Let \( \Omega \) be a population under consideration and let \( X \) be a specific measurement that we are analyzing. For example, \( \Omega = \) All U.S. households and \( X = \) Number of children (under age 18) living in the household. To study this scenario, we obtain a set of measurements \( \{ x_1, x_2, \ldots, x_n \} \) which may be either a census or simply a random sample.

### Census

In a census, we assume that we have a measurement from every member in the population under consideration. For small populations, such as students in one particular class or players on one sports team, it is not hard to obtain a census by surveying each person in that population. But for extremely large populations, such as all U.S. households, it is nearly impossible to obtain a real census even when mandated to do so every ten years by the United States Constitution.

But when we do have a census of measurements \( X \) from a population, then we can find the true values of the mean \( \mu \), the variance \( \sigma^2 \), the standard deviation \( \sigma \), as well as other population parameters.

### Mean

Given a set of \( n \) measurements \( \{ x_1, x_2, \ldots, x_n \} \), the mean (or average) of these specific values is given by

\[
\mu = \frac{x_1 + x_2 + \ldots + x_n}{n}.
\]

When the values are a census of a specific measurement \( X \) from a population \( \Omega \), then \( \mu \) is true average value. It is also called the expected value of \( X \) and may be denoted by \( \mu_X \) or \( E[X] \).

### Variance

The variance, denoted by \( \sigma^2 \), is the average squared distance from the mean and is given by

\[
\sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \ldots + (x_n - \mu)^2}{n} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2
\]

Alternately, the variance is “the average of the squares” minus “the square of average,” and can be computed by

\[
\sigma^2 = \frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n} - \mu^2
\]

The variance is sometimes denoted by \( \sigma_X^2 \) or \( Var(X) \).
Standard Deviation

We take the square root of the variance to get the standard deviation denoted by $\sigma$:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}.$$

The standard deviation gives a way of measuring the average spread from the mean. A small $\sigma$ means that measurements are consistently close to the average $\mu$.

Median and Mode

When the measurements $\{x_1, x_2, \ldots, x_n\}$ are listed in increasing order, then the median is the “middle” value, or the average of the two middle values if there are an even number of measurements. The mode is the measurement (or measurements) that occurs most often.

Example 1. Below are the number of credit hours enrolled for this semester for all students in one section of MATH 116. Find the mean, variance, standard deviation, median, and mode of these values. What percentage of these measurements are within one standard deviation of class average?

### Credit Hours Taken This Semester

<table>
<thead>
<tr>
<th>18</th>
<th>15</th>
<th>14</th>
<th>13</th>
<th>18</th>
<th>14</th>
<th>14</th>
<th>18</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>16</td>
<td>18</td>
<td>18</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>13</td>
<td>15</td>
<td>18</td>
<td>14</td>
<td>15</td>
<td>18</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>18</td>
<td>14</td>
<td>16</td>
<td>17</td>
<td>16</td>
<td>14</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

Solution. Let $\Omega$ = this specific MATH 116 class and let $X =$ Number of credit hours enrolled in this semester. Because we have a census of this class, we can find the true mean $\mu$ and the true standard deviation $\sigma$. To do so, we shall enter the data into the calculator, sort it into increasing order, and use the 1–Vars Stats command.

```
[1] Enter data into L1
[2] Sort data then enter 1–Vars Stats L1
[3] Output
[4] Scroll down
```

The mean is $\mu = \frac{18 + 15 + 18 + \ldots + 15}{36} = \frac{576}{36} = 16$ credit hrs

Note: The calculator displays this value as $\bar{x}$, which stands for sample mean. But because we have a census of this class and not merely a sample, we use $\mu$ to represent that we have the real average of $\mu = 16$ credit hours.
The variance is computed by

\[ \sigma^2 = \frac{x_1^2 + x_2^2 + \ldots + x_n^2}{n} - \mu^2 = \frac{9324}{36} - 16^2 = 3 \]

Then taking the square root gives us the standard deviation of \( \sigma = \sqrt{3} \approx 1.732 \). The true standard deviation is displayed as \( \sigma \) in the calculator output, and this value is to be used if we have a census of measurements. So now we can say that the class average is 16 credit hours with an average spread from 16 of \( \sigma \approx 1.732 \) credit hours.

The median is the “middle measurement.” But because we have an even number of measurements (36), we must take the average of the middle two measurements. After sorting the 36 values, the middle values are in the 18th and 19th positions. The 18th value is 15 while the 19th value is 16. So the median is \( \frac{15 + 16}{2} = 15.5 \), which is also displayed on the TI.

