Discrete Quantitative Data

Example: Cars are sampled from the end of the production line and inspected. To save time and money, not all cars are inspected. Below you see data on the number of blemishes (minor flaws) found on the body of a newly produced automobile. For the use of making generalizations about all cars, we would hope that the sample is taken randomly. Let’s assume the cars are identified by Vehicle Identification Number (VIN – every car has one). A data table looks like this:

<table>
<thead>
<tr>
<th>Car VIN</th>
<th>Number of blemishes</th>
</tr>
</thead>
<tbody>
<tr>
<td>19XY</td>
<td>3</td>
</tr>
<tr>
<td>39C7</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>85T2</td>
<td>1</td>
</tr>
</tbody>
</table>

The units are the cars and the variable is the number of blemishes. Notice that many of the units (cars) have the same value of the variable (number of blemishes). This typically characterizes a discrete variable: there are a good number of ties.

In our introductory studies, we will often see data arrayed in a format different from that of the data table. Like this:

3 1 2 2 7 3 4 1 2 4 2 3 4 3 3 1

Why? First, it can save space. Second, we will assume for the most part that our data is accurately measured and recorded, so that information on the units is unnecessary. (If we have data we are suspicious of, we’d want to track down the unit and confirm our value(s).)

The formal definition of a discrete variable is "one for whom a list of possible values can be systematically written out." In practice, discrete quantitative variables are almost always¹ associated with counts. In practice, large data sets of discrete variables tend to have many ties. (Variables for which ties are in theory impossible - and in practice rare - are continuous variables. Physical measurements such as height, weight, and time, are typical continuous variables.)

In statistics a good deal of effort is spent characterizing the nature of the unit-to-unit variability in whatever variable we are studying. A general description of this variability is called the distribution (of the variable). Distributions detail what values occur and how often they occur.

The first step in dealing with any variable’s distribution is to "make a picture." The appropriate picture here is the histogram. There are two, essentially equivalent, varieties of histograms: frequency and relative frequency.

List values of the variable $x$ from the minimum to the maximum, in the minimum increment the data could possibly differ by. Here, data can't differ by more than 1 (usually the case for counts), so we list from 1 to 7. (It would be OK to start at 0. We would probably agree that if we looked

¹ But not always. For instance, at a blackjack table, you might bet $10. The list of possible profits for you is: -$10, 0, $10, $15 (where a -$10 profit is a loss of $10). This is discrete, but not related to a count.

Discrete Data
at more data, there are likely to be cars with no blemishes.) Find the frequency $f$ of values taking each of the values (second row of the table).  

<table>
<thead>
<tr>
<th>number of blemishes $x$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency of cars $f$</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Plot these, as shown. Label the axes in words. (The horizontal ($x$) axis is labeled with the variable description; the vertical ($y$) axis is frequency of units.)

The data are plotted to the scale. That is: Even though there are no 5s or 6s in this data set, we include them to form a complete, sensible scale. This highlights the extreme nature of the single observation of 7. (It also admits that if 7 is a legitimate data value, 5 and 6 are probably going to be seen in a larger sample of cars.)

The observation of 7 is a (statistical) outlier.

To determine the relative frequency / proportion $p$ for a value $x$ of the variable, divide the frequency by $n$, the total number of observations (= units).  

In this situation there are $n = 16$ cars. Divide each frequency by 16. This is done in the third column of the table below.

<table>
<thead>
<tr>
<th>number of blemishes ($x$)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency of cars ($f$)</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>relative frequency of cars $p$</td>
<td>0.1875</td>
<td>0.2500</td>
<td>0.3125</td>
<td>0.1875</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0625</td>
</tr>
</tbody>
</table>

While the relative frequencies could be rounded (say to the nearest 0.01 = 1%), it is wise to keep them to the nearest 0.0001. These will be used in later calculations. Overly rounding them and then performing and combining multiple calculations on these rounding errors, may lead to substantial errors in final results.

Plotting relative frequencies rather than frequencies in a histogram is shown below. Notice the histogram has the same shape - only the vertical scale is changed.

There are "gaps" between the "bars" of the histogram. This is not necessary, but is a nice touch. It hints that only values equal to integers - and not anything between two consecutive integers - are possible. Of course the context of the problem makes this clear - and that context is expressed in the appropriately phrased label on the horizontal axis. Not all software will "gap" the "bars," so getting axes labels right makes sense.

It is permissible to express the vertical axis in term of %s. Change the labels accordingly.

