Chapter Two: Descriptive Methods (continued)
It is important to be able to quantify the degree of spread or scatter in a data set. Measures of this sort are referred to as measures of variability or dispersion. There are many such measures. We will consider four of these, (a) range, (b) mean deviation, (c) variance, and (d) standard deviation. While all of these are useful in given circumstances, the last two are by far the most important and will form the foundation of many of the methods you will study.
The range is a function of only the largest and smallest scores in a data set. Two forms of the range are often identified. Namely, the exclusive and inclusive range. The exclusive range is the more often used form.
The **exclusive range** is defined as the difference between the largest and smallest scores in the data set or more formally

\[
\text{Range (exclusive)} = x_L - x_S
\]

where $x_L$ and $x_S$ are the largest and smallest scores in the data set respectively.
The exclusive range of the numbers 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5 is

\[ 5 - 3 = 2 \]
The **inclusive range** takes into account the upper and lower real limits of the highest and lowest scores and is expressed as

\[
\text{Range (inclusive)} = URL_L - LRL_S
\]

where \(URL_L\) and \(LRL_S\) are the upper real limit of the largest and lower real limit of the smallest scores in the data set respectively.
The inclusive range of the numbers 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5 is

\[ 5.5 - 2.5 = 3 \]
Like the range, the mean deviation is a highly intuitive measure of variability. Unlike the range, however, mean deviation takes into account all the data for which variability is to be assessed thereby making it a more stable statistic.
As with some other measures of variability as well as some other statistics, mean deviation is based on what are termed **deviation scores** or simply **deviations**. A deviation score is defined as

\[ x - \bar{x} \]

where \( x \) is the score whose deviation is to be calculated and \( \bar{x} \) is the data set mean. Obviously, a deviation gives the number of units between a score and the data set mean. When data are closely “clumped” around the mean, deviations tend to be small. For data that are more spread out, deviations will be larger.
Plausibly, a reasonable representation of variability could be based on the average of these deviations. When data are more spread out, the average of the deviations would be larger than for data with less spread. The difficulty with using deviations in this manner is that they always sum to zero. That is,

\[ \sum (x - \bar{x}) = 0 \]

which makes the mean always equal to zero as well.
The summation to zero problem can be overcome by taking the absolute values of the deviations. This then is the rationale for mean deviation. **Mean deviation** ($MD$) is then the average of the absolute values of the deviations of a set of scores. Or,

$$MD = \frac{\sum |x - \bar{x}|}{n}$$
To find the mean deviation of the scores 9, 3, 3, and 1, we first note that

\[ \bar{x} = \frac{\sum x}{n} = \frac{16}{4} = 4 \]

We now calculate the deviation scores and their absolute values as shown.

| (1) \( x \) | (2) \( x - \bar{x} \) | (3) \( |x - \bar{x}| \) |
|---|---|---|
| 9 | 5 | 5 |
| 3 | -1 | 1 |
| 3 | -1 | 1 |
| 1 | -3 | 3 |

\[ \sum \quad 16 \quad 0 \quad 10 \]
The mean deviation is then

$$MD = \frac{\sum |x - \bar{x}|}{n} = \frac{10}{4} = 2.5$$
**Variance** is a less intuitive but generally much more useful measure of variability than is the range or mean deviation. As a descriptive statistic, variance is less appealing than is mean deviation but is generally more useful because of its role in inference as will be seen. Like mean deviation, variance uses deviations as its basis but squares them rather than using absolute values.
The parameter form of variance is given by

\[ \sigma^2 = \frac{\sum (x - \mu)^2}{N} \]

Thus the parameter form of variance is the average of the squared deviations of the scores that make up the population.
The statistic form of variance is given by

\[ s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} \]

Notice that the devisor for the statistic is \( n - 1 \) while that for the parameter is \( N \). This derives from the fact that \( s^2 \) is commonly used as an estimate of \( \sigma^2 \) in inferential settings. It can be shown that if the devisor of \( s^2 \) were \( n \), the resulting estimate would be biased. That is, on average the value of \( s^2 \) would be smaller than \( \sigma^2 \). By dividing by \( n - 1 \) this bias is removed making \( s^2 \) a better estimate of \( \sigma^2 \). This formulation of sample variance is termed a **conceptual form** because the expression shows the deviation score nature of \( s^2 \).
An algebraically equivalent expression for $s^2$ that is useful for computations is given by

\[
s^2 = \frac{\sum x^2 - \left(\frac{\sum x}{n}\right)^2}{n - 1}
\]
To calculate the variance of the values 9, 3, 3, and 1 via the conceptual method we arrange the data as follows.

