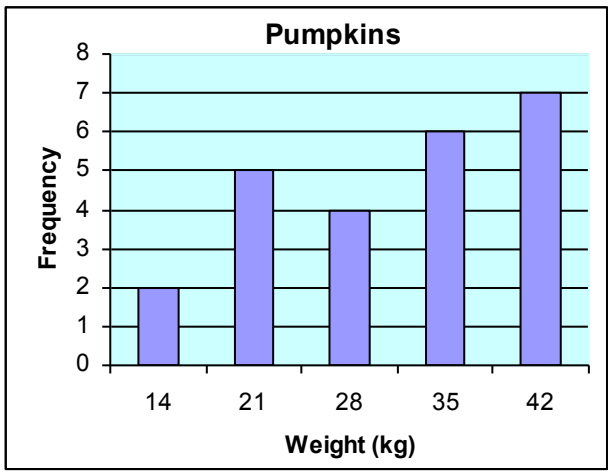
Mode - ***Most*** frequently occurring data value.

The mode can be found by looking at either the frequency table or the histogram. You can find the mode in the frequency table by seeing which 'frequency' is the highest. Then look across to see which data value it is. Or, just look at the histogram and see where the highest column ***sits***. That is, look at the horizontal axis and see which data value the column is on top of.

Example 1: the '90' appears the most number of times. It appears 5 times but we do not need to use this '5' anywhere. We just need to know that the '90' appeared the most number of times. Thus the **Mode is 90**. If you look at the histogram, the highest bar sits over the **'90.'**



Example 2: The Mode = 42kg because 42 appeared the most times (7 times). Also, the highest bar is over the 42.



Measures of Center and how they compare

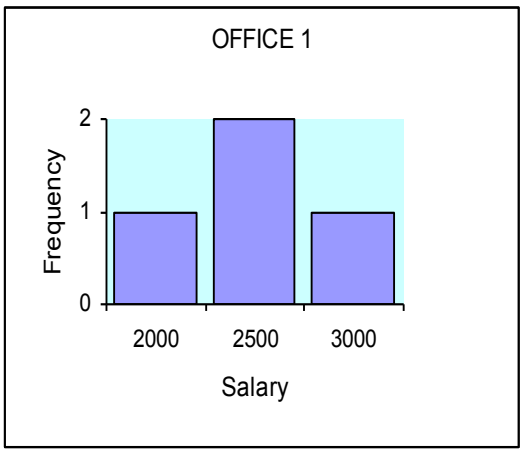
Look at the following set of monthly salaries in two small office groups. The mean, median, and mode are calculated for each one as well as showing the histogram.

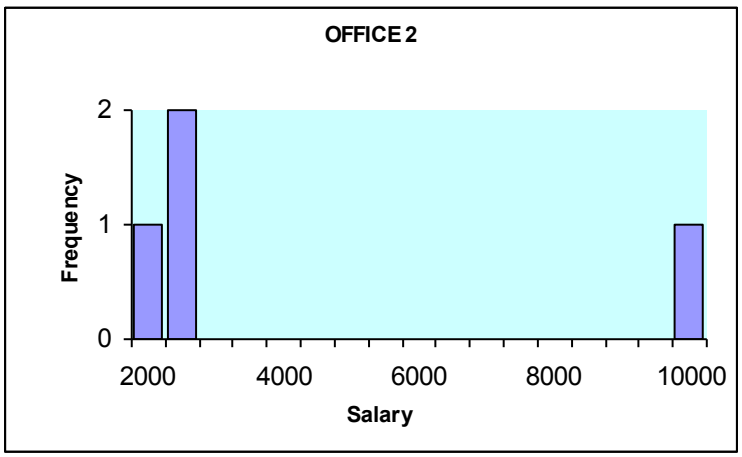
OFFICE 1

Name	*Salary*
Jane	$2,000
Rae	2,500
Rita	2,500
Pat	3,000

OFFICE 2

Name	*Salary*
Rose	$2,000
Carol	2,500
Connie	2,500
Ann	10,000





mean salary = $2,500
median salary = 2,500
mode = 2,500

mean salary = $4,250
median salary = 2,500
mode = 2,500

Notice in OFFICE 1, the mean = median = mode . The shape of the histogram is basically what we would think as 'normal.' It is *symmetric* and intuitively what we would consider 'centered.' Thus, in a NORMAL histogram, the mean, median, and mode are all about equal.