---

2 If you’ve done it right, then the first row is always a statement of what the variable is. In the second row is always “frequency of units.”
3 Relative frequency and proportion are synonyms. The symbol $p$ is used because it is shorter than $rf$, and because it is extended later to “probability.”
4 $n$ is “sample size” – the number of units for which the value of the variable is recorded.
Summary measures for a distribution:

**Mean:** Sum the data, divide by the number of observations.

Example: 
\[
\text{Sum} = 3 + 1 + 2 + 2 + 7 + 3 + 4 + 1 + 2 + 4 + 2 + 3 + 4 + 3 + 3 + 1 = 45
\]

\[
\text{Mean} = \frac{45}{16} = 2.8125 \text{ blemishes}
\]

For the time being we’ll use the symbol \( \bar{x} \) (x-bar) to denote (sample) mean.

In mathematics the greek (capital) letter \( \sum \) ("sigma") stands for “sum.” If we use \( x \) to indicate the value of a variable when measured on a unit, then

\[
\bar{x} = \frac{\sum x}{n}.
\]

Mean is a measure of center. We’ll see precisely what is meant by “center” when we discuss standard deviation. A way to visualize the mean is to place a mark on the horizontal (x) axis of the histogram at the mean: The histogram balances over this mark.

The measurement units of the mean are the same as that of the data. The mean does not have to be one of the data values. No car has 2.8125 blemishes.

**Mode:** The most frequently occurring value. There can be multiple modes.

Example: 

\[ \text{Mode} = 3 \text{ blemishes} \]

**Range:** Subtract the minimum from the maximum.

Example: 

\[ \text{Range} = 7 - 1 = 6 \text{ blemishes} \]

Range is a measure of spread or variation or variability. The measurement units of the range are the same as that of the data.

Now let’s talk about the most commonly used measure of variability in statistics, the standard deviation. To introduce it, we will work with a much smaller data set. Five students were asked to guess their instructor’s age. Here are the guesses: 36, 43, 43, 42, 39.

**Variance:** To compute variance (variance and standard deviation are related) mostly by hand, it’s good to organize your data in a table and perform these steps.

1. Obtain the Mean.

---

5 The symbol \( \mu \) is used for a population mean. Similarly, capital \( N \) (rather than small \( n \)) is used for population size. However, the computation is identical: Sum and divide by how many. The two symbols differentiate quantities that are conceptually distinct. If we knew the mean number of blemishes for all cars, the symbol \( \mu \) would correspond to that population mean. The difference is this: The population mean does not vary; a sample mean varies depending on which sample is selected.

6 In statistics, the terms “spread,” “variability,” “variation,” and “dispersion” are synonyms. Be careful however: Variance is more specific: The variance is a particular way to measure of variability.

7 We consider the sample variance here. There is a slight adjustment for obtain the variance for data from a census: divide by the number of observations, rather than one fewer…
1. Determine, for each value, the deviation from the Mean.
2. Square each of these deviations
3. Sum these squares
4. Divide this sum by one fewer than the number of observations to get the Variance

<table>
<thead>
<tr>
<th>Data</th>
<th>Deviation</th>
<th>Deviation $^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>36</td>
<td>$36 - 40.6 = -4.6$</td>
<td>$(-4.6)^2 = 21.16$</td>
</tr>
<tr>
<td>43</td>
<td>$43 - 40.6 = 2.4$</td>
<td>$2.4^2 = 5.76$</td>
</tr>
<tr>
<td>43</td>
<td>2.4</td>
<td>5.76</td>
</tr>
<tr>
<td>42</td>
<td>1.4</td>
<td>1.96</td>
</tr>
<tr>
<td>39</td>
<td>1.6</td>
<td>2.56</td>
</tr>
</tbody>
</table>

Variance = $37.2 / 4 = 9.3$ years$^2$.

We use the symbol $S^2$ for a sample variance. $S^2 = 9.3$ years$^2$.

Back to the mean for a second: The signed (positive and negative) deviations sum to 0 for any data set. In other words: The sum of distances below the mean is equal to the sum of distances above the mean. This is one property of the mean – and why it measures, in one sense, “center.” The mean “balances” the data on either side in terms of distances – which is consistent with the idea of the mean as a balance point for a histogram.

Variance is a measure of spread or variability, but has *squared* measurement units. (The variance in the example is 9.3 years squared.) Of course no one thinks in years squared…

**Standard Deviation**: The symbol for the standard deviation is $S$. It is the square root of the sample variance.

For the example, $S = \sqrt{9.3} = 3.05$ years.

