3.1 Measures of Central Tendency

Definitions:

**Mean** – the average value of a set of data. The *sample mean*, which is a statistic, is indicated by \( \bar{x} \), and for a sample of \( n \) data points \( x_1, x_2, \ldots, x_n \) is

\[
\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}
\]

The *population mean*, which is a parameter, is indicated by \( \mu \), and for a sample of \( N \) data points \( x_1, x_2, \ldots, x_N \) is

\[
\mu = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}
\]

If the mean is calculated after throwing away the largest and smallest values, then it is called the *trimmed mean*. If a frequency distribution is placed on a seesaw, the mean can be thought of as the position of the fulcrum which balances the seesaw. [See Fig. 2 on p. 119.]

**Example**: Find the mean for the sample data \{20, 13, 4, 8, 10\}.

**Solution**: Using the STAT function on the calculator, we determine that \( \bar{x} = 11 \).
Example: Find the mean for the population data \{3, 6, 10, 12, 14\}.

Solution: Using the same procedure as in the previous example, we determine that \( \mu = 9 \).

Example: A professor has recorded exam grades for 20 students in his class, but one of the grades is no longer readable. If the mean score on the exam was 82 and the mean of the 19 readable scores is 84, what is the value of the unreadable score?

Solution: Let \( x_{20} \) be the unreadable score. Then

\[
\begin{align*}
    x_1 + x_2 + \cdots + x_{19} + x_{20} &= 20(82) \\
    x_1 + x_2 + \cdots + x_{19} &= 19(84)
\end{align*}
\]

Subtracting the 2\(^{nd}\) equation from the 1\(^{st}\) gives

\[ x_{20} = 20(82) - 19(84) = 44 \]

Median – the middle value \( M \) of a set of \( n \) ordered data points \( x_1, x_2, \ldots, x_n \) given by

\[
M = \begin{cases} 
    x_{\frac{n+1}{2}} & \text{if } n \text{ is odd} \\
    \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2} & \text{if } n \text{ is even}
\end{cases}
\]

Example: Calculate the median of the data set \{13, 23, 12, 6, 24, 7, 18\}.

Solution: First, we have to put the data set in order: \{6, 7, 12, 13, 18, 23, 24\}. Since \( n = 7 \) is an odd number of data points, the median is then \( M = x_{(7+1)/2} = x_4 = 13 \).

Example: Calculate the median of the set \{6, 7, 12, 13, 18, 23, 24, 30\}.

Solution: This set is already in numerical order. Since \( n = 8 \) is an even number of data points, the median is the average of \( x_{8/2} = x_4 = 13 \) and \( x_{(8/2)+1} = x_5 = 18 \). Thus \( M = (13 + 18)/2 = 15.5 \). (Note that the median does not have to be an element in the data.
set. This will not be the case for the percentiles and deciles as we shall see later.)

**Mode** – the value in a data set that occurs most often. If no value occurs more than once, then there is no mode. If two values have the same number of occurrences, the data is *bimodal*, and we report the two values. If more than two values have the same number of occurrences, then the data is *multimodal*, and we do not report any values.

*Example*: Find the mode of the data set \{0, 0, 0, 1, 1, 2, 2\}.

*Solution*: The mode is zero.

*Example*: Find the mode of the data set \{0, 0, 1, 1, 2, 2\}.

*Solution*: There is no mode.

*Example*: Find the mode of the data set \{0, 0, 1, 1, 2\}.

*Solution*: The modes are 0 and 1.

**Measure of Central Tendency** – The median is a good measure of central tendency for skewed distributions. This is because the median is more resistant to extreme values, and is a better measure than the mean for central tendency.

*Example*: For the data set \{3, 6, 10, 12, 14\} given in an example above, we found the population mean to be \(\mu = 9\). We can see that \(M = 10\). Suppose \(x_5\) is changed from 14 to 44. How does this affect the mean and median?

*Solution*: The mean becomes \(\mu = 25\), but the median remains \(M = 10\). That is why median is more resistant than mean, and is a better measure of central tendency.

We still prefer the mean as a measure of central tendency for uniform and normal distributions because the techniques used in inferential statistics require use of the mean. For a symmetrical data distribution, the mean, median, and mode are the same. In general, for quantitative data, if the distribution is skewed right, then the mean moves further to the right than the median, while the mode remains at the peak. On the other hand, if the distribution is skewed left, the mean moves further left than the median, and the mode
remains at the peak. Therefore, for a skewed distribution of quantitative data, the median is a better measure of central tendency than the mean. [See Figs. 6 and 9 on p. 121-123, and work 3.1.17-18.] For qualitative data, the mode is a measure of central tendency.