In **OFFICE 2**, however, the highest paid employee (probably the boss), '**skews'** the histogram. The median and the mode remain the same, but the 'mean' has been affected by this high salary. We say that the **'tail'** of the histogram draws the mean in its direction, i.e. in the direction of the 'tail.'

The mean is much more sensitive to extreme values than any of the other descriptors. Look at the median salary in both cases. It has not been affected. The median is a useful measure of center or average when the data are very skewed. For example, think of the average home price in your area. The median gives a much better representation of average home values than the mean as the values of very expensive homes are likely to bring the 'average' or 'mean' up. That's because the mean uses the most arithmetic where the median only uses the 'ordering' of the data values and not their magnitude. [1]

[1] The information for this page has been taken from the Core Curriculum Quantitative Reasoning Requirement Data Text, by Gleason, Molay, Hallett, et. al., 1991.

Measures of Dispersion or 'measures of spread'

We have studied Measures of Center: Mean, Median, and Mode. These descriptors gave us different ways of 'measuring' the center of a distribution. Now we move on to the study of Measures of Dispersion: the range and standard deviation. Both these measures give an idea on how the data is 'spread' out.

The range is simply the highest data value minus the lowest. It is helpful in some respects but not others. Take for example, two towns whose average temperature is 50°. You would feel differently about these two towns if one had a 'range' of temperatures of 100° (say highest was 105° and lowest 5°) and the other town a 'range' of 40° (perhaps the highest was 50° and the lowest 10°, or the highest, 100° and lowest 60°).

But the range is not very helpful in telling us *how* the data is spread out in between these high and low extremes. We may wish to get a better sense of how the data is clustered together or where the bulk of the data are. That's where the Standard Deviation will come in handy. First, let's see how the range is calculated.

Range = Largest minus smallest data value.

Example 1. Here's our data from Example 1.
It's organized in a 'frequency table.'

Data Value	How Many
60	1
70	2
80	3
90	5
100	4

The largest Data Value (not frequency) is 100 (Bottom number in the **Data Value** column.)

The smallest Data Value is 60.

Range = Largest minus smallest *Data Value* thus

Range = 100 − 60 = **40**

Example 2. *Prized Pumpkin Weights (kg)*

Weight	Frequency
14	2
21	5
28	4
35	6
42	7

Range = 28 since largest pumpkin size −smallest is 42 − 14 = 28.

The other way of measuring spread is the standard deviation. It gives us an idea of 'how far away' the bulk of the data is 'from the mean.' In order to calculate this standard deviation we must use the mean, figure out how each individual data value 'deviates' from the mean, and then take the average of these deviations. Here's how we calculate both these *measures of dispersion*. We use our previous examples to demonstrate.

Standard Deviation - The average of deviations of each data value from the mean**.**

The formula is $\sigma = \sqrt{\frac{\Sigma(x - \bar{x})^2}{n}}$

Example 1. Let's use the same data from example 1 to demonstrate how to calculate the standard deviation. Here's the frequency table yet again.

Data Value	**How Many**
60	1
70	2
80	3
90	5
100	4

First: Set up a **table** as follows: the symbol $\bar{x}$ represents the mean and **x** itself represents the data value. **n** = total frequency. The top of each column in the table explains what that column is. Notice **Column A** is simply the 'Data Value column' from the Frequency table, and **Column D** is the 'How Many' column from the Frequency Table above.