\[
\begin{array}{ccc}
(1) & (2) & (3) \\
x & x - \bar{x} & (x - \bar{x})^2 \\
9 & 5 & 25 \\
3 & -1 & 1 \\
3 & -1 & 1 \\
1 & -3 & 9 \\
\sum & 16 & 0 & 36 \\
\end{array}
\]
We then calculate

\[ s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} \]

\[ = \frac{36}{3} \]

\[ = 12 \]
To calculate the variance of the values 9, 3, 3, and 1 via the computational method we note that 
\[ \sum x = 9 + 3 + 3 + 1 = 16 \]
and 
\[ \sum x^2 = 9^2 + 3^2 + 3^2 + 1^2 = 100 \]
then

\[
\begin{align*}
    s^2 &= \frac{\sum x^2 - \left( \frac{\sum x}{n} \right)^2}{n - 1} \\
    &= \frac{100 - \left( \frac{16}{4} \right)^2}{4 - 1} \\
    &= \frac{100 - 4}{4 - 1} \\
    &= \frac{96}{3} \\
    &= 12
\end{align*}
\]
Standard deviation is defined as the square root of variance. It follows from the equations for variance that the statistic form of standard deviation can be represented by

\[ s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \]

and

\[ s = \sqrt{\frac{\sum x^2 - (\sum x)^2}{n}} \]
To calculate the standard deviation of the values 9, 3, 3, and 1 via the conceptual method we arrange the data as follows.

<table>
<thead>
<tr>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$x - \bar{x}$</td>
<td>$(x - \bar{x})^2$</td>
</tr>
</tbody>
</table>

| 9   | 5     | 25       |
| 3   | -1    | 1        |
| 3   | -1    | 1        |
| 1   | -3    | 9        |

$\sum$ 16 0 36
We then calculate

\[ s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}} \]

\[ = \sqrt{\frac{36}{3}} \]

\[ = 3.464 \]
To calculate the standard deviation of the values 9, 3, 3, and 1 via the computational method we note that

\[ \sum x = 9 + 3 + 3 + 1 = 16 \]

and

\[ \sum x^2 = 9^2 + 3^2 + 3^2 + 1^2 = 100 \]

then

\[
\begin{align*}
    s &= \sqrt{\frac{\sum x^2 - (\sum x)^2}{n}} \\
    &= \sqrt{\frac{100 - \frac{(16)^2}{4}}{4 - 1}} \\
    &= \sqrt{\frac{100 - 16}{3}} \\
    &= \sqrt{\frac{84}{3}} \\
    &= 3.464
\end{align*}
\]
Measures of relative position are methods that locate the relative positions of observations in a distribution. To this end we will in turn examine percentiles, percentile ranks, and z scores.
There is no standard definition for percentile. We will use the following. A \textbf{percentile} is a point on the scale of measurement below which a specified percentage of the observations are located. By this definition, the (scale based) median can be defined as the fiftieth percentile.

An expression for a given percentile is provided by

\[
P_p = LRL + (w) \left[ \frac{(pr)(n) - cf}{f} \right]
\]

where \(P_p\) represents the \(p\text{th}\) percentile, \(LRL\) is the lower real limit of the interval that contains the \(p\text{th}\) percentile, \(w\) is the width of the interval calculated as the difference between the upper and lower real limits of that interval, \(pr\) is \(p\) expressed as a proportion (i.e., \(p/100\)), \(n\) is the total number of observations, \(cf\) is the cumulative frequency up to the percentile interval and \(f\) is the frequency of that interval.
When computing a percentile, three different scenarios are possible.

- When a percentile interval is identified, the percentile formula is applied.
- When the sought proportion of observations falls below a real limit with the interval above the limit having nonzero frequency, the real limit is taken as the percentile.
- When the sought proportion of observations falls below a real limit with the interval above the limit having zero frequency, the midpoint of the zero frequency interval(s) is taken as the sought percentile.
Find $P_5$ for the data provided below.

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>.5</td>
<td>22</td>
<td>60</td>
</tr>
<tr>
<td>.4</td>
<td>26</td>
<td>38</td>
</tr>
<tr>
<td>.3</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>.2</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>.1</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>.0</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

$P_5$ is the point on the measurement scale below which 5% or (.05)(60) = 3 of the observations fall. Because the cumulative frequency at the upper real limit of .05 is 3, .05 is the 5th percentile.
Example

Find \( P_{20} \) for the data provided below.

