Chapter 2: Describing Distributions with Numbers: Key Ideas

Measures of Center

- **Mean $\bar{x}$**
  - The *mean* of the n values $x_1, x_2, \ldots, x_n$ is given by the formula

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

- **Median M**
  - The **median** is the midpoint of a distribution, the number such that half the values are smaller and the other half are larger.
  - To find the median of a set of data values:
    1. Arrange the data values in order of size, from smallest to largest.
    2. If

• n **is odd**: The median M is <u>the center value</u> in the ordered list. o Find the **location of the median** by counting $\frac{(n+1)}{2}$ values from the bottom of the list.	• n **is even**: The median M is <u>the mean of the two center values</u> in the ordered list o The **location of the median** is again $\frac{(n+1)}{2}$ from the bottom of the list.

Important Notes:

- The formula $\frac{(n+1)}{2}$ gives the **location** but not the **value** of the median. Don't get confused about this!
- The median is often preferred over the mean as a measure of center because it is …
    - **Resistant:** relatively unaffected by changes in the numerical value of a small proportion of the total data set

Measures of Spread

The Quartiles Q_1 and Q_3 demarcate the middle half of a data set. They are calculated as follows:

1. Arrange the observations in increasing order (from left to right) and find the **location of the median *M*** in the overall ordered list of observations.
2. The **first (or lower) quartile Q_1 is the median** of the observations whose position in the ordered list is to the left of the location of the overall median.
3. The **third (or upper) quartile Q_3 is the median** of the observations whose position in the ordered list is to the right of the location of the overall median.

SPSS Note

- **SPSS reports the "Tukey's Hinges" quartiles which differ slightly from Q_1 and Q_3 as defined here. See the SPSS help sheet on summary statistics for more details.**
- **The Interquartile Range (IQR)** is a **resistant** measure of the spread in the middle half of the data computed from the quartiles:

$$IQR = Q_3 - Q_1.$$

- Under certain conditions, the **standard deviation, *s*,** may be viewed intuitively as the "typical distance" between a data value and its mean (though this is not technically correct).
  - ***s*** is computed as given in the display on p. 48 of Moore.
  - Like the mean $\bar{x}$, the standard deviation ***s*** is <u>not</u> resistant.
  - s=0 only when all observations are the same.
  - s has the same units of measurement as the original observations.

If the distribution of a data set is roughly bell-shaped, the **68-95-99.7 rule for data** gives intuitive meaning to s:

- About 68% of the values will fall in the interval $(\bar{x} - s, \bar{x} + s)$
- About 95% of the values will fall in the interval $(\bar{x} - 2s, \bar{x} + 2s)$
- About 99.7% of the values will fall in the interval $(\bar{x} - 3s, \bar{x} + 3s)$

Five-number summary and boxplots

- Regardless of the shape of the distribution, a concise summary of a data set is given by the **five-number** of a data set consists of the values

Minimum, Q_1, M, Q_3, Maximum

written in order from smallest to largest.

  - **Note**: The five-number summary is of limited value for very small data sets (say, for $n \leq 10$). For such small data sets, it is better to simply provide a stemplot or a listing of the entire data set.

- A **boxplot** is a graph of the five-number summary comprised of:
    - A central box representing the middle half of the data (from Q_1 to Q_3).
    - A line in the box representing the median.
    - "Whiskers" extending from the edges of the box (Q_1 and Q_3) to the minimum and maximum values.

Identifying suspected outliers

- Compute the $IQR = Q_3 - Q_1$.
- Call an observation a *suspected outlier* if it falls more than $1.5 \times$ IQR above the third quartile or below the first quartile.

Modified Boxplot

A *modified* boxplot extends the ordinary boxplot by plotting suspected outliers as separate points.

The plot can be produced by the Explore procedure in SPSS (see help sheet) or directly from SPSS using the Graphs…Boxplot procedure

Carried out by hand, the procedure is:

- Draw the box depicting the median and quartiles.
- Locate (but do not draw) the fences as follows:
  - *Lower fence*= $Q_1 - 1.5 \times IQR$
  - *Upper fence*= $Q_3 + 1.5 \times IQR$
- Extend the whiskers from the edges of the box to reach the most extreme data value *inside* the fences.
- Plot a $*$ to represent any *suspected outliers* (values *outside* the fences). [1]

[1] SPSS uses an "o" to represent points between 1.5 and 3 IQRs from the box and an "*" to present points more than 3 IQRs from the box.

Summaries of Center and Spread

- For reasonably large data sets, **the 5-number summary** is appropriate whatever the shape of the distribution of the data.
- For symmetric distributions without outliers, **the mean** $\bar{x}$ **and standard deviation** s are also useful measures of center and spread. Otherwise, use only resistant measures.

The Four-Step Process for Solving Statistical Problems (S-F-S-C) (Moore p.53)

1. **State:** What is the practical question, *in the context* of the real-world setting?
2. **Formulate:** What specific *statistical operations* does this problem call for?
3. **Solve:** Make the graphs and carry out the calculations needed for this problem.
4. **Conclude:** Give your practical conclusion *in the setting* (*in context*) of the real-world problem.