Chapter 3: Descriptive Statistics and Analytics: Numerical Methods

STAT 2601 – Business Statistics

Esam Mahdi

School of Mathematics and Statistics
Carleton University

January 21, 2026

Chapter 3 Outline

- Measures of Central Tendency: Mean, Median, Mode
- Measures of Variation: Range, Variance, Standard Deviation, CV
- z-scores, Empirical Rule, Outlier Detection
- Percentiles and Quartiles using Linear Interpolation
- Box-and-Whisker Plots

Measures of Central Tendency: Mean (Arithmetic Average)

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Example

Data: 5, 7, 9, 12, 15

$$\bar{x} = \frac{5 + 7 + 9 + 12 + 15}{5} = \frac{48}{5} = 9.6$$

Advantages:

- Uses all data points
- Algebraically tractable

Disadvantages:

- Sensitive to outliers

Example: (Calculate the Mean when Data has Outliers)

Data: 5, 7, 9, 12, 15, 1000

$$\bar{x} = \frac{5 + 7 + 9 + 12 + 15 + 1000}{6} = \frac{1048}{6} \approx 174.67$$

Measures of Central Tendency: Median

Median: Middle value after sorting the data (from smallest to largest), denoted by M_d.

Step 1: Sort the Data

Arrange all values in **ascending order** (smallest to largest).

Step 2: Find the Median Position

Calculate the median's **position** in the sorted list:

$$\text{Median Position} = \frac{n+1}{2}$$

where n = number of observations.

Step 3: Determine the Median Value

- **If n is odd:** Median is the value at position $\frac{n+1}{2}$.
- **If n is even:** Median is the average of the two middle values at positions $\frac{n}{2}$ and $\frac{n}{2}+1$.

Examples: Location and Calculation Steps

Example 1: Odd Number of Values ($n = 5$)

Data: $5, 12, 15, 7, 9$

1. Sorted: $5, 7, 9, 12, 15 \quad (n = 5)$
2. Position: $\dfrac{n+1}{2} = \dfrac{5+1}{2} = \dfrac{6}{2} = 3$
3. Median $= M_d =$ 3rd value = $\boxed{9}$

Example 2: Even Number of Values ($n = 4$)

Data: $5, 7, 9, 12$

1. Sorted: $5, 7, 9, 12 \quad (n = 4)$
2. Position: $\dfrac{n+1}{2} = \dfrac{4+1}{2} = \dfrac{5}{2} = 2.5$ (between positions 2 and 3)
3. Median $= M_d =$ average of 2nd and 3rd values = $\dfrac{7+9}{2} = \boxed{8}$

Example: (Odd Number of Observations: $n = 13$)

A retail store tracks daily sales revenue (in hundreds of dollars) over 13 days:

$$6, \; 9, \; 8, \; 11, \; 10, \; 7, \; 12, \; 9, \; 8, \; 10, \; 13, \; 11, \; 9$$

Step 1: Sort the data in ascending order

$$6, \; 7, \; 8, \; 8, \; 9, \; 9, \; 9, \; 10, \; 10, \; 11, \; 11, \; 12, \; 13$$

Step 2: Identify the middle observation

Since $n = 13$ is odd, the median is the $(n+1)/2 = 7$th observation.

$$\text{Median} = 9$$

Answer: The median daily sales revenue is $\boxed{9}$ (hundreds of dollars).

Example: (Even Number of Observations: $n = 12$)

A company records the number of minutes customers wait for service on a given day. The waiting times (in minutes) are:

$$8,\ 12,\ 15,\ 10,\ 18,\ 14,\ 20,\ 9,\ 16,\ 11,\ 13,\ 17$$

Step 1: Sort the data in ascending order:

$$8,\ 9,\ 10,\ 11,\ 12,\ 13,\ 14,\ 15,\ 16,\ 17,\ 18,\ 20$$

Step 2: Identify the location of the median

Location $= \dfrac{n+1}{2} = \dfrac{12+1}{2} = \dfrac{13}{2} = 6.7$, i.e., the median is the average of the 6th and 7th observations.

$$\text{Median} = \frac{13+14}{2} = 13.5$$

Answer: The median waiting time is $\boxed{13.5}$ minutes.