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>.5</td>
<td>22</td>
<td>60</td>
</tr>
<tr>
<td>.4</td>
<td>26</td>
<td>38</td>
</tr>
<tr>
<td>.3</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>.2</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>.1</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>.0</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Since \((.2)(60) = 12\) of the observations fall below any point between .15 and .35, the midpoint of .25 is taken as the 20th percentile.
Example

Find $P_{60}$ for the data provided below.

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>.5</td>
<td>22</td>
<td>60</td>
</tr>
<tr>
<td>.4</td>
<td>26</td>
<td>38</td>
</tr>
<tr>
<td>.3</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>.2</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>.1</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>.0</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Because the cumulative frequencies up to .35 and .45 are 12 and 38 respectively, the point below which $(.60)(60) = 36$ observations fall must be in the .35 to .45 interval.
Applying Equation 2.5 gives

\[ P_{60} = LRL + (w) \left[ \frac{(pr)(n) - cf}{f} \right] \]

\[ = .35 + (.10) \left[ \frac{(.60)(60) - 12}{26} \right] \]

\[ = .442 \]
Percentile Rank

While a percentile is a point on the scale of measurement below which a given percentage of observations fall, a **percentile rank** is the percentage of observations that fall below a given point on the scale. Thus, percentiles are points and percentile ranks are percentages. Because of their close relationship, they are often confused.
When the scale point whose percentile rank is sought falls in a nonzero frequency interval, the following equation can be used to find percentile ranks.

\[
PR_P = \frac{100 \left[ \frac{f(P-LRL)}{w} + cf \right]}{n}
\]

where \( P \) is the point on the scale for which the percentile rank is to be calculated, \( LRL \) is the lower real limit of the interval containing \( P \), \( w \) is the width of the interval calculated as the difference between the upper and lower real limits of that interval, \( n \) is the total number of observations, \( cf \) is the cumulative frequency up to the interval containing \( P \) and \( f \) is the frequency of that interval.
When the point falls at the upper (lower) real limit of an interval or in an interval with zero frequency, we use

$$PR_P = 100 \frac{cf}{n}$$

where $cf$ is the cumulative frequency up to the point.
Find the percentile rank of the point .05 for the data provided below.

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>.5</td>
<td>22</td>
<td>60</td>
</tr>
<tr>
<td>.4</td>
<td>26</td>
<td>38</td>
</tr>
<tr>
<td>.3</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>.2</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>.1</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>.0</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Noting that .05 is located at the upper real limit of the $-.05$ to $.05$, interval we calculate

\[
PR_{.05} = 100 \frac{cf}{n} = 100 \frac{3}{60} = 5
\]
Example

Find the percentile rank of the point .25 for the data provided below.

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>.5</td>
<td>22</td>
<td>60</td>
</tr>
<tr>
<td>.4</td>
<td>26</td>
<td>38</td>
</tr>
<tr>
<td>.3</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>.2</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>.1</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>.0</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Noting that .25 is located in an interval with zero frequency, we calculate

\[
PR_{.25} = 100 \frac{cf}{n} = 100 \frac{12}{60} = 20
\]
Example

Find the percentile rank of the point .442 for the data provided below.

<table>
<thead>
<tr>
<th>Score</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>.5</td>
<td>22</td>
<td>60</td>
</tr>
<tr>
<td>.4</td>
<td>26</td>
<td>38</td>
</tr>
<tr>
<td>.3</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>.2</td>
<td>0</td>
<td>12</td>
</tr>
<tr>
<td>.1</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>.0</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Noting that .442 is located in the interval .35 to .45, we calculate

\[
PR_{.442} = 100 \left[ \frac{f(P-LRL)}{w} + cf \right] = 100 \left[ \frac{26(.442-.35)}{.1} + 12 \right] = 60
\]
Unfortunately, data analysts are usually not interested in finding the percentile rank of a point on the measurement scale, but rather want to know the percentile rank of a score or observation. This presents a difficulty since, scores are not points on the scale but rather, represent intervals. How then do you find the percentile rank of a score?
Basically, you must choose a point on the scale to represent the score. Three common choices are used.