Standard deviation is a measure of spread, or variability, with the same measurement units as the data (in the example, years: this is a good thing, and is why standard deviation is almost always reported out instead of the variance).

The mean and standard deviation each have measurement units identical to that for the data (in the example, both are in years).

Note that 3.05 is “representing” for the entire set of deviations (without regard to $\pm$).

The concise mathematical expression of the standard deviation looks like this:

$$S = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

You generally want to use the built in statistics capabilities of your handheld calculator or computer to obtain the standard deviation. Computing the standard deviation is tedious for a large data set. Computations required for the blemishes data are indicated below.

---

8 The *sample* standard deviation. Again there is a slightly different method for obtain the standard deviation (and variance) when one has data describing an entire population.
Blemishes data

<table>
<thead>
<tr>
<th>Data</th>
<th>Deviation</th>
<th>Deviation²</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3 – 2.8125 = +0.1875</td>
<td>0.1875² = 0.0352</td>
</tr>
<tr>
<td>1</td>
<td>1 – 2.8125 = -1.8125</td>
<td>(-1.8125)² = 3.2852</td>
</tr>
<tr>
<td>2</td>
<td>2 – 2.8125 = -0.8125</td>
<td>(-0.8125)² = 0.6602</td>
</tr>
<tr>
<td>2</td>
<td>2 – 2.8125 = -0.8125</td>
<td>(-0.8125)² = 0.6602</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>1</td>
<td>1 – 2.8125 = -1.8125</td>
<td>(-1.8125)² = 3.2852</td>
</tr>
</tbody>
</table>

SUM 0.0000 34.4375

Variance = 34.4375 / 15 = 2.296

\[ S = \sqrt{2.1523} = 1.515. \]

A Rule of Thumb

The standard deviation is not a naturally intuitive notion. It is only through experience that you will begin to get a handle on it. In time you should be able to anticipate roughly where the data are by merely knowing the mean and standard deviation. The following rule is almost never exactly true, and is occasionally quite far from true, but it reasonably accurate in most cases.

Data sets tend to have:

- about 68% of the data within 1 standard deviation of the mean.
- about 95% (19/20ths) of the data within 2 standard deviations of the mean
- almost all of the data within 3 standard deviations of the mean.

Let’s try this out with the blemishes data. The mean is 2.81 and the standard deviation is 1.52.

Values within 1 standard deviation of the mean are within 1.52 of 2.81. 1.52 below 2.81 is 2.81 – 1.52 = 1.29; 1.52 above 2.81 is 2.81 + 1.52 = 4.33. So being within 1.52 of 2.81 is equivalent to being between 1.29 and 4.33. Here’s the data, with the values within one standard deviation of the mean underlined.

\[ \begin{array}{cccccccccccccccc}
3 & 1 & 2 & 2 & 7 & 3 & 4 & 1 & 2 & 4 & 2 & 3 & 4 & 3 & 3 & 1 \\
\end{array} \]

12 of the 16 values = 75%. 75% of the values are within one standard deviation of the mean. While this is not exactly 68%, it’s not that far off.

Values within 2 standard deviation of the mean are between 2.81 – 2×1.52 = -0.23 and 2.81 + 2×1.52 = 5.85. All but one datum (the 7) satisfy this. 15/16 = 93.75% - not too far from the rule of thumb’s 95%.

All of the data are within 3 standard deviations of the mean.

In the early stages of your studies, you should check on this repeatedly.
(Optional) Mean & Standard Deviation from Relative Frequency Tables

When there are duplicate data values, the standard deviation computation involves a lot of duplicated computations. If you organize the data in a relative frequency display, it is much quicker to obtain both the mean and standard deviation, by taking advantage of these duplications.

To find the mean

1. Determine $xp$ (multiply the value by the relative frequency) for each value.
2. Sum these. This sum is the Mean.

This is demonstrated below for the blemishes data. The relevant computations for the mean are shown in the 4th column.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$f$</th>
<th>$p$</th>
<th>$xp$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>0.1875</td>
<td>$3(0.1875) = 0.5625$</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>0.2500</td>
<td>$2(0.2500) = 0.5000$</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>0.3125</td>
<td>$3(0.3125) = 0.9375$</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>0.1875</td>
<td>$4(0.1875) = 0.7500$</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0.0000</td>
<td>$5(0) = 0.0000$</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0.0000</td>
<td>$6(0) = 0.0000$</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.0625</td>
<td>$7(0.0625) = 0.4375$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1.0000</td>
<td>$MEAN = 2.8125$</td>
</tr>
</tbody>
</table>

$MEAN = 2.1825$ blemishes

To determine the variance use the following steps.