Exercises:


2. Determine the mean, median, and mode from raw data (7 – 16, 19 – 21).

3. Determine the mean and median to identify shape of distribution (17 – 18).
3.2 Measures of Dispersion

Definitions:

Dispersion – the degree to which the data are spread out.

Range – indicated by $R$ is given as

$$R = x_{\text{max}} - x_{\text{min}}$$

Example: Calculate the range of the data \{20, 13, 4, 8, 10\}.

Solution: After putting the data in order we have

$$x_5 - x_1 = 20 - 4 = 16$$

Range is not a good measure of spread because it does not use all observations.

Deviation – the distance of an observation from the mean, $x_i - \overline{x}$, or $x_i - \mu$. Unlike the range, the sum of the deviations uses all observations, but is not a good measure of dispersion since

$$\sum_{i=1}^{n} x_i - \overline{x} = 0 = \sum_{i=1}^{N} x_i - \mu$$

Example: Calculate the sum of the deviation of the previous data set.

Solution: First we calculate the mean as $\overline{x} = 11$, then we have the deviation at the bottom of column three in the table below.

<table>
<thead>
<tr>
<th>$i$</th>
<th>$x_i$</th>
<th>$(x_i - \overline{x})$</th>
<th>$(x_i - \overline{x})^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>-7</td>
<td>49</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>-3</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>-1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>9</td>
<td>81</td>
</tr>
<tr>
<td>Sum</td>
<td></td>
<td>0</td>
<td>144</td>
</tr>
</tbody>
</table>
**Variance** – the sum of the squares of the deviations divided by $N$ for a population and $n - 1$ for a sample. Variance is indicated by $\sigma^2$ for a population and $s^2$ for a sample.

*Example:* Calculate the variance for the previous data set.

*Solution:* This is data set is treated as a sample, so using column four of the table above we have

$$s^2 = \frac{\sum_{i=1}^{n} x_i - \bar{x}^2}{n-1} = \frac{144}{4} = 36$$

**Degrees of Freedom** – the number of observational values that may be arbitrarily chosen for a given mean; i.e. $n - 1$ for a sample of $n$ observations.

*Example:* Choose a sample size $n$ such that $\bar{x} = 10$.

*Solution:* We are free to choose any $n - 1$ numbers, but the $n^{th}$ one is fixed by the equation

$$x_n = \bar{x}n - \sum_{i=1}^{n-1} x_i$$

Therefore, when $\bar{x}$ is specified, as in determining $s^2$, there are $n - 1$ degrees of freedom.

**Standard Deviation** – the square root of the variance. Standard deviation is used in conjunction with the mean to numerically describe normal distributions. The mean locates the center, and the standard deviation describes the thickness of the bell shape. The area under the distribution curve sums to one.

*Example:* Calculate the standard deviation for the previous data set.

*Solution:* Once the variance is determined, the standard deviation is simply

$$s = \sqrt{s^2} = 6$$

**Empirical Rule** – for a normal distribution, 68% lie within 1 standard deviation, 95% within 2 standard deviations, and 99.7% within 3 standard deviations. In other words, 68% of
the data lies in the interval \((\mu - \sigma, \mu + \sigma)\), 95% lies in the interval \((\mu - 2\sigma, \mu + 2\sigma)\), and 99.7% lies in the interval \((\mu - 3\sigma, \mu + 3\sigma)\). [See Fig. 13 on p. 138. Work 3.2.29-31.]

**Chebyshev’s Theorem** – for any distribution, within \(k\) standard deviations, that is, within the interval \((\mu - k\sigma, \mu + k\sigma)\), there are at least

\[
\left(1 - \frac{1}{k^2}\right)100\%
\]

observations. Note that Chebyshev’s theorem makes no guarantee for \(k = 1\). [Work 3.2.35-36.]

