3.0] **Summation Notation**

The uppercase Greek letter Σ, (sigma), is the symbol to indicate “the sum of” (summation). If \( x \) represents data values of a quantitative data set, then \( \Sigma x \) means to add all the values in the set. More generally, suppose that \( \{x_1, x_2, x_3, \ldots, x_n\} \) is a finite set of \( n \) quantitative data values, where \( x_1 \) represents the first listed data value, \( x_2 \) represents the second listed data value, \( x_3 \) represents the third listed data value, ... so on..., and \( x_n \) represents the last listed data value (the \( n^{th} \) value). Then the sum of all values in the set is written as:

\[
\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n
\]

The symbol \( i \) is called the summation index and in this case it indicates to start the sum with the first data value of the set (since \( i=1 \) for \( x_1 \)) and end the sum with the last data value (the \( n^{th} \) value, since \( i=n \) for \( x_n \)). The values 1 and \( n \) are called the **limits of the summation**, (1 is the lower limit; \( n \) is the upper limit).

**Example 1:** Find the sum of all the data values in the set \( \{4, 14, -5, 10, 4, 7, -2, 3\} \)

**Solution:** In this case, \( x_1 = 4, x_2 = 14, x_3 = -5, x_4 = 10, x_5 = 4, x_6 = 7, x_7 = -2, \) and \( x_8 = 3 \)

Since the size of the set is \( n=8 \) and we want the sum of all the data values, then the limits of the summation are 1 and 8. So:

\[
\sum_{i=1}^{8} x_i = 4 + 14 + (-5) + 10 + 4 + 7 + (-2) + 3 = 35
\]

The sum of all data values in the given set is 35.

**Example 2:** Using the same set of the data values \( \{4, 14, -5, 10, 4, 7, -2, 3\} \), find \( \sum_{i=3}^{7} x_i \)

**Solution:** Since the limits of the summation are 3 and 7, we add from the 3\(^{rd} \) (\( x_3 = -5 \)) to the 7\(^{th} \) value (\( x_7 = -2 \)): \( x_3 = -5, x_4 = 10, x_5 = 4, x_6 = 7, x_7 = -2 \)

That is: \( \sum_{i=3}^{7} x_i = x_3 + x_4 + x_5 + x_6 + x_7 = (-5) + 10 + 4 + 7 + (-2) = 14 \)

**REMARK:** Whenever there is no confusion, we will omit the summation limits if we are adding all the values in a given data set. That is, we may write \( \sum x \) instead of \( \sum_{i=1}^{n} x_i \)

**PRACTICE PROBLEMS**

Given the data set \( \{8, -3, 12, 7, 8, 1\} \), compute the following:

- a) \( \sum x \)
- b) \( \sum (3x) \)
- c) \( 3 \sum x \)
- d) \( \sum x - 3 \)
- e) \( \sum (x - 3) \)
- f) \( \sum x^2 \)
- g) \( (\sum x)^2 \)
- h) \( \sum 3x^2 \)
- i) \( \sum (3x)^2 \)
- j) \( (\sum 3x)^2 \)
- k) \( \sum x^2 - 3 \)
- l) \( \sum (x - 3)^2 \)
- m) \( (\sum x - 3)^2 \)
- n) \( \sum \frac{x}{4} \)
- o) \( \sum \frac{x}{4} \)
- p) \( \sum x^2 - (\sum x)^2 \)
- q) \( \frac{\sum x^2 - (\sum x)^2}{6} \)

**ANSWERS:** a) 33; b) 99; c) 99; d) 30; e) 15; f) 331; g) 1,089; h) 993; i) 2,979; j) 9,801; k) 328; l) 187; m) 900; n) 8,25; o) 8.25; p) -758; q) 24.9167
3.1] Descriptive Statistics Measures

1a) Measures of Central Tendency: Central tendency of a set of numerical data is a property of the data to cluster or congregate about a certain value that represents the “center” or “middle” of the distribution. The location of such a value can be visualized if the data are arranged in ascending or descending order. Typical measures of central tendency include:

a1) Arithmetic Mean: The sum of all data values divided by the size of the data set (the number of values in the set of data). It is very sensitive to extreme data values (outliers). The sample mean $\bar{x}$ as an unbiased estimator of the population mean $\mu$. We must distinguish between the two symbols: $\bar{x}$ = sample mean; $\mu$ = population mean.

For samples or for populations, the computation procedure of the mean is the same:
Sample mean: $\bar{x} = \frac{\sum x}{n}$; Population mean: $\mu = \frac{\sum x}{N}$

Example 1: (Sincich/McClave, 12th Ed., page 56, #2.54) A sample of six grade point averages of college students was recorded as follows: $3.2, 2.5, 2.1, 3.7, 2.8, 2.0$. Compute the mean. (Note: $\Sigma x = 16.3$; $n = 6$; $\bar{x} = 2.717$, rounded to the nearest 1000th)

Example 2: Compute the mean of the following population of scores: $7, 12, -2, 8, 1, 14, 9, 1$ (Note: $\Sigma x = 50$; $N = 8$; $\mu = 6.25$)

a2) Median: The value located at the exact middle of the data set, when the data are in ascending (or descending) order. The median is very resistant to outliers. It separates the top half of the data set from the lower half. Unless all data values are equal, the median is less than or equal to 50% of the data and greater than or equal to the other 50% of the data.

Example 3: a) Find the median of the population in Example 2: $7, 12, -2, 8, 1, 14, 9, 1$

3b) Find the median of the sample in Example 1: $3.2, 2.5, 2.1, 3.7, 2.8, 2.0$

Example 4: a) Find the mean and the median of the sample: $88, 89, 90, 2015$

b) Repeat without the outlier 2015. Comment.

a3) Mode: The most frequent (the most repeated) value in a data set.