Advantages and Disadvantages of Median

Advantages:

- Can be used with ordinal data
- Resistant to outliers and extreme values

Example

Consider the following dataset (already sorted):

$$1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 100$$

The dataset contains an outlier at 100.

Since there are $n = 7$ observations and n is odd, the median is the middle value:

$$\text{Median} = x_{\frac{n+1}{2}} = x_{\frac{7+1}{2}} = x_4$$

Thus,

$$M_d = 4$$

Notice that the outlier 100 does not affect the median, which remains close to the center of the main data values.

Disadvantages:

- Does not use all data values
- Less suitable, compared to the mean, for algebraic manipulation

Measures of Central Tendency: Mode

- The **mode** is the value that occurs **most frequently** in a dataset, denoted by $M_\circ$.
- A dataset may have:
  - one mode (unimodal),
  - two modes (bimodal),
  - or more than two modes (multimodal).
- The mode is especially useful for **categorical or discrete data**.

Advantages:

- Easy to understand and simple to compute.
- Can be used with **categorical, ordinal, or numerical** data.
- Not affected by extreme values or outliers.

Disadvantages:

- May not be unique (can have multiple modes).
- Not based on all data values.
- Not suitable for algebraic or mathematical analysis.

Unimodal Distribution

$$2,\ 3,\ 3,\ 4,\ 5,\ 6$$

- The value 3 appears most frequently.
- **Mode = 3**
- This dataset has **one mode**, so it is **unimodal**.

Bimodal Distribution

$$1,\ 2,\ 2,\ 3,\ 4,\ 4,\ 5$$

- The values 2 and 4 both occur most frequently.
- **Modes = 2 and 4**
- This dataset is **bimodal**.

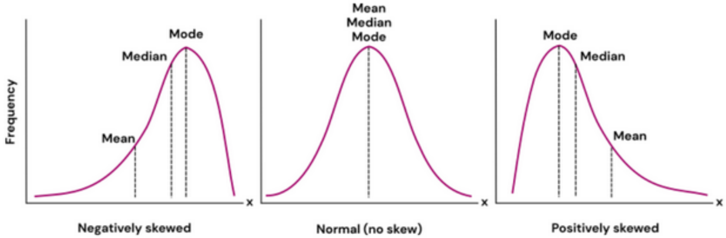
Multimodal Distribution

$$1,\ 1,\ 2,\ 2,\ 3,\ 3,\ 4$$

- The values 1, 2, and 3 all occur with the same highest frequency.
- **Modes = 1, 2, and 3**
- This dataset is **multimodal**.

Mean, Median, and Mode: Their Relationship to the Shape of a Distribution

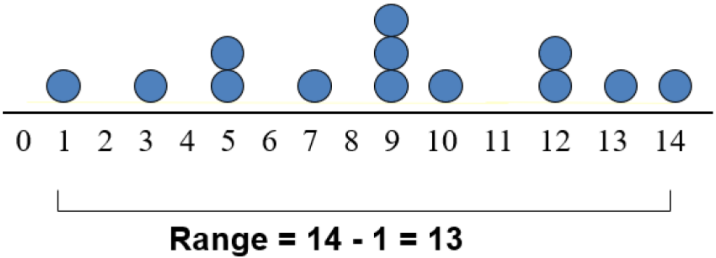
Distribution Type	Visual Shape	Mathematical Relationship
Symmetric	Bell-shaped; mirror image sides.	Mean$(\bar{x}) \approx$ Median$(M_d) \approx$ Mode$(M_\circ)$
Right-Skewed	Long tail stretches to the right.	Mean$(\bar{x}) >$ Median$(M_d) >$ Mode$(M_\circ)$
Left-Skewed	Long tail stretches to the left.	Mean$(\bar{x}) <$ Median$(M_d) <$ Mode$(M_\circ)$