1. Determine the deviation from the mean for each value
2. Square theses deviations from the mean
3. Multiply each squared deviation by the relative frequency
4. Sum the squared deviation $\times$ relative frequency entries.
5. Multiply this sum by $\left(\frac{n}{n-1}\right)$. This gives the variance.
Since $n = 16$, multiply 2.1523 by 16/15 to get the variance: \( \frac{16}{15} \times 2.1523 = 2.296 \)

The standard deviation is the square root of the variance: $SD = \sqrt{2.296} = 1.515$ blemishes.
Exercises

1. For the matinee showings of films on the final Tuesday in August, a cinema has collected data on the number of paying customers for the last 6 years.

<table>
<thead>
<tr>
<th>Year</th>
<th># Customers</th>
<th>Deviation</th>
<th>Deviation^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011</td>
<td>54</td>
<td>54 - 68.5 = -14.5</td>
<td>(-14.5)^2 = 210.25</td>
</tr>
<tr>
<td>2010</td>
<td>61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2009</td>
<td>65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2008</td>
<td>114</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2007</td>
<td>60</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2006</td>
<td>57</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SUMS\rightarrow

a) Identify the units and the variable. What is the value of n?

b) Confirm that the mean # of customers is 68.5.

c) For each unit (each row of the table), determine the deviation from the mean. Fill in the 3rd column in the table with this information. The first row is done for you.

d) What do the deviations in the third column sum to?

e) For how many of the 6 years was attendance above the mean? Below?

f) Square the deviations from the mean. Fill in the 4th column in the table with this information. The first row is done for you.

g) Use the squared deviations to determine the variance. Remember: Sum them and then divide by one fewer than the number of observations n. What symbol is used for the variance?

h) Take the square root of the variance to determine the standard deviation. What symbol is used for the standard deviation?

i) Guess the year for which there was poor weather on the last Tuesday in August? Why did you choose as you did?

j) Now, using software (in a spreadsheet; or in a statistical program if you have one) or your calculator’s built-in statistics capabilities, enter the 6 data values and compute the sample mean and standard deviation. The results should agree with your answers to b and h.

2. 25 SUNY Oswego students were asked: “How many children do you think you will have?”

3 2 0 2 4 2 3 4 2 2 2 3 3 1 2 2 2 0 2 2 0 3 3 3

a) Construct a histogram using relative frequencies. Do so neatly, and using appropriate words in the axes labels. A blank graph is provided below.

b) Determine values for the mode, mean, range and standard deviation.
c) What % of the data fall within one standard deviation of the mean? Begin by computing \( \bar{x} - s \) and \( \bar{x} + s \) (nearest 0.01). Then count how many of the observations fall between these two values. According to the rule of thumb, this percent is typically around 68%.

d) What % of the data fall within two standard deviations of the mean? Begin by computing \( \bar{x} - 2s \) and \( \bar{x} + 2s \) (nearest 0.01). Then count how many of the observations fall between these two values. According to the rule of thumb, this percent is typically around 95%.

e) What % of the data fall within three standard deviations of the mean? Begin by computing \( \bar{x} - 3s \) and \( \bar{x} + 3s \) (nearest 0.01). Then count how many of the observations fall between these two values. According to the rule of thumb, almost all the data should fall between these values.

3. Students at Brigham Young University were asked “How many children do you think you will have?” Below you see a frequency table.

<table>
<thead>
<tr>
<th>number of children ( x )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>frequency of students ( f )</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>11</td>
<td>15</td>
<td>12</td>
<td>8</td>
<td>4</td>
<td>2</td>
</tr>
</tbody>
</table>

| relative frequency \( p \) | 0.0 | 0.1 | 0.3 | 0.7 | 1.1 | 1.5 | 1.2 | 0.8 | 0.4 | 0.2 |

a) How many students were surveyed?

b) Compute relative frequencies (nearest 0.0001 for each).

c) Construct a relative frequency histogram at right. Do so neatly, and using appropriate words in the axes labels.

d) Determine the mode, mean, and range.

4. What fundamental differences are there between SUNY Oswego and Brigham Young students on the question of “How many children do you think you will have?” Notice that the scales are laid out for the histogram to allow for easy visual comparison.