Example 5: Find the mode of the sample in Example 1: $3.2, 2.5, 2.1, 3.7, 2.8, 2.0$

Example 6: Find the mode of the population in Example 2: $7, 12, -2, 8, 1, 14, 9, 1$

Example 7: Find the mode of the population in Example 4: $7, 12, -2, 8, 1, 14, 9, 1, 8$

Example 8: Find the mode of the sample: $73, 73, -21, 73, -21, -21, -21, 8, 8, 8, 8, 7, 7, 7, 7, 1, 1$

Example 9: Find the mode of the sample: $11, 11, 5, 11, 5, 5, 5, 8, 8, 8, 7, 7, 7, 7, 1, 1, 1, 1, 3, 3, 3, 3$

a4) Midrange: Arithmetic mean of the maximum and minimum values of the data set. The midrange is very sensitive to outliers.

Example 10: Find the midrange of the sample in Example 1: $3.2, 2.5, 2.1, 3.7, 2.8, 2.0$

Example 11: Find the midrange of the population in Example 2: $7, 12, -2, 8, 1, 14, 9, 1$

Remark: It is many times stated that if a set of data is skewed to the left, then the mean is less than the median and the median is less than the mode; if a set of data is skewed to the right, then the mean is larger than the median and the median is larger than the mode. While this is true in some cases, it is not true in general. It is true that in unimodal perfectly symmetric distributions (bell-shaped), the mean, median and mode have the same value.
Example 12: Consider the data set \{1, 2, 3, 4, 4, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 7, 7, 8\}. A) Construct the dot plot to verify that the distribution is skewed to the left (so we expect that the mean is less than the median and the median is less than the mode). B) Find the values of the mean, median and mode. How do they compare?

Example 13: Now consider the data set \{1, 2, 3, 4, 4, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 6, 7, 7, 8\}. A) Construct the dot plot to verify that the distribution is skewed to the left (so we expect that the mean is less than the median and the median is less than the mode). B) Find the values of the mean, median and mode. How do they compare?

Example 14: Consider the data set \{0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 6, 7, 8\}. A) Construct the dot plot to verify that the distribution is skewed to the right (so we expect that the mean is larger than the median and the median is larger than the mode). B) Find the values of the mean, median and mode. How do they compare?

Example 15: Consider the data set \{0, 0, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 6, 7, 8\}. A) Construct the dot plot to verify that the distribution is skewed to the right (so expect that the mean is larger than the median and the median is larger than the mode). B) Find the values of the mean, median and mode. How do they compare?

Example 16: Consider the data set \{0, 1, 2, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 7, 7, 8, 9, 10\}. A) Construct the dot plot to verify that the distribution is mound-shaped symmetric (so expect that the mean, median and mode are equal). B) Find the values of the mean, median and mode. How do they compare?

a5) Other Measures of Central Tendency:
a5.1) Weighted Mean: An arithmetic mean that incorporates different weights assigned to the data values. (NOTE: if the weights are the same \(\frac{1}{n}\) for all data values, the weighted mean equals the arithmetic mean). \(\text{Weighted Mean} = \sum_{i=1}^{n} w_i x_i \); where \(\sum_{i=1}^{n} w_i = 1\)

a5.2) Geometric Mean: The \(n\)th root of the product of the data values. The data values must be strictly positive values. \(\text{Geometric Mean} = \sqrt[n]{x_1 \cdot x_2 \cdots x_n} \); where \(x_i > 0\)

a5.3) Quadratic Mean: (known as the Root Mean Square – RMS): The square root of the arithmetic mean of the squares of the data values. We find the sum of the squares of the data values, divide the sum by the size of the data set \(n\), and take the square root of the result. \(\text{RMS} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}\)

a5.4) Harmonic Mean: The reciprocal of the arithmetic mean of the reciprocals of the data values. This means that we must first find the sum of the reciprocals of all the data values and then divide the size of the data set by that sum. \(\text{HM} = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \); \(x_i \neq 0\)

a5.5) Truncated Mean: (Also known as Trimmed Mean): Arithmetic mean of the remaining data values after the top given percent and the bottom given percent of the original data are discarded. For example, the arithmetic mean of the remaining values after the top 10% and the bottom 10% of the data are discarded for being outliers.
a5.6) **Midhinge**: The arithmetic mean of the upper and lower quartiles. The *upper quartile* is a value that is greater than or equal to 75% of the data values. The *lower quartile* is a value that is greater than or equal to 25% percent of the data values.

\[ \text{Midhinge} = \frac{Q_r + Q_u}{2} \]

a5.7) **Trimean**: (Also known as Tukey’s Trimean – TM): The arithmetic mean of the median and the midhinge. (NOTE: The median is the same as the *middle quartile*, which is the value greater than or equal to 50% of the data values).

a5.8) **Winsorized Mean**: An arithmetic mean in which the top given percent of data values and the bottom given percent of values (usually, extreme values or outliers) are replaced with values that are close to the median.

a5.9) **Distance-Weighted Estimator**: A weighted mean in which the weights of the data values are computed as the inverse mean distance between the data value and the rest of the data values in the set of data.

1b) **Measures of Variation (Variability, Dispersion, Spread)**: This is to measure how spread (or diverse) the data are. If the data are arranged in ascending (or descending) order or if we have a graphical representation of the data (dot plot, histogram, etc.), we may be able to visualize the spread (variability) of the data values, especially when comparing different sets of data. A measure of variability must be a positive number and should have the same units as the original data. The variability has measure zero when all data values are equal (no variation). The higher the value of the variability, the more diverse the data are. Measures of central tendency alone do not provide a complete description of quantitative data sets.

