MODULE 3: SUMMARY STATISTICS TO DESCRIBE DATA DISTRIBUTIONS

3.1 Measures of location

Chapter Objectives

Students will be able to

• Calculate measures of location and interpret the results;
• Calculate measures of variation and interpret the results.

Using Numbers to Describe Information

- We use two types of measures when describing data
  1. Measures of Location (center)
     1. Mean
     2. Median
     3. Mode
     4. Percentiles
     5. Quartiles
  2. Measures of Variability (spread)
     1. Range
     2. Interquartile Range
     3. Variance
     4. Standard Deviation
     5. Coefficient of Variation
Measures of Location (Center)

Mean
- The mean of a data set is the average of all the data values
- If the data are from a sample, the mean is denoted by 
  \[ \bar{x} = \frac{\sum x_i}{n} \]
  This represents each observation in the data set.
  \( \bar{x} \) represents the mean of a sample and NOT the mean of the whole population that the sample represents.
  The lowercase \( n \) represents the total number of observations in the data set. This just means the total size of the sample.

- If the data are from a population, the mean is denoted by 
  \[ \mu = \frac{\sum x_i}{N} \]
  \( \mu \) represents the mean of a population.
  The uppercase \( N \) represents the total number of observations in the population. This just means the total size of the population.

Measures of Location (Center)

Median
- It is the value in the middle when the data are arranged in ascending order.
- For an odd number of observations, the median is the exact middle value.
- Median is calculated using the middle value of the ordered data set.

In this example, the median is 9.
Example: Hospital Stay
Describe the average duration of hospital stays among 25 patients discharged from a local hospital.

### The Mean
\[
\bar{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} = \frac{5 + 10 + \ldots + 4}{25} = \frac{215}{25} = 8.6
\]
The mean 8.6 days indicates what the typical duration is. Thus, the typical length of hospital stay is about 8.6 days.

### The Median
To calculate the median, first arrange the data in increasing order:
3 3 3 4 4 5 5 5 5 6 7 7 8 8 9 9 9 10 11 11 11 11 14 17 30

\[ \text{Median} = 50\text{th percentile} \]
\[ i = \left( \frac{p}{100} \right) n = \left( \frac{50}{100} \right) 25 = 12.5 \]

\[ \text{Median} = 8 \text{ days} \]
The median 8 days indicates where the center of the data is located in the sense that 50% of hospital durations are shorter than 8 days and 50% are longer than 8 days.
Measures of Location

1. **Mean**
2. **Median**
3. **Mode**
4. **Percentiles**
5. **Quartiles**

**Mode** = The value that occurs with greatest frequency
- The greatest frequency can occur at 2 or more values
  - If the data have exactly two modes, the data are bimodal
  - If the data have >2 modes, the data are multimodal
  - If the data have one mode, the data are unimodal

3 3 4 4 5 5 5 5 6 7 7 8 8 9 9 9 10 11 11 11 11 14 17 30

There are two data points (5 days and 11 days) which have the most frequencies, each 4 times. Thus, the modes are 5 days and 11 days, giving a bimodal distribution.

Using Excel

- Excel can be used to calculate these descriptive statistics.

Excel formula
- `=AVERAGE(B2:B26)` *Calculates the mean
- `=MEDIAN(B2:B26)` *Calculates the median
- `=MODE(B2:B26)` *Calculates the mode

Example: Admission test scores for colleges and universities are frequently reported in terms of percentiles.
Percentiles

- The $p^{th}$ percentile of a data set is a value such that at least $p$ percent of the items take on this value or less and at least $(100 - p)$ percent of the items take on this value or more.

Steps in determining percentile

1. Arrange the data in ascending order.
2. Compute index $i$, the position of the $p^{th}$ percentile.
   
   $$i = \left(\frac{p}{100}\right)n$$

3. If $i$ is not an integer, round up. The $p^{th}$ percentile is the value in the $i^{th}$ position.
4. If $i$ is an integer, the $p^{th}$ percentile is the average of the values in positions $i$ and $i+1$.

Computing Percentiles

1. Arrange the data in ascending order. This is data from the hospital stay example.

   | 3 3 3 3 4 4 5 5 5 5 6 7 7 8 8 9 9 9 10 11 11 11 11 14 17 30 |

2. Compute index $i$, the position of the $p^{th}$ percentile. (This can be whatever percentile you want. Here we want to find the 90th percentile.
   
   $$i = \left(\frac{p}{100}\right)n = \left(\frac{90}{100}\right)25 = 22.5$$

3. If $i$ is not an integer, round up. The $p^{th}$ percentile is the value in the $i^{th}$ position.
   
   The 23rd data value represents the 90th percentile
   
   90th Percentile = 14 days

4. If $i$ is an integer, the $p^{th}$ percentile is the average of the values in positions $i$ and $i+1$.

Quartiles

- Quartiles are specific percentiles
  - First Quartile ($Q_1$) = 25th Percentile
  - Second Quartile ($Q_2$) = 50th Percentile = Median
  - Third Quartile ($Q_3$) = 75th Percentile

   | 3 3 3 4 4 5 5 5 5 6 7 7 8 8 9 9 9 10 11 11 11 11 14 17 30 |

   $Q_1$ $Q_2$ $Q_3$

   Third quartile ($Q_3$) = 75th percentile
   
   $$i = \left(\frac{p}{100}\right)n = \left(\frac{75}{100}\right)25 = 18.75 = 19$$

   $Q_3$ = 11 days
Percentiles in Excel

Excel formula

- \text{PERCENTILE.EXE}(B2:B26, A34)
  - Calculates the 25\textsuperscript{th} percentile (i.e., Q1)
- \text{PERCENTILE.EXE}(B2:B26, A35)
  - Calculates the 50\textsuperscript{th} percentile (i.e., Q2)
- \text{PERCENTILE.EXE}(B2:B26, A36)
  - Calculates the 75\textsuperscript{th} percentile (i.e., Q3)

The A34 value, located in column A row 34, points to the percentile you want to be calculated (e.g., 25\textsuperscript{th}, 50\textsuperscript{th}, 75\textsuperscript{th})