Measures of Variation: Range

Range:: Difference between the largest and the smallest observations:

$$R = X_{\text{Max}} - X_{\text{Min}}$$



Range: Simple but sensitive to outliers

Measures of Variation: Variance and Standard Deviation (Population)

- **Population Variance (Parameter):**
  - Definitional Formula: $\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$
  - Working Formula (recommended): $\sigma^2 = \dfrac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}$
  - Has the square unit of the original data (cm^2, kg^2, $\$^2$).
- **Population Standard Deviation (Parameter):**
  - $\sigma = \sqrt{\sigma^2} = \sqrt{\dfrac{\sum x^2 - \frac{(\sum x)^2}{N}}{N}}$
  - Has the same unit as the original data (cm, kg, $\$$).

Measures of Variation: Variance and Standard Deviation (Sample)

- **Sample Variance (Statistic):**
  - Definitional Formula: $s^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$.
  - Working Formula (recommended): $s^2 = \dfrac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$.
  - Has the square unit of the original data (cm^2, kg^2, $\$^2$).

- **Sample Standard Deviation (Statistic):**
  - $s = \sqrt{s^2} = \sqrt{\dfrac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}}$.
  - Has the same unit as the original data (cm, kg, \$).

Note: The sum of deviations of data points from their arithmetic mean is always zero: $\sum_{i=1}^{2}(x_i - \bar{x}) = 0$

Example (Definitional Formula): Sample Variance and Standard Deviation

Calculate the variance and standard deviation for the data: 5, 12, 6, 8, , 14

Solution: Using the definitional formula:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \text{ and } s = \sqrt{s^2}, \text{ where } \bar{x} = \frac{\sum x}{n} = \frac{45}{5} = 9$$

	x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
	5	-4	16
	12	3	9
	6	-3	9
	8	-1	1
	14	5	25
Sum	**45**	**0**	**60**

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = \frac{60}{4} = 15$$

$$s = \sqrt{s^2} = \sqrt{15} = 3.87$$

Example (Recommended Formula): Sample Variance and Standard Deviation

Solution: Using the recommended formula:

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

	x_i	x_i^2
	5	25
	12	144
	6	36
	8	64
	14	196
Sum	**45**	**465**

$$s^2 = \frac{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}{n-1} = \frac{465 - \frac{45^2}{5}}{4} = 15$$

$$s = \sqrt{s^2} = \sqrt{15} = 3.87$$

Measures of Variation: Variance and Standard Deviation (Sample)

Some Notes

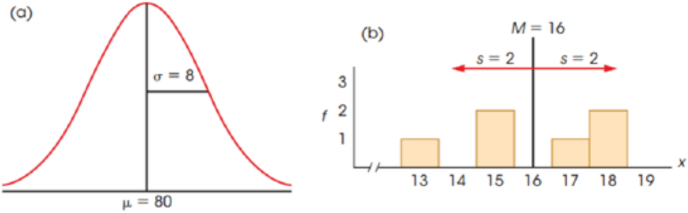
- The value of s^2 or s is ALWAYS positive.
- The larger the value of s^2 or s, the larger the variability of the data set.

Why dividing by $n - 1$?

- The sample standard deviation s is often used to estimate the population standard deviation σ.
- Dividing by $n - 1$ gives us a better estimate of σ.
- Degrees of freedom: Number of scores $(n - 1)$ in sample that are independent and free to vary.