**Example 17**: Display the dot plots and discuss the variability of the sets of data:
a) \{49, 50, 50, 51\};  b) \{0, 49, 51, 100\};  c) \{50, 50, 50, 50\};

Measures of variability include:

b1) **Range**: The difference of the maximum data value minus the minimum data value.

The range is very sensitive to outliers.

**Example 18**: In **Example 1**, find the range of the sampled data: 3.2, 2.5, 2.1, 3.7, 2.8, 2.0

**Example 19**: In **Example 2**, find the range of the population data: 7, 12, −2, 8, 1, 14, 9, 1

b2) **Variance**: Arithmetic mean of the squared deviations from the mean of the data. To compute the variance of a population, divide the sum of the squared deviations by the population size \(N\). To compute the variance of a sample, divide the sum of the squared deviations by one less than the sample size \(n−1\). We must distinguish between the symbols:

Sample variance \(s^2 = \frac{\sum(x-\bar{x})^2}{n-1}\)  Population variance \(\sigma^2 = \frac{\sum(x-\mu)^2}{N}\)

A problem with the variance is that its units are the square of the units of the original data. Because of this, we take the square root of the variance and it gives us the standard deviation, which has the same units as the original data. Analogous to the sample mean \(\bar{x}\), the sample variance \(s^2\) as an unbiased estimator of the population variance \(\sigma^2\).
Alternative computational formulas derive from the above definition formulas:

**Sample Variance**

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

**Population Variance**

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

**Example 20:** In *Example 1*, find the variance of the sampled data: 3.2, 2.5, 2.1, 3.7, 2.8, 2.0 (Note: \(\sum x=16.3; n=6; \sum (x-x)^2=2.148333; \sum x^2=46.43\));

**Example 21:** In *Example 2*, compute the variance of the population data: 7, 12, -2, 8, 1, 14, 9, 1 (Note: \(\sum x=50; N=8; \sum (x-\mu)^2=227.5; \sum x^2=540\))

Again, to compute the sample variance, we divide the sum of the squared deviations \(\sum (x - \bar{x})^2\) by \(n-1\) and not by \(n\). A strong reason is that the division by \(n-1\) makes the sample variance \(s^2\) a better estimator of the population variance \(\sigma^2\) because if we divide by \(n\), the value of the sample variance \(s^2\) under-estimates the value of the population variance \(\sigma^2\).

**b3) Standard Deviation:** Positive square root of the variance. The standard deviation has the same units as the original data. In general (not always), the value of sample standard deviation \(s\) tends to underestimate the value of the population standard deviation \(\sigma\). So, \(s\) is not an unbiased estimator of \(\sigma\). We must distinguish between the two symbols: \(s=\)sample standard deviation; \(\sigma=\)population standard deviation

Computational formulas: \(s = \sqrt{s^2} = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} ; \sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum (x - \mu)^2}{N}}\)

**Example 22:** In *Example 1*, find the standard deviation of the sample: 3.2, 2.5, 2.1, 3.7, 2.8, 2.0

**Example 23:** In *Example 2*, find the standard deviation of the population: 7, 12, -2, 8, 1, 14, 9, 1

**b4) Coefficient of Variation:** For a set of non-negative data values (population or sample), the coefficient of variation (CV) is the standard deviation expressed as a percentage of the mean of the set of data: *Sample CV* = \(\frac{s}{\bar{x}} \cdot 100\%\); *Population CV* = \(\frac{\sigma}{\mu} \cdot 100\%\)

**Example 24:** In *Example 1*, find the CV of the sample: 3.2, 2.5, 2.1, 3.7, 2.8, 2.0

**Example 25:** Compute the CV of the population data: 7, 12, 5, 8, 1, 14, 9, 1

The CV is very useful to compare the variability of sets of data with different units or different scales or with same scales but substantially different means. We could use the standard deviation to compare the variability of sets of data, but the data sets must have the same scales, the same units, and equal means (or means that are not too different in value).
Example 26: A real estate agent needs to know whether appraised values of some properties have more variability than their lot sizes. Suppose that that in the past 2 months appraisal values had mean $110,000 and standard deviation $20,000 while lot sizes had mean 1,400 square feet and standard deviation 400 square feet. Use the CV to determine which data set has higher variability.

Example 27: Sy is an investor who plans to purchase shares of one of two companies, A or B, which are listed in the American Stock Exchange. Neither company offers dividends to stockholders and both companies are rated equally high (by various investment services) in terms of potential growth. So, Sy must consider the volatility (variability) of each of the two stocks to decide in which company he will invest. Sy finds that the mean price per share of Company A stock is $60 over the past six months with standard deviation $12, while the mean price per share of Company B is $18 over the same period of time with standard deviation $3. Use the CV to determine which stock price has greater variability. Which stock is more stable? Which one is more volatile?

b5) Mean Absolute Deviation: Mean of the absolute value of the deviations from the mean of the data. The mean absolute deviation (MAD) has the same units as the original data.

Sample: \[ \text{MAD} = \frac{1}{n} \sum |x_i - \bar{x}| \]

Population: \[ \text{MAD} = \frac{1}{N} \sum |x_i - \mu| \]

Example 28: In Example 1, find the MAD of the sample: 3.2, 2.5, 2.1, 3.7, 2.8, 2.0

Example 29: In Example 2, find the MAD of the population: 7, 12, -2, 8, 1, 14, 9, 1

Many beginning statistics students ask why we use the standard deviation and not just the mean absolute deviation (MAD). A strong reason is that the computation of the MAD involves the use of absolute values and absolute value is not an algebraic operation. You may recall that algebraic operations are defined as any one of the operations of addition, subtraction, multiplication, division, raising to an integer power or to a fractional power, and taking roots. Algebraic operations are performed on algebraic variables, terms, or algebraic expressions similarly to arithmetic operations (4 + 3 = 7 is an arithmetic operation; \( y + y = 2y \) is an algebraic operation). Another strong reason is that the sample MAD is not an unbiased estimator of the population MAD (we will see in inferential statistics). In addition, variances possess the additive property that when we have two independent populations, if we randomly select one value from each population and add them, the variance of such sums equals the sum of the variances of the two populations. This property is crucial in the development of a method for making inferences about the means of two populations. The MAD does not possess this property.