**Exercises:**

2. Compute the range, variance, and standard deviation of a variable from raw data (5 – 28).
3. Use the empirical rule to describe data that are bell shaped (29 – 34).
4. Use Chebyshev’s inequality to describe any set of data (35 – 36).
3.3 Measures of Central Tendency and Dispersion from Grouped Data

**Grouped Data** – the mean and standard deviation may be calculated from a frequency distribution with \( k \) classes of data by the formulae

\[
\mu = \bar{x} = \frac{\sum_{i=1}^{k} x_i f_i}{\sum_{i=1}^{k} f_i}
\]

\[
s^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 f_i}{\sum_{i=1}^{k} f_i - 1}, \quad \sigma^2 = \frac{\sum_{i=1}^{k} (x_i - \mu)^2 f_i}{\sum_{i=1}^{k} f_i}
\]

where \( x_i \) is a value of a variable and \( f_i \) is that value’s frequency. Note that for the sample standard variance \( s^2 \), we divide by the degrees of freedom.

**Example:** Calculate the mean and standard deviation of the sample with the following distribution. According to the Empirical Rule, 68\% of the data will be between what two numbers?

<table>
<thead>
<tr>
<th>data</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>16</td>
</tr>
<tr>
<td>3.5</td>
<td>13</td>
</tr>
<tr>
<td>5.5</td>
<td>24</td>
</tr>
<tr>
<td>7.5</td>
<td>19</td>
</tr>
<tr>
<td>9.5</td>
<td>30</td>
</tr>
</tbody>
</table>

**Solution:** We first construct a table containing all the data values \( x_i \) and frequencies \( f_i \) and the terms needed in the equations above. This table is shown on the next page. We then can write
Chapter II

I Numerically Summarizing Data

3-9

\[
\bar{x} = \frac{\sum_{i=1}^{5} x_i f_i}{\sum_{i=1}^{5} f_i} = \frac{629}{102} = 6.167
\]

\[
\sum_{i=1}^{5} x_i - \bar{x}^2 f_i = \frac{818.67}{101} = 8.106 \Rightarrow
\]

\[s = 2.847\]

<table>
<thead>
<tr>
<th>(x_i)</th>
<th>(f_i)</th>
<th>(x_i f_i)</th>
<th>((x_i - \bar{x})^2)</th>
<th>((x_i - \bar{x})^2 f_i)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.5</td>
<td>16</td>
<td>24.0</td>
<td>21.78</td>
<td>348.44</td>
</tr>
<tr>
<td>3.5</td>
<td>13</td>
<td>45.5</td>
<td>7.11</td>
<td>92.44</td>
</tr>
<tr>
<td>5.5</td>
<td>24</td>
<td>132.0</td>
<td>0.44</td>
<td>10.67</td>
</tr>
<tr>
<td>7.5</td>
<td>19</td>
<td>142.5</td>
<td>1.77</td>
<td>33.78</td>
</tr>
<tr>
<td>9.5</td>
<td>30</td>
<td>285.0</td>
<td>11.11</td>
<td>333.33</td>
</tr>
<tr>
<td></td>
<td>102</td>
<td>629</td>
<td>42.22</td>
<td>818.67</td>
</tr>
</tbody>
</table>

This may also be done on the TI -83/84 by putting the data into L1, the frequencies into L2, and using the command STAT | CALC | 1-Var Stats L1, L2.

To find the interval containing 68% of the data, we find \(\bar{x} \pm s = 6.167 \pm 2.847\). Hence, 68% of the data is found in the interval (3.320, 9.014).

**Weighted Mean** – the weighted mean \(\bar{x}_w\) of a variable is calculated in the same manner as grouped data, but using weights instead of frequencies.

**Example:** Calculate the grade point average of a student who earned a B in a 5-hour calculus class, an A in a 3-hour social work class, an A in a 4-hour
biology class, and a C in a 3-hour American literature class.

**Solution:** The categories are 4 for A, 3 for B, 2 for C, 1 for D, and 0 for F. The weights are the class hours. Entering the data on the calculator and using 1-Var Stats L1, L2 gives

Reading the mean, the GPA is 3.27.

**Example:** In a course, attendance counts for 5% of the grade, quizzes count for 10%, exams count 60%, and the final exam 25%. If a student scored 100% on attendance, 93% on quizzes, 86% on exams, and 85% on the final exam, what is the student’s course average?

**Solution:** The categories are the student’s scores, and the weights are the percentages of the grade due to attendance, quizzes, exams, and the final converted to proportions.

The student’s course average is 87.15%.