Measures of Variation

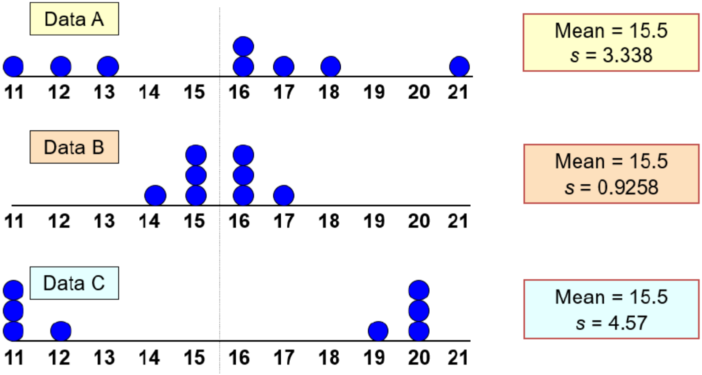
- Means and standard deviations are particularly useful in clarifying graphs of distributions.
- Means and standard deviations together provide extremely useful descriptive statistics for characterizing distributions.
- For both populations and samples, it is easy to represent mean and standard deviation.
  - Vertical line in the "center" denotes location of mean.
  - Horizontal line to right, left (or both) denotes the distance of one standard deviation.



Measures of Relative Variation: Coefficient of Variation

Relative Measures of Variability:

- Same mean but different standard deviations.
- Although the centre is the same, the way the data are distributed is different.



Measures of Relative Variation: Coefficient of Variation

Coefficient of Variation (CV): The ratio of the standard deviation to the mean expressed as a percentage. Lower the CV, the more stable or consistent the data set is.

$$\text{Population CV} : \frac{\sigma}{\mu} \times 100\%$$

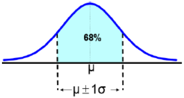
$$\text{Sample CV} : \frac{s}{\bar{x}} \times 100\%$$

Example

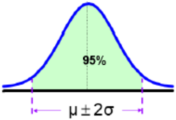
- Stock A:
  - Average price last year = \$50
  - Standard deviation = \$5
  - $CV_A = \frac{s}{\bar{x}} \times 100\% = \frac{\$5}{\$50} \times 100\% = 10\%$
- Stock B:
  - Average price last year = \$100
  - Standard deviation = \$5
  - $CV_B = \frac{s}{\bar{x}} \times 100\% = \frac{\$5}{\$100} \times 100\% = 5\%$
- Both stocks have the same standard deviation, but stock B is less variable (or more stable) relative to its price.

Empirical Rule (68%, 95%, 99.7%)

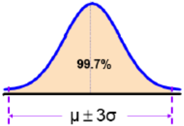
- If the data distribution is bell-shaped or symmetric, the tolerance interval $\mu \pm 1\sigma$ contains about 68% of the values in the population or the sample.



- If the data distribution is bell-shaped or symmetric, the tolerance interval $\mu \pm 2\sigma$ contains about 95% of the values in the population or the sample.



- If the data distribution is bell-shaped or symmetric, the tolerance interval $\mu \pm 3\sigma$ contains about 99.7% of the values in the population or the sample.



Example (Empirical Rule)

The length of time for a worker to complete a specified operation averages 12.8 minutes with a standard deviation of 1.7 minutes. If the distribution of times is approximately mound-shaped, what proportion of workers will take longer than 16.2 minutes to complete the task?

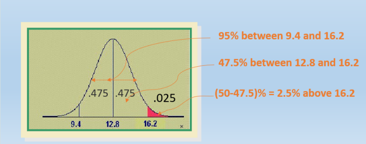
Solution:

Let X be the time to complete the task (in minutes). We are given $\mu = 12.8, \sigma = 1.7$ and the distribution is approximately mound-shaped (symmetric and bell-shaped). We need the proportion of workers with $X > 16.2$.