After sorting the values, it is easy to make a frequency chart from which we see that the mode is 18 hours. That is, in this class more students are registered for 18 hours than for any other number of hours.

```
# Hours | # Students
13      | 2
14      | 6
15      | 10
16      | 3
17      | 3
18      | 11
19      | 1

N = 36
```

To find the pct. within one standard deviation of average, we first compute \( \mu \pm \sigma = 16 \pm 1.732 \), which is about 14.268 to 17.732. So students taking 15, 16, or 17 hours fall in this range. There are \( 10 + 3 + 3 = 16 \) students in this range. Thus, \( 16/36 \), or 44.44% of the students in this class are within one standard deviation of class average.

**Question:** Is this class representative of all students on campus? Representative of just undergraduates? Representative of all students taking a Gen. Ed. math class this semester? Or perhaps representative of just MATH 116 students this semester?

The most we can say is that this class is representative of all MATH 116 students this semester. If you want a sample that is representative of a larger portion of the student body, then you must sample accordingly from among that entire group of students. But you should never take an existing sample and try to say that it is representative of a larger group that was not represented in the sample.
Sample Mean and Sample Deviation

Often a collection of measurements is just a sample from a larger population. In this case, we cannot find the real average $\mu$. Instead we can only compute the sample mean denoted by $\bar{x}$. However, $\bar{x}$ is computed the same way as we computed $\mu$ by adding up the values and dividing by $n$; we just denote it now by $\bar{x}$ to specify that we are only working with a sample.

The sample deviation, denoted by $S$, is computed similarly to $\sigma$; however, we use $\bar{x}$ in the formula, rather than $\mu$, and we average the squared differences by dividing by $n - 1$ rather than $n$.

$$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2}$$

For a census

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

For a sample

By dividing by $n - 1$, the sample variance $S^2$ becomes an unbiased estimator of the true unknown variance $\sigma^2$. That is, the average of all possible $S^2$ from all possible samples of size $n$ will equal the true variance $\sigma^2$.

Quartiles and 1.5 IQR

The first quartile $Q_1$ is the median of just the measurements that are below the overall median. The third quartile $Q_3$ is the median of just the measurements that are above the overall median. These values are displayed, along with the minimum, median, and maximum, in the 1-Vars Stats output. Together, the values min – $Q_1$ – med – $Q_3$ – max make up the five-number summary.

The 1.5 IQR (or 1.5 Interquartile Range) is the interval $Q_1 - 1.5 \times (Q_3 - Q_1)$ to $Q_3 + 1.5 \times (Q_3 - Q_1)$. Values from a sample that are outside this range are called outliers and are often excluded from samples so as not to throw off the average too much.

Example 2. Below are data on city mpg from a sample of two-seater cars:

<table>
<thead>
<tr>
<th>Model</th>
<th>City MPG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acura NSX</td>
<td>17</td>
</tr>
<tr>
<td>Audi TT Quattro</td>
<td>20</td>
</tr>
<tr>
<td>Audi TT Roadster</td>
<td>22</td>
</tr>
<tr>
<td>BMW M Coupe</td>
<td>17</td>
</tr>
<tr>
<td>BMW Z3 Coupe</td>
<td>19</td>
</tr>
<tr>
<td>BMW Z3 Roadster</td>
<td>20</td>
</tr>
<tr>
<td>BMW Z8</td>
<td>13</td>
</tr>
<tr>
<td>Chevrolet Corvette</td>
<td>18</td>
</tr>
<tr>
<td>Chrysler Prowler</td>
<td>18</td>
</tr>
<tr>
<td>Ferrari 360 Modena</td>
<td>11</td>
</tr>
<tr>
<td>Ford Thunderbird</td>
<td>17</td>
</tr>
<tr>
<td>Honda Insight</td>
<td>57</td>
</tr>
<tr>
<td>Honda S2000</td>
<td>20</td>
</tr>
<tr>
<td>Lamborghini Murcielago</td>
<td>9</td>
</tr>
<tr>
<td>Mazda Miata</td>
<td>22</td>
</tr>
<tr>
<td>Mercedes-Benz SL500</td>
<td>16</td>
</tr>
<tr>
<td>Mercedes-Benz SL600</td>
<td>13</td>
</tr>
<tr>
<td>Mercedes-Benz SLK230</td>
<td>23</td>
</tr>
<tr>
<td>Mercedes-Benz SLK320</td>
<td>20</td>
</tr>
<tr>
<td>Porsche 911 GT2</td>
<td>15</td>
</tr>
<tr>
<td>Porsche Boxter</td>
<td>19</td>
</tr>
<tr>
<td>Toyota MR2</td>
<td>25</td>
</tr>
</tbody>
</table>
Use your calculator for the following:

(i) Find the sample mean and sample deviation, the median, the mode, and the five-number summary. What percentage of these mileages are within one sample deviation of sample average?