b6) Median Absolute Deviation: Median of the absolute value of the deviations from the arithmetic mean. This is a robust measure of variability in the sense that it is very resistant to extreme data values (outliers).

b7) Interpretation of a Standard Deviation: We may interpret the value of a standard deviation by using general characteristics of sets of data. Some of such characteristics are: (1) the Empirical Rule; (2) Chebyshev’s Theorem.
(1) The Empirical Rule can be stated as follows: for sets of data having a bell-shaped distribution (or an approximate bell-shaped distribution), the following are true:
- Approximately, 68% of the data values lie within one standard deviation of the mean.
- Approximately, 95% of the data values lie within two standard deviations of the mean.
- Approximately, 99.7% of the data values lie within three standard deviations of the mean.

Figure 3.1: The Empirical Rule summarized graphically

The Empirical Rule is based on empirical evidence accumulated over many years of data analysis and applies to sets of data with bell-shaped symmetric distributions. Data values that fall between two standard deviations and three standard deviations above or below the mean may be considered suspicious outliers. Data values that exceed three standard deviations above or below the mean may be considered definite outliers.

Example 30: The heights of adult women (20–40 years old) have mean of about 64 inches and standard deviation 2.4 inches. Assume that the distribution of such heights is approximately bell-shaped symmetric, use the Empirical Rule to estimate the intervals where we can find:
- a) approximately 68% of the heights;
- b) approximately 95% of the heights;
- c) approximately 99.7% of the heights;
- d) which of the heights {55, 57, 60, 70 inches} may be definite outliers?

(2) Chebyshev’s Theorem: for any set of quantitative data and any positive number \( k > 1 \), the proportion of values lying within \( k \) standard deviations of the mean is at least \( 1 - \frac{1}{k^2} \).

In particular, the following are true:
- The percentage of data values within one standard deviation of the mean is undetermined.
- At least \( \frac{3}{4} \) (or 75%) of the data values lie within two standard deviations of the mean.
- At least \( \frac{8}{9} \) (or 88.9%) of the data values lie within three standard deviations of the mean.
**Example 31:** In *Example 30*, suppose that the shape of the distribution of the heights is unknown. Use *Chebyshev’s Theorem* to find the least percentage of heights between:

a) 61.6 and 66.4 inches; b) 59.2 and 68.8 inches; c) 56.8 and 71.2 inches.

Chebyshev’s Theorem was proved by the Russian mathematician Pafnuty L. Chebyshev (1821−1894). It applies to any set of data (with bell-shaped symmetric distribution or not).

Empirical evidence accumulated over all years of data analysis shows that in a very large percentage of data sets (with bell-shaped symmetric distributions or not), about 95% of the values lie within two standard deviations of the mean. This fact, together with the Empirical Rule and Chebyshev’s Theorem, leads to conclude that the standard deviation should not be much larger than ¼ of the range and if the range exceeds six standard deviations, then the data may contain outliers. Given the mean and the standard deviation of a data set, we may estimate the minimum and the maximum data values by using: minimum ≈ $x - 2s$; maximum ≈ $x + 2s$. Similarly, given a data set, we may estimate the standard deviation by dividing the range by 4. We emphasize that these properties do not replace the calculation of the exact value of the standard deviation using the computational formulas described earlier.

**Example 32:** (Sincich/McClave, 12th Ed., page 71, #2.94) The data show a sample of 25 measurements: {7, 6, 6, 11, 8, 9, 11, 9, 10, 8, 7, 7, 9, 10, 7, 7, 7, 9, 12, 10, 10, 8, 6}

a) construct the dot plot to view the shape of the distribution; b) find the mean, median, mode, midrange, range, variance, standard deviation, coefficient of variation; c) count the number of measurements in the intervals $x ± s$, $x ± 2s$, $x ± 3s$. Express each count as a percentage of the total number of measurements; d) compare the percentages found in part c with the percentages given by the Empirical Rule and Chebyshev’s Theorem; e) use the range to obtain a rough approximation for $s$. Does the result compare favorably with the actual value of $s$ found in part b? (Note: $\Sigma x = 206$; $\Sigma x^2 = 1778$)

**Example 33:** (1) For mound-shaped symmetric distributions of scores, what can we say about the percentage of scores contained in the following intervals: (a) $\overline{x} - s$ to $\overline{x} + s$? (b) $\overline{x} - 2s$ to $\overline{x} + 2s$? (c) $\overline{x} - 3s$ to $\overline{x} + 3s$? (2) For non-symmetric distributions of scores, what can we say about the percentage of scores contained in the following intervals: (a) $\overline{x} - s$ to $\overline{x} + s$? (b) $\overline{x} - 2s$ to $\overline{x} + 2s$? (c) $\overline{x} - 3s$ to $\overline{x} + 3s$?

**1c) Measures of Relative Standing:** quantitative description of the location of a data value relative to the rest of the data or quantitative description of the relationship of a data value to the rest of the data. We will briefly discuss *z-scores* and *percentiles*.

**c1) z-score** (or *standardized score*): the number of standard deviations that a given data value $x$ is away from the mean. A data value with positive *z-score* is above (greater than) the mean of the distribution. A data value with negative *z-score* is below (less than) the mean of the distribution. A data value with *z-score* zero is identically equal to the mean of the distribution.