- One can notice that 16.2 is 2 standard deviations above the mean:

$$16.2 = \mu + 2\sigma = 12.8 + 2(1.7) = 12.8 + 3.4.$$

- Based on empirical rule, 95% of observations fall within two standard deviations of the mean, i.e., between 9.4 and 16.2. Consequently, 5% of observations fall outside this interval.
- Because the distribution is symmetric, half of that 5% is above 16.2 and half is below 9.4.
- Thus, the required proportion is $\dfrac{0.05}{2} = 0.025$. i.e., about 2.5% of workers will take longer than 16.2 minutes to complete the task.



Measures of Relative Standing: Z-scores

- Z-scores Measure relative standing of every score in a distribution.
- Take different distributions and make them equivalent and comparable by standardizing and providing a common metric.

Formula of z-score

$$\textbf{Population: } Z = \frac{X - \mu}{\sigma} \Rightarrow X = \mu + Z\sigma$$

$$\textbf{Sample: } Z = \frac{X - \bar{x}}{s} \Rightarrow X = \bar{x} + Zs$$

Interpretation:

(i) Sign $(+, -)$ of Z score tells whether the score (X) is located above (if $+$) or below (if $-$) the mean.

(ii) Value of z-score tells distance between score and mean in standard deviation units.

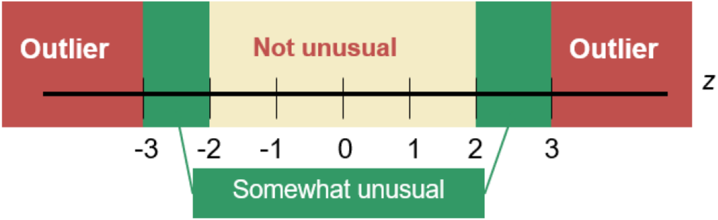
Example: IQ scores with $\mu = 100$ and $\sigma = 15$. Find z-score for $X = 121$.

$$Z = \frac{X - \mu}{\sigma} = \frac{121 - 100}{15} = 1.4$$

Interpretation: The person with an IQ of 121 is 1.4 standard deviations above the mean.

Detection of Outlier (Using z-scores and Empirical Rule)

Note: No matter the original data, after conversion to z-scores, the new mean is 0 and the new standard deviation is 1.



Measures of Relative Standing: Percentiles

What Are Percentiles?

A **percentile** is a value below which a given percentage of observations fall.

Example

If you score at STAT 2601 (say 67 out of 100) at the 85th percentile on a test:

- Your score 67 is higher than 85% of all test-takers
- Only 15% scored higher than you (67 out of 100)

Common Percentiles

- 50th percentile = **Median** (middle value)
- 25th percentile = **First Quartile = Lower Quartile (Q1)**
- 75th percentile = **Third Quartile = upper Quartile (Q3)**
- 0th percentile = Minimum (X_{min})
- 100th percentile = Maximum (X_{max})

Linear Interpolation Method for Percentiles

Given sorted data $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$:

Step 1: Calculate position L_p for p-th percentile:

$$L_p = \frac{p}{100} \times (n+1)$$

Step 2: Determine the percentile value:

- If L_p is an integer: Percentile = value at position L_p
- If L_p is NOT an integer:
  - Let k = integer part of L_p
  - Let d = decimal part of L_p
  - Percentile = $x_{(k)} + d \times (x_{(k+1)} - x_{(k)})$

For example, suppose $n = 9$, the location of the 20th Percentile is $L_{20} = \frac{20}{100} \times (9+1) = 2$, is the 2nd value $= x_{(2)}$ in the sorted list, whereas the position of the 75th Percentile is $L_{75} = \frac{75}{100} \times (9+1) = 7.5$ is $P_{75} = x_{(7)} + 0.5 \times (x_{(8)} - x_{(7)})$ after ranking the data from the smallest to largest values.

Calculating Percentiles: 3 Simple Steps

Dataset: Monthly Car Sales (units)

$$12, \quad 15, \quad 18, \quad 22, \quad 25, \quad 35, \quad 40, \quad 28, \quad 30, \quad 45 \quad (n = 10)$$

Step 1: Sort Data

$12, 15, 18, 22, 25, 28, 30, 35, 40, 45$

Always sort first!