(ii) Make a histogram with range of [5, 60] divided into bins of length 5. Which bin has the most measurements? The second most?

(iii) Give the $1.5 \times IQR$ and specify the outliers.

Solution. (i) We first enter the data into a list in the STAT EDIT screen. For this problem we shall use L2. After entering the data, we sort the data with the command SortA(L2). Then we compute the statistics with the command 1-Var Stats L2.

Because the data are only a sample of measurements from the population $\Omega$ of all two-seater makes of cars, the value of $\bar{x} \approx 19.59$ is the sample mean. The sample deviation is displayed as $S \approx 9.22$.

The minimum value is shown to be 9 while the maximum value is 57. The median is given as 18.5. That is, 18.5 is the average of the two “middle” measurements when in increasing order (the 11th and 12th with this data set of even-size 22). The 11th value is 18 while the 12th value is 19. So the median is $(18 + 19)/2 = 18.5$. The first quartile is $Q_1 = 16$, which is the median of all values below 18.5. And $Q_3 = 20$, which is the median of all values above 18.5. So the five-number summary is $9 - 16 - 18.5 - 20 - 57$.

By scrolling down the sorted list, we see that the mode is 20 which occurs most often at 4 times.

The range $\bar{x} \pm S$ is $19.59 \pm 9.22$, which is 10.37 to 28.81, contains 20 out of 22 measurements. So we can say that 90.9% of these mileages are within one sample deviation of sample average.

(ii) Adjust the WINDOW and STAT PLOT settings to see a histogram (3rd type). Press GRAPH, then TRACE and scroll to see the bin range values. The range [15, 20] has 9 values while the range [20, 25] has 7 values.
(iii) The 1.5 \textit{IQR} is the interval \( Q_1 - 1.5 \times (Q_3 - Q_1) \) to \( Q_3 + 1.5 \times (Q_3 - Q_1) \), where \( Q_3 - Q_1 = 20 - 16 = 4 \). So the 1.5 \textit{IQR} is 16 - 1.5 \times 4 \) to 20 + 1.5 \times 4 \), or \textbf{10 to 26}. Thus, the outliers are those values outside of this range which are 9 mpg and 57 mpg.

\section*{Frequency Charts}

Often measurements are given in a frequency chart that states how many times each measurement occurs.

\begin{tabular}{|c|c|c|c|c|c|}
\hline
Measurement & \( x_1 \) & \( x_2 \) & \( x_3 \) & \ldots & \( x_m \) \\
Frequency & \( k_1 \) & \( k_2 \) & \( k_3 \) & \ldots & \( k_m \) \\
\hline
\end{tabular}

Now we let \( n = k_1 + \ldots + k_m \) = total number of measurements. Then the mean \( \mu \) is actually a \textit{weighted average} given by

\[ \mu = \frac{k_1 x_1 + \ldots + k_m x_m}{n}. \]

When using the calculator, enter the measurements into one list and enter the frequencies into another list.

\textbf{Example 3.} A survey on the number of children per household was taken throughout a neighborhood. Here are the results from the \textit{sample} that was obtained.

\begin{tabular}{|c|c|c|c|c|c|c|}
\hline
Number of children & 0 & 1 & 2 & 3 & 4 & 5 & 6 \\
Number of households & 60 & 42 & 86 & 59 & 22 & 4 & 2 \\
\hline
\end{tabular}

(i) Find the mean and deviation, the median, the mode, and the five-number summary for the number of children in this sample of households. What percentage of these households are within a deviation of average?

(ii) Make a histogram with bins of length 1. (iii) Give the 1.5 \textit{IQR} and state the outliers.