**z-score Computation Formula** for a *sample* data value $x$: \[ z = \frac{x - \overline{x}}{s} \] (round to nearest 100th)

**z-score Computation Formula** for a *population* data value $x$: \[ z = \frac{x - \mu}{\sigma} \] (round to nearest 100th)
**Example 34:** IQ scores have a mean of 100 and a standard deviation of 15. Compute and interpret the following IQ scores: 

- **a)** 125; 
- **b)** 80; 
- **c)** Find the IQ score with \( z = 2.24 \); 
- **d)** Find the IQ score with \( z = -0.78 \).

**Example 35:** Suppose that 10 and 100 are scores of a sample set of data and their respective \( z \)-scores are –2 and 3. Find the sample mean and standard deviation. If not possible, explain why.

A \( z \)-score is dimensionless (has no units of measurement). It simply measures the relative standing (position) of a data value with respect to the rest of the values in the data set. A large positive \( z \)-score indicates that the data value is larger than all or almost all other values in the data set. A large negative \( z \)-score indicates that the data value is smaller than all or almost all other values in the data set. A \( z \)-score of zero or near zero indicates that the data value is equal to the mean or it is near the mean of the data set.

Since \( z \)-scores do not include any units, they are also useful to compare individual data values from different sets of data, even if the values are measured on different scales with different units.

In addition, based on the **Empirical Rule** and the definition of \( z \)-scores, we can say that for sets of data having a bell-shaped distribution (or an approximate bell-shaped distribution), the following are true:

- Approximately, 68% of the data values have \( z \)-scores between –1 and +1.
- Approximately, 95% of the data values have \( z \)-scores between –2 and +2.
- Approximately, 99.7% of the data values have \( z \)-scores between –3 and +3.

Data values with \( z \)-scores between –3 and –2 or between +2 and +3 may be considered **suspicious outliers** (unusual values). Data values with \( z \)-scores less than –3 or greater than +3 may be considered **definite outliers** (or **highly unusual values**).

**Figure 3.2:** The **Empirical Rule** in terms of \( z \)-scores

\[
\begin{align*}
-3 & \quad -2 \quad -1 \quad 0 \quad +1 \quad +2 \quad +3 \\
-0.15 & \quad 2.35 \quad 13.5 \quad 34 \quad 34 \quad 13.5 \quad 2.35 \quad 0.15
\end{align*}
\]

\[
\begin{align*}
\sim 68\% & \\
\sim 95\% & \\
\sim 99.7\%
\end{align*}
\]
**Example 36:** Xin scored a 72 on her first statistics exam. Her husband Kāluòsī scored 83 on his first statistics exam (same course, same professor, different class section). Xin’s class mean exam score was 64 with standard deviation 3. Kāluòsī’s class mean exam score was 76 with standard deviation 4. Who had better class standing?

**Example 37:** Example 26 described a real estate agent trying to determine whether appraised values of some properties have more variability than their lot sizes. In a sample from the past 2 months appraisal values had mean $110,000 and standard deviation $20,000 while lot sizes had mean 1,400 square feet and standard deviation 400 square feet. We used the coefficient of variation (CV) to determine which data set had higher variability. Now we want to compare individual values from the two data sets. Which of the following values is more extreme: $145,000 in the set of appraisal values or 2,300 square feet in the set of lot sizes?

**Example 38:** (Triola, 12th Ed., page 124, #9) The Wechsler Adult Intelligence Scale measures IQ scores with a test designed so that the mean is 100 and the standard deviation is 15. Consider the group of IQ scores that are unusual; a) what are the z-scores that separate the unusual IQ scores from those that are usual? b) what are the IQ scores that separate the unusual IQ scores from those that are usual?

**c2) Percentiles:** When a set of quantitative data is arranged in ascending or descending order, the $k^{th}$ percentile (denoted $P_k$) is a number whose value is greater than or equal to $k\%$ of the data values in the set and less than or equal to $(100-k)\%$ of the data values in the set. So, the value of $P_k$ separates the lowest $k\%$ of the data values from the highest $(100-k)\%$ of the data values. Percentiles (denoted $P_1,P_2,P_3,...,P_{99}$) partition a set of quantitative data into 100 ranked groups or categories and each group contains 1% of the data values in the set.

**Estimate of Percentile rank of a data value $x = \frac{\text{Number of values less than or equal to } x}{\text{Size of the data set}} \times 100**

Round to the nearest whole number.

**Estimate the value of $P_k$ in a data set:** a) The data must be arranged in ascending order; b) Compute $i = \frac{k}{100} \times (1 + \text{size of the data set})$; c) If $i$ is a whole number, the value of $P_k$ is located at position $i$. If $i$ is not a whole number, the value of $P_k$ equals the data value at the position given by the $\text{int}(i)$ plus the decimal part of $i$ multiplied times the difference of the data values at positions $\text{int}(i)$ and $\text{int}(i)+1$.

**Example 39:** Scores on a statistics final exam were {15, 24, 26, 45, 48, 53, 55, 55, 58, 60, 64, 66, 67, 68, 69, 70, 75, 80, 82, 84, 89, 90, 91, 98}. A) Find the percentile rank of 53; B) Find the percentile rank of 89; C) Find $P_{55}$; D) Find $P_{30}$

**Example 40:** For the class of 2013, the mean mathematics SAT score was 514. A) Interpret: the 67% percentile is 570; B) If 30% of all scores are more than 580, give the percentile rank for the score 580; C) If 89% of all scores fall below 670, give the percentile rank for the score 670.
c3) **Quartiles**: Quartiles are particular types of percentiles (denoted $Q_1, Q_2, Q_3$) that partition a set of quantitative data into four ranked groups or categories and each group contains 25% of the data values in the set.