Step 2: Find Position

$$L_p = \frac{p}{100} \times (n + 1)$$

For 30th percentile:

$$L_{30} = \frac{30}{100} \times (10 + 1) = 3.3$$

Step 3: Calculate Value

Position $3.3 = 3\text{rd} + 0.3(4\text{th} - 3\text{rd})$

$$P_{30} = 18 + 0.3(22 - 18) = 18 + 1.2$$

$$\boxed{P_{30} = 19.2}$$

Interpretation (note car sale units is an integer number!)

30% of months had sales $\leq$ 19 cars; 70% had sales > 19 cars

Example: Calculate the 80th Percentile

Dataset: Employee Commute Times (minutes)

$$15, \quad 20, \quad 25, \quad 30, \quad 35, \quad 40, \quad 45, \quad 50, \quad 60, \quad 70 \quad (n = 10)$$

Step-by-Step Solution

1. Already sorted
2. $L_{80} = \frac{80}{100} \times (10 + 1) = 0.8 \times 11 = 8.8$
3. Position 8.8 is between 8th and 9th values (50 and 60)
4. $P_{80} = 50 + 0.8 \times (60 - 50) = 58$

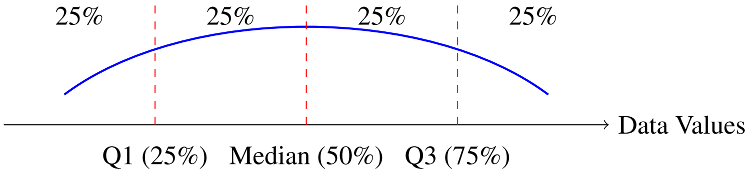
Interpretation

- 80% of employees have commute times ≤ 58 minutes
- 20% have commute times > 58 minutes
- If your commute is 58 minutes, you're at the 80th percentile

Quick Check!

What does this mean for an employee with a 70-minute commute?

Quartiles: Dividing Data into Quarters



Always sort the data first!

The Three Quartiles

- **Q1**: 25th percentile
- **Q2**: 50th percentile = Median
- **Q3**: 75th percentile

Interquartile Range (IQR)

$$IQR = Q3 - Q1$$

- Spread of middle 50% of data
- Not affected by outliers

Location of Q1

$$L_{Q1} = \frac{n+1}{4}$$

Location of Q2

$$L_{M_d} = \frac{n+1}{2}$$

Location of Q3

$$L_{Q3} = \frac{3}{4}(n+1)$$

Calculating Quartiles: Car Sales Example

Dataset (already sorted)

$$12, \quad 15, \quad 18, \quad 22, \quad 25, \quad 28, \quad 30, \quad 35, \quad 40, \quad 45$$

Q1 (25th Percentile)

$$\begin{aligned} L_{25} &= 0.25 \times 11 = 2.75 \text{ (position)} \\ Q1 &= x_2 + 0.75(x_3 - x_2) \\ &= 15 + 0.75(18 - 15) \\ &= 15 + 2.25 \\ &= \boxed{17.25} \end{aligned}$$

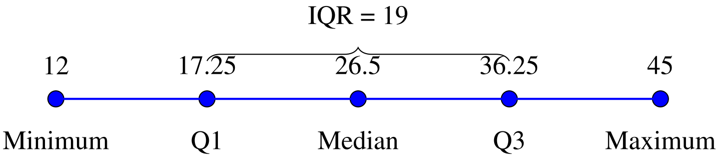
Q3 (75th Percentile)

$$\begin{aligned} L_{75} &= 0.75 \times 11 = 8.25 \text{ (position)} \\ Q3 &= x_8 + 0.25(x_9 - x_8) \\ &= 35 + 0.25(40 - 35) \\ &= 35 + 1.25 \\ &= \boxed{36.25} \end{aligned}$$

Median ($Q2 = P_{50}$) and IQR

- Median = average of 5th and 6th values $= \frac{25+28}{2} = \boxed{26.5}$
- $IQR = Q3 - Q1 = 36.25 - 17.25 = \boxed{19}$

Five-Number Summary: Complete Picture



Why Five Numbers?