**Exercises:**

1. Compute the weighted mean (9 – 12).
3.4 Measures of Position and Outliers

**Percentile** - a data value that locates a specified percentage of the data. For the \( k \)th percentile, written \( P_k \), \( k \% \) of the data is less than or equal to \( P_k \).

*Example:* What is meant by the following statements? (a) The 15th percentile of the head circumference of males 3 to 5 months of age is 41.0 cm. (b) The 90th percentile of the waist circumference of females 2 years of age is 52.7 cm.

*Solution:* (a) 15% of males 3 to 5 months old have head circumferences of 41 cm or less. (b) 90% of two-year-old females have waist circumferences of 52.7 cm or less.

**Quartile** – divides the data into 4 parts. The second quartile \( Q_2 \) locates two-fourths or one-half of the data and is just the median \( M \). The first quartile \( Q_1 \) locates one-fourth of the data and is the median of the lower half of the data. The third quartile \( Q_3 \) locates three-fourths of the data and is the median of the upper half of the data.

*Example:* For the data set \( \{1, 2, 5, 5, 5, 6, 7\} \), determine the quartiles.

*Solution:* There are 7 data points, so the median is item number \((7 + 1)/2 = 4\). Therefore, \( Q_2 = x_4 = 5 \). There are 3 data points in the lower half, so the median of that portion is item number \((3 + 1)/2 = 2\). Therefore, \( Q_1 = x_2 = 2 \). There are also 3 data points in the upper half, whose first item number is item number 5. So the median of that portion is item number \((3 + 1)/2 = 2\), the second item number of the upper half, which is the sixth item number of the data set. Therefore, \( Q_3 = x_6 = 6 \).

*Example:* For the data set \( \{10, 21, 33, 45, 56, 63, 75, 81, 101\} \), determine the quartiles.

*Solution:* There are 10 data points, so the median is the average of item numbers 5 and 6. Therefore, \( Q_2 = (x_5 + x_6)/2 = (56 + 63)/2 = 59.5 \). [Note that unlike deciles and percentiles, the quartiles need not be a member of the data set.] There are 5 data points.
in the lower half, so the median of that portion is item number 3. Therefore, \( Q_1 = x_3 = 33 \). There are also 5 data points in the upper half, so the median of that portion is item number 8. Therefore, \( Q_3 = x_8 = 75 \).

**Inter-Quartile Range** – the range of the middle 50% of the data; \( IQR = Q_3 - Q_1 \). The IQR is the preferred measure of dispersion over the standard deviation when the data is skewed or has outliers. The disadvantage of the IQR is that it is a range, and therefore does not include all data values like \( s \).

<table>
<thead>
<tr>
<th>shape</th>
<th>normal</th>
<th>skewed</th>
</tr>
</thead>
<tbody>
<tr>
<td>center</td>
<td>( \mu, \bar{x}, M )</td>
<td></td>
</tr>
<tr>
<td>spread</td>
<td>( \sigma, s, IQR )</td>
<td></td>
</tr>
</tbody>
</table>

**Outlier** – an extreme observation outside the fences:

Lower fence = \( Q_1 - 1.5(IQR) \)

Upper fence = \( Q_3 + 1.5(IQR) \)

**Example**: 3.4.21 (b-d), 3.4.23

**z-Score** – the number of standard deviations of a data point from the mean, given by

\[
z = \frac{x - \bar{x}}{s} \quad \text{or} \quad z = \frac{x - \mu}{\sigma}
\]

**Example**: 3.4.5, 3.4.13

**Exercises**:

2. Determine and interpret z-scores (5 – 14).
3. Explain a percentile (15 – 17).
4. Determine and interpret quartiles; check a set of data for outliers (18 – 31).
3.5 The Five-Number Summary and Boxplots

Definitions:

**Five-Number Summary** – the points $x_{\text{min}}, Q_1, M, Q_3,$ and $x_{\text{max}}$.

**Box Plot** – a plot of the quartiles depicted on p. 166. The whiskers do not extend beyond the fences, which are marked with brackets. Outliers are marked with asterisks. If $Q_2$ is left of center and the right whisker is substantially longer than the left one, the distribution is skewed right. If $Q_2$ is right of center and the left whisker is substantially longer than the right one, the distribution is skewed left.

Example: 3.5.7, 3.5.12

Exercises:


2. Compute the five-number summary; draw and interpret box plots (3 – 18).