- $Q_1$ is the 25th percentile. It is called **First Quartile** or **Lower Quartile** with symbol $Q_L$. That is, $Q_1 = P_{25} = Q_L$
- $Q_2$ is the 50th percentile. It is called **Second Quartile** or **Middle Quartile** with symbol $Q_M$. That is, $Q_2 = P_{50} = Q_M$. The **Middle Quartile** is the **Median** of the data set.
- $Q_3$ is the 75th percentile. It is called **Third Quartile** or **Upper Quartile** with symbol $Q_U$. That is, $Q_3 = P_{75} = Q_U$

**Example 41**: In Example 39, the data show 24 scores on a statistics final exam {15, 24, 26, 45, 48, 53, 55, 55, 58, 60, 64, 66, 67, 68, 69, 70, 75, 80, 82, 84, 89, 90, 91, 98}. Find $Q_L, Q_M, Q_U$.

**Example 42**: Investigate the definitions of **quantiles**, **fractiles**, **tertiles**, **quintiles**.

c4) **Other Descriptive Statistics Measures Using Quartiles and Percentiles**: In addition to the descriptive statistics measures discussed earlier, there are other statistics that are defined in terms of quartiles and percentiles:

- **Midhinge** (or **Midquartile**): $\frac{Q_L + Q_U}{2}$
- **Interquartile Range** (or **IQR**): $Q_U - Q_L$
- **Semi-Interquartile Range**: $\frac{Q_U - Q_L}{2}$
- **10–90 Percentile Range**: $P_{90} - P_{10}$

**Example 43**: In Example 39, the data show 24 scores on a statistics final exam {15, 24, 26, 45, 48, 53, 55, 55, 58, 60, 64, 66, 67, 68, 69, 70, 75, 80, 82, 84, 89, 90, 91, 98}. A) Compute the midhinge; B) Compute the IQR; C) Compute the semi-IQR; D) Compute the 10–90 percentile range.

c5) **Five-Number Summary**: Given a set of quantitative data, the **five-number summary** consists of the five values: the minimum data value, $Q_L, Q_M, Q_U$, the maximum data value.

**Example 44**: Find the five-number summary for the set of 24 exam scores in Example 39, {15, 24, 26, 45, 48, 53, 55, 55, 58, 60, 64, 66, 67, 68, 69, 70, 75, 80, 82, 84, 89, 90, 91, 98}.

c6) **Box-Whisker Plots** (or **boxplots**): A **box-whisker plot** (or boxplot) is a diagram that shows the **five-number summary** as follows: one **whisker** (straight line segment) emanates from each of two parallel sides (hinges) of a rectangular box; the starting point of the lower whisker corresponds to the minimum data value; the ending point of the lower whisker (at the lower hinge) corresponds to $Q_L$; the starting point of the upper whisker (at the upper hinge) corresponds to $Q_U$; the ending point of the upper whisker corresponds to the maximum data value; the line between the lower and the upper hinges corresponds to $Q_M$.

**Example 45**: Use the **five-number summary** from Example 44 to construct the boxplot.

**Solution**:

<table>
<thead>
<tr>
<th>Value</th>
<th>15</th>
<th>53.5</th>
<th>66.5</th>
<th>81.5</th>
<th>98</th>
</tr>
</thead>
<tbody>
<tr>
<td>(minimum)</td>
<td>(Q_L)</td>
<td>(Q_M)</td>
<td>(Q_U)</td>
<td>(maximum)</td>
<td></td>
</tr>
</tbody>
</table>
Example 46: Use the given boxplot to find: a) the five-number summary; b) the midquartile; c) the IQR; d) the semi-IQR

8 14 17 22 30

46

(7) Inner Fences and Outer Fences: boxplots include two more components that are very useful for detecting outliers.

> Lower Inner Fence = \( Q_L - 1.5(IQR) \)

> Upper Inner Fence = \( Q_U + 1.5(IQR) \)

> Lower Outer Fence = \( Q_L - 3(IQR) \)

> Upper Outer Fence = \( Q_U + 3(IQR) \)

Detecting Outliers: Data values falling between the inner and outer fences are considered suspicious outliers (or unusual values). Data values falling beyond outer fences are considered definite outliers (or highly unusual values). To reflect the presence of fences and outliers, we construct a modified boxplot. The suspicious outliers are shown as a tiny unshaded circle (●). The definite outliers are shown as an asterisk (*).

Boxplots are also useful for comparing different data sets. The values in the data sets must be on the same scale for comparisons to be easier.

Example 47: a) Find the inner and outer fences for the set of 24 exam scores in Example 39, \{15, 24, 26, 45, 48, 53, 55, 55, 58, 60, 64, 66, 67, 68, 69, 70, 75, 80, 82, 84, 89, 90, 91, 98\}; b) Use the fences to find any outliers; c) Use \( z \)-scores to find any outliers;

(NOTE: \( \sum x = 1532; \sum x^2 = 108566; \bar{x} = 63.833333; s^2 = 468.405797; s = 21.642685 \))

Example 48: (Sincich/McClave, 12th Ed., page 86, #2.136) Estimating the age of glacial drifts. Tills are glacial drifts consisting of a mixture of clay, sand, gravel, and boulders. Engineers from the University of Washington’s Department of Earth and Space Sciences studied the chemical makeup of buried tills in order to estimate the age of the glacial drifts in Wisconsin (American Journal of Science, Jan. 2005). The ratio of the elements aluminum (Al) and beryllium (Be) in sediment is related to the duration of burial. The Al-Be ratios for a sample of 26 buried till specimens are shown in the given set.

\{3.75, 4.05, 3.81, 3.23, 3.13, 3.30, 3.21, 3.32, 4.09, 3.90, 5.06, 3.85, 3.88, 4.06, 4.56, 3.60, 3.27, 4.09, 3.38, 3.37, 2.73, 2.95, 2.25, 2.73, 2.55, 3.06\}; a) Use \( z \)-scores to find outliers (if any); b) use a boxplot to find outliers (if any); c) Find \( P_{63} \); d) compute the midhinge.

