5.2 Measuring Center

a. middle of a set of data

- mean
- median
- mode

1. Mean

the sum of the observed values divided by the number of observations.

**Notation:**

- $X_1, X_2, \ldots, X_n$ denote a sample of $n$ observations.

$$\bar{X} = \frac{\sum X_i}{n} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- Mean of the sample

$$\mu = \text{Mean of the population}$$
Ex 5.1  Mean number of children per household in the sample, \( n = 10 \)

Data set: 2, 3, 0, 2, 1, 0, 3, 3, 0, 1

The mean of these 10 observations is

\[
\bar{X} = \frac{\sum X_i}{n} = \frac{2 + 3 + 0 + 2 + 1 + 0 + 3 + 3 + 0 + 1 + 4}{10} = 1.6
\]

40 instead of 4.

\[
\bar{X} = \frac{2 + 3 + 0 + 2 + 1 + 0 + 3 + 0 + 1 + 40}{10} = 5.2
\]

Mean is sensitive to extreme observations.

2. Median (Q2)

Is the sample value such that at least \( \frac{1}{2} \) of the sample values are less than or equal to that value and at least \( \frac{1}{2} \) of the sample values are greater than or equal to it. Median is more resistant to extreme values.

1. Order the data set, (increasing order)

\[ x_{(1)} = \text{minimum} \rightarrow \ldots \rightarrow x_{(n)} = \text{maximum} \]

2. \( n \) is odd, \( Q_2 = X_{(\frac{n+1}{2})} \)

3. \( n \) is even, \( Q_2 = \frac{X_{(\frac{n}{2})} + X_{(\frac{n+1}{2})}}{2} \)
Ex. 5.1 \( n=10 \)

2, 3, 0, 2, 1, 0, 3, 0, 1, 4

1. Order the data set

\[ 0, 0, 0, 1, 1, 2, 2, 3, 3, 4 \]

\[ x_{(0)} \]

2. \( \bar{x} = \frac{1+2}{2} = 1.5 \) 

(= \( \frac{x_{(1)}+x_{(10)}}{2} \))

3. Mode is the most frequently occurring value(s).

Ex. 5.1 \( 0, 0, 0, 1, 1, 2, 2, 3, 3, 4 \)

the mode is 0

Ex. \( \{0, 0, 0, 1, 1, 2, 2, 3, 3, 4\} \)

0, 2 are modes. (Bimodal)

Ex. \( \{0, 1, 2, 4, 5, 8\} \)

no mode.

mode usually computed for qualitative data.

5.2.4 Which Measure of Center to Use?

Symmetric
Bell-shaped
mean = median = mode

[Diagram of a bell curve]
Summary:

- **Mean**: roughly symmetric, unimodal, no outliers.
- **Median**: skewed distributions, with outliers.
- **Mode**: qualitative values.
5.3 Measuring Variation on Spread

Ex.: Data Set 1: 5, 6, 7, 7, 8, 9
Data Set 2: 3, 5, 7, 7, 9, 11

- Mean = median = mode = 7
- Range
- Interquartile range
- Standard deviation

1. Range (Overall range)

\[ \text{Range} = \text{Maximum} - \text{Minimum} = X_{(n)} - X_{(1)} \]

- Rough measure of spread

2. Interquartile range (IQR)

\[ \text{IQR} = Q_3 - Q_1 \]

- \( Q_1 \): First quartile = the median of observations that fall below the median.
  (about 25% of the values are on or below \( Q_1 \))

- \( Q_3 \): Third quartile = the median of observations that fall above the median.
  (about 75% of the values are on or below \( Q_3 \)).
Notes:

1. If \( n \) is odd, \( Q_2 \) is the middle observation.
   This observation is not included in either of the halves when computing \( Q_1 \) and \( Q_3 \).

\[ Q_1 = \begin{cases} 
\frac{X(m+1)}{2}, & \text{if } m \text{ is odd} \\
\frac{X(\frac{m}{2}) + X(\frac{m+1}{2})}{2}, & \text{if } m \text{ is even} 
\end{cases} \\
m = \frac{n}{2} \text{ rounded down.}
\]

2. \( Q_3 \) is the median of the upper half.

