2.3 Describing Distributions with Numbers

The Mean (for measuring the center):

The **(arithmetic) mean** of n observations $x_1, x_2, x_3, \ldots, x_n$ is denoted by $\overline{x}$ (read as "x bar") and is defined as

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example:

- "Most Widely Held Stocks":

- "The Cost of Victories":

Note:

There are other "means" in addition to the arithmetic mean defined above.

The Median (for measuring the center):

The **median** of n observations $x_{(1)}, x_{(2)}, x_{(3)}, \ldots, x_{(n)}$ is denoted by M or $\tilde{x}$ (read as "x tilde") and is defined as

$$M = \tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{if } n \text{ is odd} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}, & \text{if } n \text{ is even} \end{cases}$$

Note:

When we use the notation $x_{(i)}$, we assume that the data is ordered from smallest to largest and $x_{(i)}$ is the i^{th} smallest observation, i.e., $x_{(1)}$ refers to the smallest observation, $x_{(2)}$ to the second smallest, and so on, and finally $x_{(n)}$ to the largest observation.

When we have an odd number of observations, e.g., $n = 11$, then $\frac{n+1}{2} = \frac{11+1}{2} = 6$. In the definition of the median above, the expression $x_{\left(\frac{n+1}{2}\right)}$ turns out to be $x_{(6)}$ which is the 6^{th} smallest as well as the 6^{th} largest observation, i.e., the observation in the middle.

Similarly, when we have an even number of observations, the formula above states nothing else than that the median is the arithmetic mean of the 2 observations in the middle.

Example:

- We have observations $x_1 = 7, x_2 = 5, x_3 = 4, x_4 = 10, x_5 = 4$.
  then $x_{(1)} =$ $, x_{(2)} =$ $, x_{(3)} =$ $, x_{(4)} =$ $, x_{(5)} =$.
  $n = 5$ is odd.
  The median is
  $$M = \tilde{x} = x_{(\frac{n+1}{2})} = ...$$

- "Most Widely Held Stocks":
  $$\begin{aligned} x_{(1)} &= 31\tfrac{9}{16} \\ x_{(2)} &= 37\tfrac{7}{8} \\ &\vdots \\ x_{(8)} &= 73\tfrac{5}{8} \\ &\vdots \\ x_{(15)} &= 125\tfrac{13}{16} \end{aligned}$$
  $n = 15$ is odd.
  The median is
  $$M = \tilde{x} = x_{(\frac{n+1}{2})} = ...$$

- "The Cost of Victories":
  $$\begin{aligned} x_{(1)} &= 8 \\ x_{(2)} &= 11 \\ &\vdots \\ x_{(15)} &= 19 \\ x_{(16)} &= 19 \\ &\vdots \\ x_{(30)} &= 29 \end{aligned}$$
  $n = 30$ is even.
  The median is
  $$M = \tilde{x} = \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} = ...$$

Note:

The **median** is not effected by a few "unusual/extreme" values in a data set. The **mean** is effected by such values. The median $\tilde{x}$ is called a "robust" statistic.

The Range (for measuring spread):
The **range** of n observations $x_{(1)}, \ldots, x_{(n)}$ is denoted by R and is defined as

$$R = x_{(n)} - x_{(1)},$$

i.e., the largest observation minus the smallest observation.

Example:

- "Most Widely Held Stocks":

$$R =$$

- "The Cost of Victories":

$$R =$$

The Quartiles (for measuring spread):
The **quartiles** of n observations $x_{(1)}, \ldots, x_{(n)}$ are denoted by Q_1 and Q_3 and can be (roughly) determined as the median of the observations that are listed left of $\tilde{x}$, i.e., observations that are smaller or equal to the median, (for Q_1) and the median of the observations that are listed right of $\tilde{x}$, i.e., observations that are larger or equal to the median (for Q_3). We often call Q_1 the lower (or first) quartile and Q_3 the upper (or third) quartile.

Example:

- Recall the example where we have observations $x_1 = 7, x_2 = 5, x_3 = 4, x_4 = 10, x_5 = 4$. then $x_{(1)} = 4, x_{(2)} = 4, x_{(3)} = 5, x_{(4)} = 7, x_{(5)} = 10$.