(NOTE: \( \sum x = 91.18; \sum x^2 = 329.8224; \bar{x} = 3.506923; s^2 = 0.402446; s = 0.634386; Q_L = 3.1125; Q_U = 3.9375; IQR = 0.825; \) SORTED DATA: \{2.25, 2.55, 2.73, 2.73, 2.95, 3.06, 3.13, 3.21, 3.23, 3.27, 3.30, 3.32, 3.37, 3.38, 3.60, 3.75, 3.81, 3.85, 3.88, 3.90, 4.05, 4.06, 4.09, 4.09, 4.56, 5.06\})

Answers: a) the interval (2.238, 4.776) contains the data values with \( z \)-scores between -2 and 2; the interval (1.604, 5.410) contains the data values with \( z \)-scores between -3 and 3; so 5.06 is a suspicious outlier. Since there are no data values less than 1.604 or greater than 5.410, there are no definite outliers; b) lower inner fence: 1.875; upper inner fence: 5.175; lower outer fence: 0.6375; upper outer fence: 6.4125; using a boxplot, there are no suspicious outliers; no definite outliers.
Example 49: For the sample: −12, −4, 6, 18, 26, 35, 41, 51, 54, 55, 59, 59, 60, 61, 67, 69, 70, 70, 72, 83, 97, 99, 129, 146, 162; a) find \( P_{36} \); b) find the 5-point summary; c) compute the midhinge; d) compute the IQR; e) compute the inner fences and the outer fences; f) list all suspicious outliers and definite outliers (if any); g) use \( z \)-scores to find suspicious outliers and definite outliers (if any).

a) 52.32; b) minimum = −12, \( Q_L = 42.5, Q_M = 59.5, Q_U = 71.5, \) maximum = 162; c) 57; d) 29; e) the lower inner fence and the upper inner fence form the interval (−44.5, 158.5); the lower outer fence and the upper outer fence form the interval (−148.5, 158.5); f) suspicious outliers: −12, −4, 129, 142; definite outlier: 162; g) (NOTE: \( \sum x = 1741, \sum x^2 = 151171; \bar{x} = 62.178571; s^2 = 1589.559524; s = 39.869280 \); the interval (−17.56, 141.92) contains the data values with \( z \)-scores between −2 and 2; the interval (−57.43, 181.79) contains the data values with \( z \)-scores between −3 and 3; suspicious outliers: 146, 162; no definite outliers.

Example 50: For the sample: −75, −9, 26, 33, 48, 52, 55, 56, 56, 63, 67, 70, 76, 79, 83, 85, 89, 90, 95, 99, 101, 123, 159, 222, 231; a) find \( P_{81} \); b) find the 5-point summary; c) compute the midhinge; d) compute the IQR; e) compute the inner fences and the outer fences; f) list all suspicious outliers and definite outliers (if any); g) use \( z \)-scores to find suspicious outliers and definite outliers (if any).

Example 51: Consider the following box plot:

a) List the five-point summary; b) Find the midhinge; c) Find the median; d) find the inner fences; e) find the outer fences; f) list the suspicious outliers (if any); g) list the definite outliers (if any); h) Is the distribution positively skewed, negatively skewed, or symmetric?

Example 52: Consider the following box plot:

a) List the five-point summary; b) Find the midhinge; c) Find the median; d) find the inner fences; e) find the outer fences; f) list the suspicious outliers (if any); g) list the definite outliers (if any); h) Is the distribution positively skewed, negatively skewed, or symmetric?

Example 53: (Triola, 12th Ed., page 109, #24) Customer Waiting Times. Waiting times (in minutes) of customers at the Jefferson Valley Bank (where all customers enter a single waiting line) and the Bank of Providence (where customers wait in individual lines at three different teller windows) are: Jefferson Valley (single line): 6.5; 6.6; 6.7; 6.8; 7.1; 7.3; 7.4; 7.7; 7.7; 7.7

Providence (individual lines): 4.2; 5.4; 5.8; 6.2; 6.7; 7.7; 7.7; 8.5; 9.3; 10.0

a) Use the CV to compare the two data sets; b) Use boxplots to compare the two data sets.
1d) Measures of “Shape”: Skewness and Kurtosis

In addition to using measures of center (or location), measures of variability, and measures of relative standing to characterize numerical data, we can further characterize the data by including measures of **skewness** and measures of **kurtosis**. Formulas for the skewness coefficient and kurtosis coefficient will not be discussed here. These coefficients are available in most statistical software programs.

**Skewness** is a unitless measure that describes the degree of asymmetry of a distribution around its mean. A mound-shaped perfectly symmetric distribution has a skewness of zero. An asymmetrical distribution with a long tail to the right (toward the larger data values) has a positive skewness. An asymmetrical distribution with a long tail to the left (toward the smaller data values) has a negative skewness.

![Skewness Graph](image)

**Kurtosis** is a unitless measure that describes whether the shape of the distribution of the data is the same as the normal distribution (mound-shaped symmetric) or flatter than the normal distribution or more peaked than the normal distribution. The kurtosis of the normal distribution is zero. If a distribution of data has the same shape as the normal distribution, it is said to be **mesokurtic** and has zero kurtosis. If a distribution of data is flatter than the normal distribution, it is said to be **platykurtic** and has negative kurtosis. If a distribution of data is more peaked than the normal distribution, it is said to be **leptokurtic** and has positive kurtosis. The normal distribution is also said to have kurtosis 3 (called **excess kurtosis**). So, a platykurtic distribution has kurtosis less than 3 and a leptokurtic distribution has kurtosis greater than 3.

There exist alternative definitions of skewness and kurtosis. Customarily, a histogram is the one of the most popular and effective graphical technique for showing the skewness and the kurtosis of a data set.

