Chapter 3, Notes

Range Rule of Thumb for Identifying Significant Values

- **Significantly low** values are $\mu - 2\sigma$ or lower.
- **Significantly high** values are $\mu + 2\sigma$ or higher.
- **Values not significant** are between ($\mu - 2\sigma$) and ($\mu + 2\sigma$).

To roughly estimate the standard deviation from a collection of known sample data, use

$$s \approx \frac{\text{range}}{4}$$

s = **sample** standard deviation

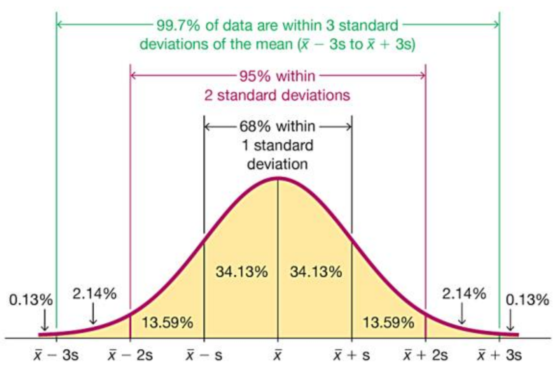
s^2 = **sample** variance

σ = **population** standard deviation

σ^2 = **population** variance

The **empirical rule** states that **for data sets having a distribution that is approximately bell-shaped,** the following properties apply.

- About 68% of all values fall within 1 standard deviation of the mean.
- About 95% of all values fall within 2 standard deviations of the mean.
- About 99.7% of all values fall within 3 standard deviations of the mean.



The **coefficient of variation** (or **CV**) for a set of nonnegative sample or population data, expressed as a percent, describes the standard deviation relative to the mean, and is given by the following:

Sample

$$CV = \frac{s}{\bar{x}} \cdot 100$$

Population

$$CV = \frac{\sigma}{\mu} \cdot 100$$

- *z* Score

***z* score** (or **standard score** or **standardized value**) is the number of standard deviations that a given value *x* is above or below the mean. The *z* score is calculated by using one of the following:

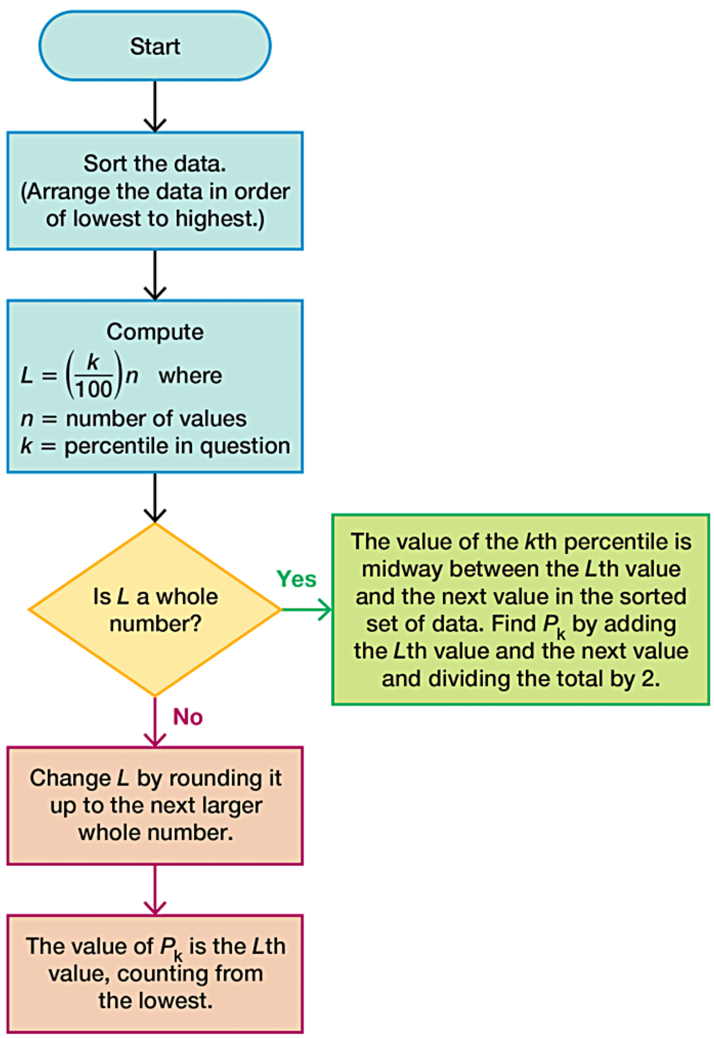
$$z = \frac{x - \bar{x}}{s} \qquad z = \frac{x - \mu}{\sigma}$$

1. A z score is the number of standard deviations that a given value x is above or below the mean.
2. z scores are expressed as numbers with no units of measurement.
3. A data value is **significantly low** if its z score is less than or equal to −2 or the value is **significantly high** if its z score is greater than or equal to +2.
4. If an individual data value is less than the mean, its corresponding z score is a negative number.

The process of finding the percentile that corresponds to a particular data value x is given by the following (round the result to the nearest whole number):

$$\text{Percentile of value } x = \frac{\text{number of values less than } x}{\text{total number of values}} \cdot 100$$

Converting a Percentile to a Data Value

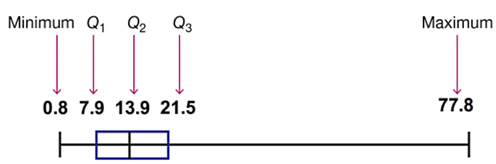


Quartiles are measures of location, denoted Q_1, Q_2, and Q_3, which divide a set of data into four groups with about 25% of the values in each group.

- 5-Number Summary
  - For a set of data, the **5-number summary** consists of these five values:
  - Minimum
  - First quartile, Q_1
  - Second quartile, Q_2 (same as the median)
  - Third quartile, Q_3
  - Maximum
- Boxplot (or Box-and-Whisker Diagram)

A **boxplot** (or **box-and-whisker diagram**) is a graph of a data set that consists of a line extending from the minimum value to the maximum value, and a box with lines drawn at the first quartile Q_1, the median, and the third quartile Q_3.

Example:



A boxplot can often be used to identify skewness. A distribution of data is **skewed** if it is not symmetric and extends more to one side than to the other.

Identifying Outliers for Modified Boxplots

1. Find the quartiles Q_1, Q_2, and Q_3.
2. Find the interquartile range (IQR), where IQR = $Q_3 - Q_1$.
3. Evaluate 1.5 × IQR.
4. In a modified boxplot, a data value is an **outlier** if it is above Q_3, by an amount greater than 1.5 × IQR or below Q_1, by an amount greater than 1.5 × IQR.