Central Tendency and Dispersion

Md. Aminul Islam Shazid

Outline

1. Central Tendency
2. Dispersion
3. Shape of Distribution
4. Outlier
5. Boxplot

Central Tendency

Central Tendency

- Observations of a variable tend to gather around a single value, this is known as central tendency
- Central tendency is a descriptive measure that represents the center or typical value of a variable
- It provides a summary of the values of the variable

Central Tendency (cont.)

- Mean:
  - Arithmetic mean
  - Geometric mean
  - Harmonic mean
- Median
- Mode

These are different *measures* of central tendency. They represent the "average" value of a dataset in different ways.

Depending on the shape of the distribution and the presence of outliers, different measures are used.

Characteristics of a Good Measure

- Clear and unambiguous definition so that the same data provides the same value of the measure
- Easy to understand and calculate
- Based on all or most of the observations in the sample
- Not unduly affected by outliers so that a few outliers does not distort the result too much
- Representative of the distribution so that the value lies within the range of the data and and describe its central location
- Capable of further mathematical treatment so that it can be used for further analysis

Arithmetic Mean

- The arithmetic mean is the sum of all observations divided by the number of observations
- The population mean is denoted by μ
- For a sample of values $x_1, x_2, \ldots, x_n$ of a variable X, the sample mean is

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

- It uses all observations in the dataset
- The arithmetic mean is easy to compute and interpret
- It is *sensitive* to extreme values (outliers)
- Therefore, it is most appropriate for numerical data that are symmetrically distributed

Example: Arithmetic Mean From Frequency Table

Value, x_i	**Frequency**, f_i	$f_i \cdot x_i$
55	7	385
60	10	600
62	6	372
65	4	260
67	3	201
Total:	30	1818

The mean, $\bar{x} = \frac{\sum_{i=1}^{n} f_i x_i}{\sum_{i=1}^{n} f_i} = \frac{1818}{30} = 60.6$.

If the data is grouped, then the class midpoints are treated as x_i.

Arithmetic Mean for Grouped Data

Class interval	**Class midpoint**, x_i	**Frequency**, x_i	$f_i \cdot x_i$
5-30	17.5	7	122.5
30-55	42.5	10	425
55-80	67.5	6	405
80-105	92.5	4	370
105-130	117.5	3	352.5
	Total:	30	1675

Mean = $1675/30 = 55.83$

Weighted Mean

- When caluclating average, sometimes some values may be more important than other values
- In the previous example, the observations appeared different number of times
- Therefore, each value has different level of influence over the center of the distribution
- This is called the weight of each value
- Another example is the calculation of CGPA where the total credit of each semester is the weight of the corresponding GPA

Geometric Mean

- The geometric mean is a measure of central tendency defined as the n-th root of the product of n positive observations
- For positive data $x_1, x_2, \ldots, x_n$, the geometric mean is

$$G = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

- It is only defined for positive values
- The geometric mean is appropriate for data involving ratios, rates, or growth factors
- It reduces the influence of very large values compared to the arithmetic mean
- The geometric mean is commonly used for percentage changes and financial returns

Geometric Mean for Grouped Data

- For grouped data, the geometric mean is calculated using class frequencies
- Let $x_1, x_2, \ldots, x_k$ be the class midpoints and $f_1, f_2, \ldots, f_k$ the corresponding frequencies
- The geometric mean is given by

$$G = \left(\prod_{i=1}^{k} x_i^{f_i} \right)^{1/n},$$

  where $n = \sum_{i=1}^{k} f_i$
- In practice, the computation is often simplified using logarithms:

$$\log G = \frac{1}{n} \sum_{i=1}^{k} f_i \log x_i$$

Harmonic mean

- The harmonic mean is a measure of central tendency defined as the reciprocal of the arithmetic mean of reciprocals
- For positive data $x_1, x_2, \dots, x_n$, the harmonic mean is

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

- It is only defined for positive values
- The harmonic mean gives more weight to smaller observations
- It is appropriate for averaging rates or ratios, such as speeds or densities
- The harmonic mean is strongly affected by very small values

Harmonic Mean for Grouped Data

- For grouped data, the harmonic mean is calculated using class frequencies
- Let $x_1, x_2, \ldots, x_k$ be the class midpoints and $f_1, f_2, \ldots, f_k$ the corresponding frequencies.
- The harmonic mean is given by

$$H = \frac{n}{\sum_{i=1}^{k} \frac{f_i}{x_i}},$$

where $n = \sum_{i=1}^{k} f_i$

Mode

- The mode is the value that occurs most frequently in a dataset
- A dataset may have:
  - one mode (unimodal),
  - two modes (bimodal), or
  - more than two modes (multimodal)
- The mode can be used for both numerical and categorical data
- A dataset may have no mode if all values occur with the same frequency
- The mode is not affected by extreme values
- For grouped data, the mode is estimated using the modal class

Mode for Grouped Data

$$\text{Mode} = L_0 + \frac{l_1}{l_1 + l_2} \times c,$$

where:

- L_0 is the lower limit of the modal class (class with the highest frequency)
- l_1 is the difference in Frequency between the modal class and the pre-modal class
- l_2 is the difference in Frequency between the modal class and the post-modal class
- c is the class interval

Example: Mode for Grouped Data

Class interval	Frequency
5-30	7
30-55	10
55-80	6
80-105	4
105-130	3

Mode = $30 + \frac{3}{3+4} \times 30 = 40.71$

Median

- The median is the middle value of a dataset when the observations are arranged in ascending or descending order
- If the number of observations n is odd, the median is the $(\frac{n+1}{2})^{\text{th}}$ observation
- If n is even, the median is the average of the $(\frac{n}{2})^{\text{th}}$ and $(\frac{n}{2}+1)^{\text{th}}$ observations
- The median divides the dataset into two equal halves
- Therefore, it is the value below which 50% of the data lies
- It is not affected by extreme values (outliers)
- Therefore, it is useful for skewed distributions or data with outliers

Median for Grouped Data

$$\text{Median} = L_m + \frac{\frac{n}{2} - F_c}{f_m} \times c,$$

where:

- L_m = lower limit of the median group, it is the group in which relative cumulative frequency is equal to 0.5 (50%) or the first group in which relative cumulative frequency exceeds 0.5
- n = sample size
- F_c = cumulative frequency of the pre-median class
- f_m = frequency of the median class

Example: Median for Grouped Data

Class interval	Frequency	Cumulative Frequency
5-30	7	7
30-55	10	17
55-80	6	23
80-105	4	27
105-130	3	30

Here, sample size is 30. Since 50% of 30 is 15, the second group is the median class.

Median = $L_m + \frac{\frac{n}{2} - F_c}{f_m} \times c = 30 + \frac{\frac{30}{2} - 7}{10} \times 25 = 50.$

Trimmed Mean

- The trimmed mean is a measure of central tendency obtained by removing a fixed proportion of the smallest and largest observations
- After trimming, the arithmetic mean is computed using the remaining data
- A $p\%$ trimmed mean removes the lowest $p\%$ and highest $p\%$ of the data
- It is less sensitive to extreme values than the arithmetic mean
- The trimmed mean provides a balance between the mean and the median
- It is useful when the data contain outliers or are moderately skewed

Quantile

- Quantiles are values that divide an ordered dataset into equal parts
- Each part contains the same proportion of observations
- Common quantiles include:
  - Quartiles: divide the data into four equal parts
  - Deciles: divide the data into ten equal parts
  - Percentiles: divide the data into one hundred equal parts
- The median is the second quartile (Q_2) or the 50th percentile, it divides the data in two parts
- Quantiles are useful for describing the distribution and spread of data

Quartile

- Quartiles are values that divide an ordered dataset into four equal parts
- Each part contains approximately 25% of the observations
- The three quartiles are:
  - First quartile (Q_1): 25th percentile
  - Second quartile (Q_2): 50th percentile (the median)
  - Third quartile (Q_3): 75th percentile
- Quartiles are used to describe the spread and position of data
- Sometimes, the minimum value is referred to as the 0th quartile and the maximum value as the 4th quartile

Percentile

- Percentiles divide an ordered dataset into 100 equal parts
- Each percentile represents 1% of the observations
- The p-th percentile is the value below which $p\%$ of the data lie
- The median is the 50^{th} percentile, the first quartile is the 25^{th} percentile, the third quartile is the 75^{th} percentile
- Percentiles are widely used in examinations, test scores, and rankings

Someone's IQ score being 90th percentile means that the score is above 90% of the population.

Calculating Percentile

- First, arrange the data in ascending order
- In order to find the i^{th} percentile, p_i, calculate: $\frac{i \cdot n}{100}$
- If $\frac{i \cdot n}{100}$ is an integer, then the average of the $\left(\frac{i \cdot n}{100}\right)^{\text{th}}$ value and the value to its right, is p_i
- If $\frac{i \cdot n}{100}$ is not an integer, then the integer to the right of $\frac{i \cdot n}{100}$ is p_i

Example: Calculating Percentile

Find the 90th and 20th percentiles from the test scores of 25 students: $43, 54, 56, 61, 62, 66, 68, 69, 69, 70, 71, 72, 77, 78, 79, 85, 87, 88, 89, 93, 95, 96, 98, 99, 99$.

90^{th} percentile: $\frac{90 \times 20}{25} = 22.5$. Therefore, the 23^{rd} value is the 90^{th} percentile, which is 98.

20^{th} percentile: $\frac{20 \times 20}{25} = 5$. Therefore, the average of the 5^{th} and the 6^{th} values, is the desired percentile, which is $\frac{62+66}{2} = 64$.