5. Find the mean and then standard deviation for the following data: 84, 128, 90, 94. Are the data split equally above and below the mean? If not, how then is the mean a measure of center? Explain. (Hint: For each observation below the mean, determine how far below. Sum these. Now compare to how far above the mean for the remaining observation.)
6. Quizzes are given in six classes (scores range from 0 to 8). On page 12 you see a partially completed grid of histograms for quiz scores for the six classes.

Mostly we’re investigating standard deviation here. Remember: Standard deviation measures variability (or spread) – the unit to unit variability in the data.

For Class A, here’s the data: 3, 2, 1, 0, 4, 2, 4, 7, 6, 1, 5, 8, 4, 2, 3, 6, 3, 4, 5, 7, 5, 6, 3.

a) Obtain the relative frequency table for Class A. Draw the frequency histogram in the grid.

b) Determine the mean and standard deviation for Class A.

c) Among classes A, B and C (look at the histograms – think about what lists of data would look like too), which do you think has the most variability in quiz scores? Which has the least? Or are they all the same? Explain.

d) Between Classes D and E which do you think has more variability in quiz scores? Explain.

e) For Class E Complete the frequency / relative frequency table:

<table>
<thead>
<tr>
<th>x</th>
<th>f</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

f) Determine the mean and standard deviation for Class E. You’ll have to list all the data. These computations don’t depend on the order of the data, so you can simply calculate using this 0 0 0 0 0 0 4 4 4 4 4 4 4 4 4 4 4 4 4 8 8 8 8 8 8 (there are six 0s, twelve 4s and six 8s).

g) Do scores in Class F tend to have more or less variation than those for the other classes?

h) The following table displays the standard deviation for each class (for all of them Mean = 4.00):

<table>
<thead>
<tr>
<th>Class</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>St Dev</td>
<td>2.04</td>
<td>3.05</td>
<td>2.63</td>
<td>3.33</td>
<td>2.89</td>
<td>1.20</td>
</tr>
</tbody>
</table>

Now you can check your standard deviation computations, and your expectations about which class(es) have more / less variability. (Larger standard deviation implies more variability.) Based on these values, do you need to rethink any of your answers about comparisons among standard deviations? If “Yes,” then start thinking! (It’s most helpful to think about standard deviation in terms of the very tedious computations required to do it by hand from a unit-by-unit list of data.)
Histogram of Class A (25 students)

Histogram of Class B (29 students)

Histogram of Class C (27 students)

Histogram of Class D (27 students)

Histogram of Class E (24 students)

Histogram of Class F (26 students)
Finally: Here’s the data for Class G.

<table>
<thead>
<tr>
<th>x</th>
<th>f</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

i) Plot the frequency histogram for Class G.

j) How do you expect the means Class G and Class F to compare? (Which is larger? Or – are they the same?)

k) How does the variability for Class G compare to that for Class F? (Which is larger? Or – are they the same?)

l) Compute the mean and standard deviation for Class G. Compare these to the values for Class F.

m) For data set A only, determine the percent of values within 1, 2 and 3 standard deviations of the mean. Compare these to the rule of thumb which suggests 68%, 95% and (almost) 100% as typical values. Complete the table that follows.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Rule of Thumb</th>
<th>Data Set A</th>
</tr>
</thead>
<tbody>
<tr>
<td>% w/in 1 SD of Mean</td>
<td>≈68%</td>
<td></td>
</tr>
<tr>
<td>% w/in 2 SDs of Mean</td>
<td>≈95%</td>
<td></td>
</tr>
<tr>
<td>% w/in 3 SDs of Mean</td>
<td>≈100%</td>
<td></td>
</tr>
</tbody>
</table>
Solutions

1. a) The units are the “last Thursdays in August of a year”. The variable is “number of paying customers.” \( n = 6 \).

   “Number of paying customers varies from last Thursday in August of one year to that for another year.”

   \[
   \begin{array}{ccc}
   \text{Year} & \# \text{ Customers} & \text{Deviation} \\
   \hline
   2010 & 54 & 54 - 68.5 = -14.5 \\
   2009 & 61 & -7.5 \\
   2008 & 65 & -3.5 \\
   2007 & 114 & 45.5 \\
   2006 & 60 & -8.5 \\
   2005 & 57 & -11.5 \\
   \hline
   \text{SUM} & & 0.0
   \end{array}
   \]

   \[
   \begin{array}{ccc}
   & \text{ Deviation}\^2 \\
   (-14.5)^2 & = 210.25 \\
   -7.5^2 & = 56.25 \\
   -3.5^2 & = 12.25 \\
   45.5^2 & = 2070.25 \\
   -8.5^2 & = 72.25 \\
   -11.5^2 & = 132.25 \\
   \hline
   \text{SUM} & & 2553.5
   \end{array}
   \]

   d) These sum to zero (they always do).