- "Most Widely Held Stocks":

$$\begin{aligned}
x_{(1)} &= 31\tfrac{9}{16} \\
x_{(2)} &= 37\tfrac{7}{8} \\
x_{(3)} &= 43\tfrac{1}{4} \\
x_{(4)} &= 57\tfrac{5}{16} \\
x_{(5)} &= 68\tfrac{3}{16} \\
x_{(6)} &= 71\tfrac{5}{16} \\
x_{(7)} &= 72\tfrac{3}{16} \\
x_{(8)} &= 73\tfrac{5}{8} \\
x_{(9)} &= 77\tfrac{3}{4} \\
x_{(10)} &= 84\tfrac{1}{8} \\
x_{(11)} &= 84\tfrac{9}{16} \\
x_{(12)} &= 88\tfrac{15}{16} \\
x_{(13)} &= 91\tfrac{1}{8} \\
x_{(14)} &= 118\tfrac{1}{16} \\
x_{(15)} &= 125\tfrac{13}{16}
\end{aligned}$$

Note:

There are two slightly different interpretations of what is meant by "left of" and "right of" the median — different computer software may yield results that are different from our hand–calculated results. However, the more observations we have, the smaller these differences typically are.

The Interquartile Range (for measuring spread):

The **interquartile range** of n observations $x_{(1)}, \ldots, x_{(n)}$ is denoted by IQR and is defined as

$$IQR = Q_3 - Q_1,$$

i.e., the third quartile minus the first quartile.

Example:

- "Most Widely Held Stocks":

$$IQR =$$

Note:
An observation that falls more than $1.5 \times IQR$ above the third quartile or an observations that falls more than $1.5 \times IQR$ below the first quartile is a suspected **outlier**.

Putting all together:
Five–Number Summary and Boxplots

The **five–number summary** consists of

$$\text{Minimum} \quad Q_1 \quad \tilde{x} \quad Q_3 \quad \text{Maximum}$$

and provides a reasonably complete overview of the center and spread of a data set. A **boxplot** is a graphical representation of these numbers and is constructed as follows:

- Draw a central box that spans the quartiles.
- Draw a line in the box that marks the median.
- Draw each observation that falls more than $1.5 \times IQR$ outside the central box separately using a $*$ (denoting a possible outlier).
- Draw a line from the box to the smallest and largest observations that are not marked as possible outliers (often, but not always, these are minimum and maximum).

Example: "Most Widely Held Stocks"

Five–Number Summary:

$IQR =$,$1.5 \times IQR =$

Possible outlier would fall below $Q_1 - 1.5 \times IQR =$ or above $Q_3 + 1.5 \times IQR =$.

Boxplot:

The Variance and Standard Deviation (for measuring spread):

The **(sample) variance** of a sample of n observations $x_1, \ldots, x_n$ is denoted by s^2 and is defined as

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \ldots + (x_n - \overline{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

The **standard deviation** is denoted by s and is defined as

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

Note:

(i) If $X_1, \ldots, X_N$ represents an entire **population** of N individuals, we define the **(population) variance** σ^2 as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$$

where μ is the known population mean.

(ii) Check your pocket calculator! If it says σ^2 (or σ), it devides by N. If it says s^2 (or s), it devides ny $n-1$.

(iii) The variance is a kind of "mean squared deviation" from the center of a data set.

(iv) s^2 (and σ^2) is never negative. It can be 0 only if all data values are equal.

(v) Shortcut formula for hand calculations:

$$SS(x) = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}x_i^2 - \frac{\left(\sum_{i=1}^{n}x_i\right)^2}{n}$$

where $SS(x)$ stands for "Sum of Squares" (of the x–values).
Then, the sample variance is calculated as

$$s^2 = \frac{1}{n-1}SS(x).$$

Example: "The Cost of Victories"

$n = 30$ (or $N = 30$)

x_i	x_i^2

$$\sum_{i=1}^{n} x_i =$$

$$\sum_{i=1}^{n} x_i^2 =$$

$$SS(x) = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$=$$

If we assume that $N = 30$, i.e., these 30 teams represent the whole population, i.e., all possible teams, we have to calculate the population variance and standard deviation:

$$\sigma^2 = \frac{1}{N} SS(x)$$

$$=$$

$$\sigma = \sqrt{\sigma^2}$$

$$=$$

If we assume that $n = 30$, i.e., these 30 teams represent a sample of a larger population, i.e., a subset of all possible teams (and there exist other teams that have not been listed in this newspaper data set), we have to calculate the sample variance and standard deviation:

$$s^2 = \frac{1}{n-1} SS(x)$$

$$=$$

$$s = \sqrt{s^2}$$

$$=$$