A quick summary of the data distribution:

- **Min, Max**: Range and extremes
- **Q1, Q3**: Spread of middle 50% (IQR)
- **Median**: Center of distribution
- **Complete**: Shows center, spread, and extremes

Building a Box-and-Whisker Plot

Ingredients

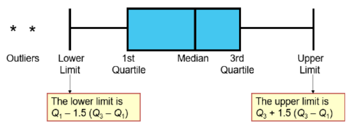
- Five-number summary
- $IQR = Q3 - Q1$
- Whisker = $1.5 \times IQR$
- Lower fence: $Q1 - 1.5 \times IQR$
- Upper fence: $Q3 + 1.5 \times IQR$

For Our Car Sales

- $Q1 = 17.25, Q3 = 36.25$
- IQR = 19
- Lower fence $= 17.25 - 28.5 = -11.25$
- Upper fence $= 36.25 + 28.5 = 64.75$
- All data within fences, so no outliers.

Construction Steps

1. Draw number line
2. Draw box from Q1 to Q3
3. Draw line at median
4. Plot outliers (values outside the lower/upper fence) and denoted the outliers (if exist!) by $\star$
5. Draw whiskers to min/max within fences (smallest/largest data point that is not counted as an outlier)



What Boxplots Tell Us About Distribution Shape

✓Median line in center of box and whiskers of equal length—symmetric distribution

- Approximately symmetric, Mean = median
- i.e. $(Q_2\text{-}Q_1) \approx (Q_3\text{-}Q_2)$

✓Median line left of center and long right whisker—skewed right

- Shifted to the right, mean > median
- i.e. $(Q_2\text{-}Q_1) < (Q_3\text{-}Q_2)$

✓Median line right of center and long left whisker—skewed left

- Shifted to the left, mean < median
- i.e. $(Q_2\text{-}Q_1) > (Q_3\text{-}Q_2)$

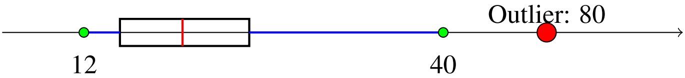
Detecting Outliers: The 1.5×IQR Rule

- Lower fence $= Q1 - 1.5 \times IQR$
- Upper fence $= Q3 + 1.5 \times IQR$
- Values outside fences (less than lower fence or greater than upper fence) are Potential outliers

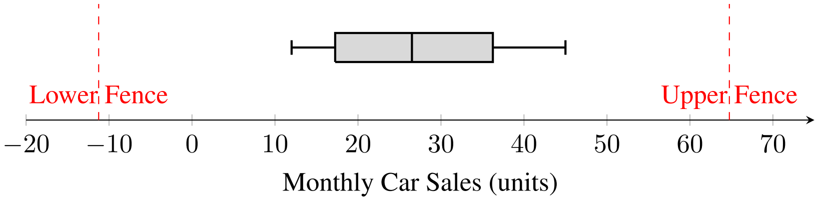
Example: Sales with Potential Outlier

Dataset: $12, 15, 18, 22, 25, 28, 30, 35, 40, 80$

- $Q1 = 17.25, Q3 = 36.25, IQR = 19$
- Upper fence $= 36.25 + 1.5 \times 19 = 64.75$
- $80 > 64.75 \rightarrow$ Outlier detected!
- Lower fence $= 17.25 - 1.5 \times 19 = -11.25 \rightarrow$ no outlier detected