![Kurtosis Graph](image)
**Answers to Selected Examples**

1) $\bar{x} = 2.717$; 2) $\mu = 6.25$; 3) a) 8; b) 2.65; 4) a) mean = 570.5; median = 89.5; b) mean = 89; median = 89. The median is almost unaffected by the outlier; 5) no mode (amodal); 6) 1 (unimodal); 7) 1, 8 (bimodal); 8) -21, 7, 8 (trimodal); 9) 1, 3, 5, 7, 8 (multimodal); 10) 2.85; 11) 6; 12) mean = 5.11; median = 5; mode = 6; 13) mean = 5.19; median = 6; mode = 6; 14) mean = 2.84; median = 3; mode = 2; 15) mean = 2.92; median = 2; mode = 2; 16) mean = 5; median = 5; mode = 5; 18) range = -1.7; 19) range = -16; 20) $s^2 = 0.4297$; 21) $\sigma^2 = 28.4375$; 22) $s = 0.6555$; 23) $\sigma = 5.3327$; 24) 24.1%; 25) 61.8%; 26) lot sizes: CV = 28.6%; (higher than the CV = 18.2% for appraised values); 27) B has higher variability with CV = 16.7%, so B is more volatile; A is more stable with CV = 20%; 28) 0.5167; 29) 4.6875; 30) 61.6-66.4 inches; b) 59.2-68.8 inches; c) 56.8-71.2 inches; d) 55 inches; 31) a) the percentage of scores cannot be determined; b) at least 75% 4) at least 88.9%; 33) (1) approximately 68% of the scores lie in the interval $(\bar{x} - s, \bar{x} + s)$; b) approximately 95% of the scores lie in the interval $(\bar{x} - 2s, \bar{x} + 2s)$; c) approximately 99.7% of the scores lie in the interval $(\bar{x} - 3s, \bar{x} + 3s)$; (2) a) 25th percentile; b) 88th percentile; c) 81.5; d) 55; 40) a) the score 570 is greater than or equal to 67% of all the scores in the class of 2013; b) 580 is the 70th percentile; c) the score 670 is the 89th percentile; 41) a) $Q_1 = 53.5; b) Q_M = 66.5; c) Q_2 = 81.5; 43) a) 67.5; b) 28; c) 14; d) 65.5; 44) minimum = 15, $Q_1 = 53.5, Q_M = 66.5, Q_2 = 81.5, maximum = 98; 46) a) minimum = 8, $Q_L = 14, Q_M = 17, Q_U = 22, maximum = 30; b) 18; c) 8; d) 4; 47) a) lower inner fence: 11.5; upper inner fence: 123.5; lower outer fence: -30.5; upper outer fence: 165.5; b) there are no suspicious outliers; no definite outliers; c) the interval (20.55, 107.12) contains the data values with z-scores between -2 and 2; the interval (-1.09, 128.76) contains the data values with z-scores between -3 and 3; suspicious outlier: 15; there are no definite outliers; 48) a) the interval (2.238, 4.776) contains the data values with z-scores between -2 and 2; the interval (1.604, 5.410) contains the data values with z-scores between -3 and 3; 5.06 is a suspicious outlier. Since there are no values less than 1.604 or greater than 5.410, there are no definite outliers; b) lower inner fence: 1.875; upper inner fence: 5.175; lower outer fence: 0.6375; upper outer fence: 6.4125; using a boxplot, there are no suspicious outliers; no definite outliers; 50) a) 100.74; b) minimum = -75. $Q_L = 54.25, Q_M = 77.5, Q_U = 96, maximum = 231; c) 75.125; d) 41.75; e) the lower inner fence and the upper inner fence form the interval (-8.375, 158.625); the lower outer fence and the upper outer fence form the interval (-71.00, 221.25); f) suspicious outliers: -9.159; definite outliers: -75, 222, 231; 49) $\Sigma x = 2069; \Sigma x^2 = 258393; \bar{x} = 79.576923; s^2 = 3749.933846; s = 61.236703; the interval (-42.90, 202.05) contains the data values with z-scores between -2 and 2; the interval (-104.13, 263.29) contains the data values with z-scores between -3 and 3; suspicious outliers: -75, 222, 231; no definite outliers; 52) a) minimum = -8.9, $Q_L = 22, Q_M = 30, Q_U = 32, maximum = 66; b) 27; c) 30; d) 7; 47) e) [-8, 62]; f) 48; g) -10, 64, 66; h) negatively skewed; 53) a) Jefferson Valley: $\Sigma x = 71.5; \Sigma x^2 = 513.27; \bar{x} = 7.15; s^2 = 0.227222; s = 0.476778; the interval (6.20, 8.10) contains the data values with z-scores between -2 and 2; the interval (5.72, 8.58) contains the data values with z-scores between -3 and 3; there are no suspicious outliers; no definite outliers; Providence: $\Sigma x = 71.5; \Sigma x^2 = 541.09; \bar{x} = 7.15; s^2 = 3.318333; s = 1.821629; the interval (3.51, 10.79) contains the data values with z-scores between -2 and 2; the interval (1.69, 12.61) contains the data values with z-scores between -3 and 3; no suspicious outliers; no definite outliers; a) Jefferson Valley CV = 6.7%; Providence CV = 25.5%; data values with z-scores between -2 and 2 are in the interval (3.51, 10.79); data values with z-scores between -3 and 3 are in the interval (1.69, 12.61); Jefferson Valley’s waiting times variation is much less than Providence’s; b) Jefferson Valley: (J. Valley’s 5 numbers: 6.5, 6.6, 7.2, 7.7, 7.7) (Providence’s 5 numbers: 4.2, 5.7, 7.2, 8.7, 10.0)