Example. Data set 1: \( \{1, 2, 3, 4, 5, 6, 7\} \)

- \( Q_2 = 4 \)
- \( Q_1 = 3 \)
- \( Q_3 = 6 \)

IQR = \( Q_3 - Q_1 = 4 \)

Data set 2: \( \{1, 2, 3, 4, 5, 6, 7, 8\} \)

- \( Q_2 = \frac{4+5}{2} = 4.5 \)
- \( Q_1 = 2.5 \)
- \( Q_3 = 6.5 \)

IQR = \( Q_3 - Q_1 = 4 \)

percentiles.
of the observations fall at or below that value and (100 - p)% of the observations fall at or above that value.

Q_1 25th percentile

3. Five number Summary
   - **median** — center
   - **IQR** — spread / variation

Five number summary:
- minimum, Q_1, median(Q_2), Q_3, maximum

**Boxplot:**

```
       +-------------------+
       |                  |
min   +-------------------+ max
       | Q_1 Q_2 Q_3       |
       +-------------------+
```

To **Build a Basic** boxplot:
1. Order the data set (increasing order)
2. Find the five number summary
3. Locate Q_1, Q_2, Q_3 on the scale. These values determine the "Box" part.
4. Draw lines from the midpoints of the ends of the box out to the min and max.
Ex: Data Set 2: \{5, 2, 3, 4, 5, 6, 7, 8\}

- $Q_2 = 4.5$, $Q_1 = 2.5$, $Q_3 = 6.5$
- $\text{min} = 1$, $\text{max} = 8$
- $\text{IQR} = 4$
- $Q_3 - Q_1$

*Boxplot (basic)*

Further analysis:

Using the IQR Rule to identify outliers:

1. Order the data set (increasing order)
2. Five Number Summary:
   - $Q_1$
   - $Q_2$
   - $Q_3$
   - $Q_1 - Q_3$
   - $Q_3 - Q_1$

3. $\text{IQR} = Q_3 - Q_1$
4. **STEP** = $1.5 \times \text{IQR}$
5. Locate $Q_1$, $Q_2$, $Q_3$, "box" part.
6. **Inner fences:**
   - Lower inner fence $= Q_1 - \text{STEP}$
   - Upper inner fence $= Q_3 + \text{STEP}$
1. Draw outliers from the midpoints of the ends of box out to those two values. (rather out to the min and max)

3. Observations that fall outside the inner fences are considered potential outliers. Plotted individually along the scale using a solid dot.

Example: 25, 53, 70, 75, 90, 90, 91, 95, 95, 95, 96, 100, 105, 110, 115, 120

Five number summary:
- min = 25
- Q1 = 82.5
- Q2 = 95
- Q3 = 102.5
- max = 120

IQR = Q3 - Q1 = 20

STEP = 1.5 x IQR = 30

Lower inner fence = Q1 - STEP = 82.5 - 30 = 52.5
Upper inner fence = Q3 + STEP = 102.5 + 30 = 132.5

Basic boxplot:
Modified boxplot:

- Inner fences
- Potential outliers

Symmetric distribution → symmetric boxplot

Ex. 5.7, 5.8.

4. Standard deviation

- Mean = center
- Standard deviation = variation
  
  *is the spread in observations from the mean.*

   1. Deviation of a value = \( \bar{x} - \overline{x} \)
   
   2. Squared deviation = \( (x_i - \overline{x})^2 \)
   
   3. Sample variance = \( \frac{\sum (x_i - \overline{x})^2}{n-1} = s^2 \)

   4. Sample standard deviation
   
   \[ s = \sqrt{\frac{\sum (x_i - \overline{x})^2}{n-1}} = \sqrt{s^2} \]

   5. Population standard deviation

\[ \sigma = \frac{s}{\sqrt{n}} \]
Example 5.9
\[ \{0, 5, 7, 4\} \]
\( \bar{x} = 4 \)
\( x_1, x_2, x_3 \)
\( n = 3 \)

Deviations: -4, 1, 3
Squares of deviations: 16, 1, 9
Sample variance: \( \frac{16 + 1 + 9}{3-1} = \frac{26}{2} = 13 \)
Sample standard deviation: \( \sqrt{13} \approx 3.6 \)

\[ S = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{\sum x_i^2 - (\sum x_i)^2}{n(n-1)} \]

Population variance

Summary:
- Median: skewed distribution, with outliers
- IQR: symmetric (roughly), unimodal
- Mean: no outliers