Boxplot: Monthly Car Sales



Interpretation

- Middle 50% of sales: from 17.25 to 36.25 units
- Median sales: 26.5 units
- No outliers (all points within fences)
- Distribution appears symmetric (median centered in box)

Example: Amount of sodium in 8 brands of cheese

Create a box plot to display the distribution of sodium levels in the following 8 cheese brands:

$$260, \quad 290, \quad 300, \quad 320, \quad 330, \quad 340, \quad 340, \quad 520$$

The five numbers are:

$\textbf{min} = 260, \quad \textbf{max} = 520, \quad Q_1 = 292.5 \quad Q_2 = 325 \quad Q_3 = 340$

$$L_{Q_1} = \frac{n+1}{4} = \frac{9}{4} = 2.25 \Rightarrow Q_1 = x_{(2)} + 0.25(x_{(3)} - x_{(2)}) = 290 + 0.25(300 - 290) = 292.5$$

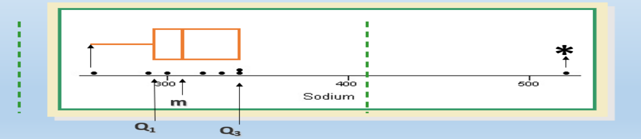
$$L_{Q_2} = \frac{n+1}{2} = \frac{9}{2} = 4.5 \Rightarrow Q_2 = (x_{(4)} + x_{(5)})/2 = (320 + 330)/2 = 325$$

$$L_{Q_3} = \frac{3}{4}(n+1) = \frac{27}{4} = 6.75 \Rightarrow Q_3 = x_{(6)} + 0.75(x_{(7)} - x_{(6)}) = 340 + 0.75(340 - 340) = 340$$

$IQR = Q_3 - Q_1 = 340 - 292.5 = 47.5, \quad \text{Whisker } = 1.5 \times IQR = 1.5 \times 47.5 = 71.25$

Lower fence $= Q_1 - 1.5 \times IQR = 292.5 - 1.5(47.5) = 221.25$ (no outliers below this value)

Upper fence $= Q_3 + 1.5 \times IQR = 340 + 1.5(47.5) = 411.25$ (520 is an outlier)



Distribution is right skewed.

Central Tendency

- **Mean**: $\bar{x} = \frac{\sum x_i}{n}$
- **Median**: Middle value when sorted
- **Mode**: Most frequent value(s)

Variation

- **Range**: Max $-$ Min
- **Variance**:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$$

- **St. Dev.**: $s = \sqrt{s^2}$
- **CV**: $CV = \frac{s}{\bar{x}} \times 100\%$

Relative Standing

- **z-score**: $z = \frac{x - \bar{x}}{s}$
- **Percentiles**: Relative position
- **Quartiles**: Q1, Q2, Q3

Key Formulas

Mean: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$

Variance (sample): $s^2 = \dfrac{\sum(x_i - \bar{x})^2}{n-1} = \dfrac{\sum x^2 - \frac{(\sum x)^2}{n}}{n-1}$

z-score: $z = \dfrac{x - \bar{x}}{s}$

CV: $CV = \dfrac{s}{\bar{x}} \times 100\%$

Percentile position: $L_p = \dfrac{p}{100}(n+1)$

Q1 position: $L_{Q1} = \dfrac{n+1}{4}$

Q2 position: $L_{Q2} = \dfrac{n+1}{2}$

Q3 position: $L_{Q3} = \dfrac{3(n+1)}{4}$

IQR: $IQR = Q3 - Q1$

Lower fence: $Q1 - 1.5 \times IQR$

Upper fence: $Q3 + 1.5 \times IQR$

Empirical Rule (Normal Distribution)

68% within $\mu \pm 1\sigma$, 95% within $\mu \pm 2\sigma$, 99.7% within $\mu \pm 3\sigma$

Five-Number: Min, Q1, Median, Q3, Max | **Boxplot shows:** Shape, Center, Spread, Outliers