1. The lower real limit of the score interval.
2. The midpoint of the score interval.
3. The upper real limit of the score interval.
Percentiles locate points relative to all the observations in a data set. By contrast, $z$ scores locate points relative to the mean of the data. More precisely, a $z$ score indicates the distance and direction of a point from the mean in terms of standard units.
A sample $z$ score is calculated by

$$z = \frac{x - \bar{x}}{s}$$

While the population equivalent is given by

$$Z = \frac{x - \mu}{\sigma}$$
Convert the set of scores 1, 3, 3, and 9 to z scores. Then find the mean and standard deviation of the z scores. The mean and standard deviation of the original data are, respectively, 4 and 3.46. Using these values in Equation 2.2.3 gives

\[ z_1 = \frac{1 - 4}{3.46} = -0.867 \]
\[ z_3 = \frac{3 - 4}{3.46} = -0.289 \]
\[ z_3 = \frac{3 - 4}{3.46} = -0.289 \]
\[ z_9 = \frac{9 - 4}{3.46} = 1.445 \]
Because the sum of these scores is zero, the mean is also zero. With mean zero, Equation 2.15 simplifies to

\[ s = \sqrt{\frac{\sum z^2}{n-1}} = \sqrt{\frac{3}{3}} = 1 \]
Certain aspects of distribution shapes can be characterized numerically. The two most common of these, skew and kurtosis, will be discussed here.
Various methods have been developed to numerically describe the amount of skew (or lack thereof) that characterizes a distribution. Skew is generally defined as the degree of asymmetry in a distribution.
A common measure of skew is given by

\[
\text{Skew} = \frac{\sum z^3}{n}
\]

where \( z \) is the standardized deviation as described by Equation 2.2.3 and \( n \) is the sample size. Stated differently, this expression of skew is simply the average of the cubed \( z \) scores.
Skew (continued)

When this expression results in a negative or positive value, the distribution is said to be negatively or positively skewed respectively. When the value is zero the distribution is said to be symmetric. The following are depictions of each type.
Skew (continued)

![Graph showing skewness types: Symmetric, Negatively Skewed, Positively Skewed](image)

- **Symmetric**: The distribution is balanced on both sides of the center.
- **Negatively Skewed**: The tail is longer on the left side, indicating more frequent data on the right.
- **Positively Skewed**: The tail is longer on the right side, indicating more frequent data on the left.

---

2.6 Numerical Methods (continued)
Calculate skew for the values 1, 3, 3 and 9.

**Solution**

The z scores of the data and their cubes are as follows.

<table>
<thead>
<tr>
<th>Score</th>
<th>z</th>
<th>$z^3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-.867</td>
<td>-0.652</td>
</tr>
<tr>
<td>3</td>
<td>-.289</td>
<td>-0.024</td>
</tr>
<tr>
<td>3</td>
<td>-.289</td>
<td>-0.024</td>
</tr>
<tr>
<td>9</td>
<td>1.445</td>
<td>3.017</td>
</tr>
</tbody>
</table>

\[
\sum z^3 = 2.317 \\
Skew = \frac{\sum z^3}{n} = \frac{2.317}{4} = .579
\]
Kurtosis refers to the peakedness of a distribution relative to the length and size of its tails. Distributions with sharp peaks are said to be **leptokurtic** while those that have flattened middles are said to be **platykurtic**. Kurtosis pertains to distributions with no more than one mode.
Distribution A is more peaked and has longer tails than does distribution B and therefore has greater kurtosis.
Kurtosis (continued)

\[
\text{Kurtosis} = \frac{\sum z^4}{n}
\]

Notice that while skewness was expressed as the average of the cubed \( z \) scores in a distribution, kurtosis is the average of \( z \) scores raised to the fourth power. In general, larger kurtosis values reflect sharper peaks than do smaller values.
Example

Calculate kurtosis for the values 1, 3, 3 and 9.

**Solution**
The $z$ scores of the data with their values raised to the fourth power are as follows.

<table>
<thead>
<tr>
<th>Score</th>
<th>$z$</th>
<th>$z^4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-.867</td>
<td>0.565</td>
</tr>
<tr>
<td>3</td>
<td>-.289</td>
<td>0.007</td>
</tr>
<tr>
<td>3</td>
<td>-.289</td>
<td>0.007</td>
</tr>
<tr>
<td>9</td>
<td>1.445</td>
<td>4.360</td>
</tr>
</tbody>
</table>

\[
\sum z^4 = 4.939
\]

Kurtosis = \[ \frac{\sum z^4}{n} = \frac{4.939}{4} = 1.235 \]