Section 3-4
Measures of Relative Standing

This section introduces measures that can be used to compare values from different data sets, or to compare values within the same data set. The most important of these is the concept of the \( z \text{ score} \).

\[ z = \frac{x - \mu}{\sigma} \]

Round \( z \) to 2 decimal places.

**Definition**

\( z \text{ Score} \) (or standardized value)
the number of standard deviations that a given value \( x \) is above or below the mean.

**Measures of Position**

**Sample**

\[ z = \frac{x - \bar{x}}{s} \]

**Population**

\[ z = \frac{x - \mu}{\sigma} \]

Whenever a value is less than the mean, its corresponding \( z \) score is negative.

Ordinary values: \( z \) score between -2 and 2
Unusual Values: \( z \) score < -2 or \( z \) score > 2

**Definition**

- \( Q_1 \) (First Quartile) separates the bottom 25% of sorted values from the top 75%.
- \( Q_2 \) (Second Quartile) same as the median; separates the bottom 50% of sorted values from the top 50%.
- \( Q_3 \) (Third Quartile) separates the bottom 75% of sorted values from the top 25%.

**Quartiles**

\( Q_{11}, Q_2, Q_3 \)

divide ranked scores into four equal parts

\[
\begin{align*}
25\% & \quad 25\% & \quad 25\% & \quad 25\% \\
(Q_1) & \quad (Q_2) & \quad (Q_3) & \quad (Q_3) \\
\text{(minimum)} & \quad & \quad & \quad \text{(maximum)}
\end{align*}
\]
Percentiles

Just as there are three quartiles separating data into four parts, there are 99 percentiles denoted $P_1, P_2, \ldots, P_{99}$, which partition the data into 100 groups.

Finding the Percentile of a Given Score

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

Converting from the $k$th Percentile to the Corresponding Data Value

Notation

$$L = \frac{k}{100} \times n$$

$n$ total number of values in the data set

$k$ percentile being used

$L$ locator that gives the position of a value

$P_k$ $k$th percentile

Some Other Statistics

- Interquartile Range (or IQR): $Q_3 - Q_1$
- Semi-interquartile Range: $\frac{Q_3 - Q_1}{2}$
- Midquartile: $\frac{Q_3 + Q_1}{2}$
- 10 - 90 Percentile Range: $P_{90} - P_{10}$

Section 3-5

Exploratory Data Analysis (EDA)

This section discusses outliers, then introduces a new statistical graph called a boxplot, which is helpful for visualizing the distribution of data.
Definition

- Exploratory Data Analysis (EDA)
  the process of using statistical tools (such as graphs, measures of center, and measures of variation) to investigate data sets in order to understand their important characteristics

- An outlier is a value that is located very far away from almost all of the other values.

Important Principles

- An outlier can have a dramatic effect on the mean.
- An outlier can have a dramatic effect on the standard deviation.
- An outlier can have a dramatic effect on the scale of the histogram so that the true nature of the distribution is totally obscured.

Definitions

- For a set of data, the 5-number summary consists of the minimum value; the first quartile, $Q_1$; the median (or second quartile, $Q_2$); the third quartile, $Q_3$; and the maximum value.
- 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, $Q_3$.

Boxplots

- Minimum $Q_1$ Median $Q_3$ Maximum
  - 21 28 35.5 39.5 80

(a) Normal (bell-shaped) distribution
1000 heights (in.) of women
Boxplots - cont

(b) Uniform distribution
1000 rolls of a die

(c) Skewed distribution
Incomes (thousands of dollars) of 1000 statistics professors

Modified Boxplots

Some statistical packages provide modified boxplots which represent outliers as special points.

A data value is an outlier if it is ...
- above Q₃ by an amount greater than 1.5 X IQR
- below Q₁ by an amount greater than 1.5 X IQR

Modified Boxplot Construction

A modified boxplot is constructed with these specifications:
- A special symbol (such as an asterisk) is used to identify outliers.
- The solid horizontal line extends only as far as the minimum data value that is not an outlier and the maximum data value that is not an outlier.

Modified Boxplots - Example

A boxplot for pulse rates by sex, showing outliers as special symbols.