We had the mean: $\bar{x} = 86$ and **n = 15. NOTE: The asterisk * means to multiply.**

Data Value **A**	Deviation **B**	Deviation^2= (Column B)^2 **C**	Frequency **D**	Column C*D **E**
x	$x - \bar{x}$	$(x - \bar{x})^2$	**freq**	$(x - \bar{x})^2 * f$
60	60-86= –26	$(-26)^2=676$	1	676
70	70–86=−16	$(-16)^2=256$	2	512
80	80–86=−6	36	3	108
90	90−86=4	16	5	80
100	100–86= 14	196	4	784
			n=15	**Σ = 2160**

Next, to find the standard deviation, do the following steps:

Step 1: Sum the last column - this is symbolized by the **Σ** so Σ=2160

Step 2: Divide this total by the total frequency (the sum of Column D = **n**) $\frac{2160}{15}$

Step 3: Take the Square Root of this answer. Written in the formula it looks like this:

$$\sigma = \sqrt{\frac{2160}{15}} = \sqrt{144} = 12$$

Let's calculate the standard deviation for the pumpkin data.

Example 2. *Prized Pumpkin Weights (kg)*

Weight	**Frequency**
14	2
21	5
28	4
35	6
42	7

$$\textbf{Mean} = \frac{2(14)+5(21)+4(28)+6(35)+7(42)}{24} = \frac{28+105+112+210+294}{24} = \frac{749}{24} = 31.2$$

$\bar{x}$ **= 31.2 kg**

The mean weight of this group of pumpkins is 31.2 kg. (I rounded the mean to 1 decimal place for this example.)

Data Value **A**	Deviation **B**	Deviation^2= (Column B)^2 **C**	Frequency **D**	Column C*D **E**
x	$\mathbf{x - \bar{x}}$	$\mathbf{(x - \bar{x})^2}$	**freq**	$\mathbf{(x - \bar{x})^2 * f}$
14	14–31.2= –17.2	$(-17.2)^2$=295.84	2	295.84*2=591.68
21	21–31.2= –10.2	$(-10.2)^2$=104.04	5	104.04*5=520.20
28	28–31.2= –3.2	10.24	4	40.96
35	35–31.2= 3.8	14.44	6	86.64
42	42–31.2= 10.8	116.64	7	816.48
			n=24	**Σ = 2055.96**

$$\sigma = \sqrt{\frac{2055.96}{24}} = \sqrt{85.665} = \textbf{9.2553. . .} \quad \textbf{Rounds to 9.3}$$

In order to memorize the process, you **must** set up a few of these tables. Try the "Examples with solutions for practice" on the next page yourself.

Examples with solutions for practice: Try to duplicate the results at home. (These are set up to give you practice doing a standard deviation calculation. They are especially important so you will remember to take the square root of your answer near the end of your calculation. Carry out to two decimal places. **(Not to be handed in.)**

Given: A certain organization at a well-known university generally surveys students in each major course at the end of each term. Students were asked, "On a scale of 1 (awful) to 5 (great) how would you rate this course overall? The results of three such surveys are given below.

Notice: The Data here are small values. Think of them, if you get confused, as a grade for one problem where the teacher only gives a 1,2,3,4, or a 5 for a grade on that problem.

Calculate: the mean, median, mode (if there is one), range, and standard deviation of the responses in the following data sets.

Ex**ample 1**: Biology 4303: Frogs and You

Response	Frequency
1	19
2	3
3	1
4	2
5	20

Results: n = 45

Shape = (may be considered bimodal)
Mean = 3.02
Median = 3
Mode = ----(No value really stands out over another.)
Range = 4
Standard Deviation = 1.89

Example 2: Philosophy 1881: The Purpose of Thought

Response	Frequency
1	3
2	12
3	21
4	14
5	4

Results: n = 54

Shape = normal
Mean = 3.07
Median = 3
Mode = 3
Range = 4
Standard Deviation = 0.997 = 1.0 (rounded)

Example 3: Math 46,917.2: Abstract and Useless Mathematical Elements

Response	Frequency
1	2
2	14
3	5
4	15
5	2

Results: n = 38

Shape = bimodal
Mean = 3.03
Median = 3
Mode = ---
Range = 4
Standard Deviation = 1.09