   e) 1 above the mean (by 45.5); 5 below (by a total of 45.5).

   g) The variance is \( 2553.5/5 = S^2 = 510.7 \) customers\(^2 \) (yes: customers squared).

   h) The square root of 510.7 is \( S = 22.60 \) customers.

   i) 2007. When weather is poor – especially in the afternoon for matinee performances – movie theaters do better, because people seek out indoor entertainment.

2. a) Here is the relative frequency table. The histogram is shown as part of the solution for #3 given below.

   \[
   \begin{array}{c|ccccc}
   \text{number of children } x & 0 & 1 & 2 & 3 & 4 \\
   \hline
   \text{relative frequency } p & 0.12 & 0.04 & 0.44 & 0.32 & 0.08 \\
   \end{array}
   \]

   b) The mode is 2, the mean is 2.20, the range is 4, and the standard deviation is 1.08.

   c) Between 1.12 and 3.28. That’s 19 of 25 or 76%.

   d) Between 0.04 and 4.36. That’s 22 of 25 or 88%.

   e) 100%.

3. a) 63. b) Here’s relative frequency table.

   \[
   \begin{array}{c|ccccc}
   \text{number of children } x & 0 & 1 & 2 & 3 & 4 \\
   \hline
   \text{relative frequency } p & 0.0000 & 0.0159 & 0.0476 & 0.1111 & 0.1746 \\
   \end{array}
   \]

   \[
   \begin{array}{c|ccccc}
   \text{number of children } x & 5 & 6 & 7 & 8 & 9 \\
   \hline
   \text{relative frequency } p & 0.2381 & 0.1905 & 0.1270 & 0.0635 & 0.0317 \\
   \end{array}
   \]
c) See the solution for #3 below in the solution for #4.

d). The mean is 5.16. The mode is 5. The range is 8.

4. We see that the distribution for SUNY Oswego students differs in two primary ways:
1) It has a lower center (as measured by the mean), 2) it has considerably less variability (as measured by the range).

When comparing two distributions
   Use the same scale on both axes.
   If the collections of data are of different sizes, use relative frequencies, not frequencies.
   This allows for quick, meaningful, and direct comparison.

5. Mean = 99. The standard deviation is 19.77. The deviations from the mean are -15, 29, -9, -5. The data are not split equally above and below the mean. There are 3 values below the mean and 1 above. In this sense, the mean is not a measure of center. However, if you examine how far below the mean the three values are, and total these distances, you get 15 + 9 + 5 = 29. That’s exactly how far above the mean the remaining value is.

6. Many of the answers are shown in part h.
   a) See at right. b) The mean is 4.00, the standard deviation is 2.04.
   e)

   \[
   \begin{array}{c|cc}
   x & f & p \\
   \hline
   0 & 6 & 0.25 \\
   4 & 12 & 0.50 \\
   8 & 6 & 0.25 \\
   \end{array}
   \]

   f) The mean is 4. The standard deviation is 2.89.
i) See at right. Class G is just Class F shifted 2 units to the left. Scores are lower, but variability in scores is the same.

1) The mean is 2.00 (2 less than for Class F). The standard deviation is 1.20 – exactly as it is for class F.

m) 18 of the 25 values are within 1 standard deviation of the mean (between 4.00 – 2.04 = 1.96 and 4.00 + 2.04 = 6.04). That’s 76% (not that close to 67%, but not that far either). All 25 of the 25 are within 2 standard deviations of the mean (between 4.00 – 2×2.04 = -0.08 and 4.00 + 2×2.04 = 8.08). That’s 100% (compared to 95%; basically a difference of 1 value out of 25, as 24/25 = 96% which is as close as one could get to the 95% with a sample of size 25).

<table>
<thead>
<tr>
<th>Rule of Thumb</th>
<th>Data Set A</th>
</tr>
</thead>
<tbody>
<tr>
<td>% w/in 1 SD of Mean</td>
<td>≈67%</td>
</tr>
<tr>
<td>% w/in 2 SDs of Mean</td>
<td>≈95%</td>
</tr>
<tr>
<td>% w/in 3 SDs of Mean</td>
<td>≈100%</td>
</tr>
</tbody>
</table>