\textbf{Solution.} Here \( \Omega = \text{All households in this neighborhood} \) and \( X = \text{Number of children in household}. \) We shall use list \texttt{L3} for the measurements and list \texttt{L4} for the frequencies, then enter the command \texttt{1-Var Stats L3, L4}.

\begin{itemize}
\item [(i)] Because we have a \textit{sample}, \( \bar{x} \approx 1.86 \) children with \( S \approx 1.34 \); the median is 2 and the mode is 2. The five-number summary is 0 – 1 – 2 – 3 – 6.
\end{itemize}
Next we compute \( \bar{x} \pm S = 1.86 \pm 1.34 \), which is 0.52 to 3.2. This range includes all households having 1, 2, or 3 children. There are \((42 + 86 + 59) = 187\) out of 275 such households, or 68% within a sample deviation of sample average.

The 1.5 IQR is from \(1 - 1.5 \times 2\) to \(3 + 1.5 \times 2\), or -2 to 6; thus, there are no outliers because all measurements are within this range.

**Exercise 1.** Consider the Verbal ACT scores from a group of English majors at WKU:

\[16, 18, 20, 21, 22, 23, 24, 25, 26, 27, 30, 34\]

(a) Make a histogram with range \([15, 36]\) and bins of length 3. Which bin range has the most scores?

(b) Assuming this group is the entire population under consideration:

(i) Find the true mean.  
(ii) Find and explain the median and the mode.  
(iii) Find the true standard deviation.  
(iv) Compute the percentage of these students whose Verbal ACT score is within a standard deviation of average.

(c) Assuming this group is only a sample from a larger population \(\Omega\):

(i) Find the sample mean and sample deviation. 
(ii) Give the boundaries of the 1.5 IQR and state the outliers. 
(iii) In this case, what is the appropriate larger population \(\Omega\) that this sample could represent?

**Exercise 2.** A group of WKU freshman were asked to give the number of hours taken during their first semester. The results were:

<table>
<thead>
<tr>
<th>Hrs</th>
<th>13</th>
<th>14</th>
<th>14.5</th>
<th>15</th>
<th>15.5</th>
<th>16</th>
<th>16.5</th>
<th>17</th>
<th>18</th>
</tr>
</thead>
<tbody>
<tr>
<td># Fr</td>
<td>4</td>
<td>5</td>
<td>8</td>
<td>14</td>
<td>8</td>
<td>23</td>
<td>12</td>
<td>18</td>
<td>8</td>
</tr>
</tbody>
</table>

(a) Make a histogram with bins of length 1. Which bin range has the most values?

(b) Assuming this group is the entire population under consideration:

(i) Find the true mean.  
(ii) Find and explain the median and the mode.  
(iii) Find the true standard deviation.  
(iv) Compute the percentage of these students whose Verbal ACT score is within a standard deviation of average.

(c) Assuming this group is only a sample from a larger population \(\Omega\):

(i) Find the sample mean and sample deviation. 
(ii) Give the boundaries of the 1.5 IQR and state the outliers. 
(iii) In this case, what is the appropriate larger population \(\Omega\) that this sample could represent?
Solutions

1. (a) [21, 24) has 5 scores

(b) (i) $\mu = 23.5$ (ii) Because there are 14 scores, the median is the average of the 7th and 8th scores, which is $(22 + 23)/2 = 22.5$. The modes are 21 and 22 (both occur twice, and no other score occurs more than once).

(iii) $\sigma \approx 4.547$ (iv) $\mu \pm \sigma$ is 18.953 to 28.047 and contains 10/14 or 71.43% of the scores.

(c) Assuming this group is only a sample from a larger population $\Omega$, then

(i) $\bar{x} = 23.5$ and $S \approx 4.719$

(ii) The 1.5 IQR is from $21 - 1.5(26 - 21)$ to $26 + 1.5(26 - 21)$, or 13.5 to 33.5. The only outlier is 34. (iii) $\Omega = All$ English majors at WKU.

2. [16, 17) has 35 scores

(b) (i) $\mu = 15.88$ hours. (ii) Because there are 100 measurements, the median is the average of the 50th and 51st measurement, which is $(16 + 16)/2 = 16$. The mode is also 16 hrs because it occurs most often at 23 times.

(iii) $\sigma \approx 1.1898$. (iv) $\mu \pm \sigma$ is 14.69 to 17.07, which contains all students taking 15, 15.5, 16, 16.5, or 17 hours. Thus, there are 75 out of 100 or 75% of the students within one standard deviation of average.

(c) (i) $\bar{x} = 15.88$ and $S = 1.19578$ (ii) The 1.5 IQR is from $15 - 1.5 \times 2$ to $17 + 1.5 \times 2$, or 12 to 20 which contains all measurements. There are no outliers.

(iii) $\Omega = All$ WKU Freshmen.