Decile

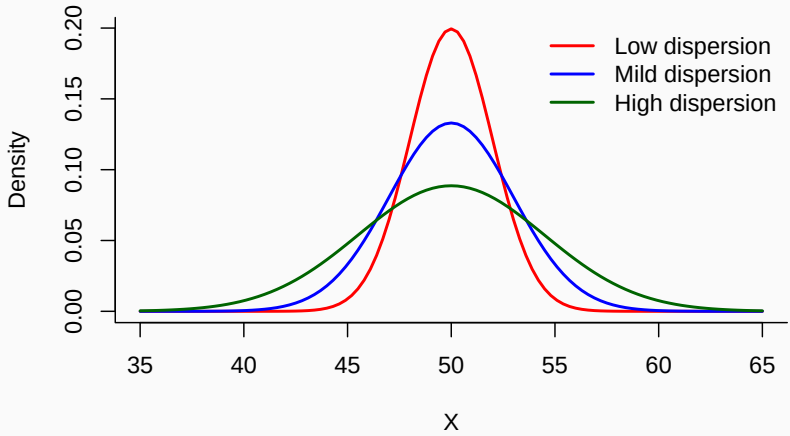
- Deciles divide an ordered dataset into 10 equal parts
- Each decile represents 10% of the observations
- The k^{th} decile (D_k) is the value below which $k \times 10\%$ of the data lie
- The fifth decile (D_5) coincides with the median
- Deciles are useful for studying the distribution of data in broader groups
- Can be calculated the same way as percentiles are calculated

Dispersion

Dispersion

- Dispersion describes the extent to which data values are spread out
- It indicates the variability or consistency present in a dataset
- A low dispersion indicates that data values are close to each other, while a high dispersion indicates greater variability
- Measures of dispersion complement measures of central tendency
- Common measures of dispersion include:
  - Range
  - Interquartile range
  - Mean deviation, Mean absolute deviation
  - Variance and standard deviation
- Dispersion is essential for comparing datasets with similar central tendencies.

Comparing Different Levels of Dispersion



Range

- The range is the simplest measure of dispersion
- It gives a quick idea of the overall spread of the data
- It is defined as the difference between the largest and smallest observations in a dataset
- If the minimum value is $\min(x)$ and the maximum value is $\max(x)$, then

$$\text{Range} = \max(x) - \min(x)$$

- It is highly sensitive to extreme values (outliers)
- The range does not use all observations and provides only a rough measure of variability

Inter-quartile Range

- The interquartile range (IQR) is a measure of dispersion based on quartiles
- It is defined as the difference between the third and first quartiles

$$\text{IQR} = Q_3 - Q_1$$

- The IQR measures the spread of the middle 50% of the data
- Threfore it is not affected by extreme values or outliers
- The IQR is particularly useful for skewed distributions
- It is commonly used to identify outliers

Mean Deviation

- The mean deviation is defined as the arithmetic mean of deviations from the mean
- For observations $x_1, x_2, \ldots, x_n$ with mean $\bar{x}$,

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

- The positive and negative deviations cancel each other out
- As a result, the value of this mean deviation is always zero
- Because it is always zero, it is not useful as a measure of dispersion
- This limitation motivates the use of absolute deviations or squared deviations

Mean Absolute Deviation

- Mean absolute deviation is defined as the arithmetic mean of the absolute deviations of observations from a central value
- For observations $x_1, x_2, \dots, x_n$ with mean $\bar{x}$,

$$\text{Mean Deviation} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|.$$

- Mean deviation uses all observations in the dataset
- It is less affected by extreme values than variance and standard deviation

Variance

- It is based on squared deviations from the mean and measures how far observations spread out from the mean, on average
- For a population, the variance is

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

- For a sample, the variance is

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

- Squaring the deviations removes cancellation of positive and negative values
- Variance uses all observations and is sensitive to extreme values
- The units of variance are the square of the original data units

Standard Deviation

- Standard deviation is a widely used measure of dispersion
- It measures the average spread of observations around the mean
- It is defined as the square root of the variance.
- For a population, the standard deviation is

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

- For a sample, the standard deviation is

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

- Standard deviation uses all observations and is sensitive to extreme values
- It is expressed in the same units as the original data

Coefficient of Variation

- The coefficient of variation is a relative measure of dispersion (the ones before this were absolute measures)
- It expresses variability relative to the mean of the dataset
- The coefficient of variation is defined as

$$\text{CV} = \frac{\text{standard deviation}}{\text{mean}}.$$

- It is often expressed as a percentage:

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

- The coefficient of variation is unit-free
- It is useful for comparing variability between datasets with different units or different means
- The coefficient of variation is meaningful only when the mean is non-zero

Example

Calculate measures of dispersion from the weight of ten newborn babies (in pounds): $7.5, 4.5, 10.1, 9.6, 5.5, 6.6, 7.8, 5.9, 6.0, 5.5$.

Range = $10.1 - 4.5 = 5.6$ pounds

Mean, $\bar{x} = \frac{7.5+4.5+\ldots+5.5}{10} = 6.9$ pounds

Variance, $s^2 = \frac{(7.5-6.9)^2+(4.5-6.9)^2+\ldots+(5.5-6.9)^2}{10-1} = \frac{30.28}{9} = 3.36\,\text{pound}^2$

Standard deviation (SD), $s = \sqrt{3.36} = 1.83$ pound

Coefficient of Variation, CV $= \frac{s}{\bar{x}} = \frac{1.83}{6.9} = 0.2652$

Shape of Distribution

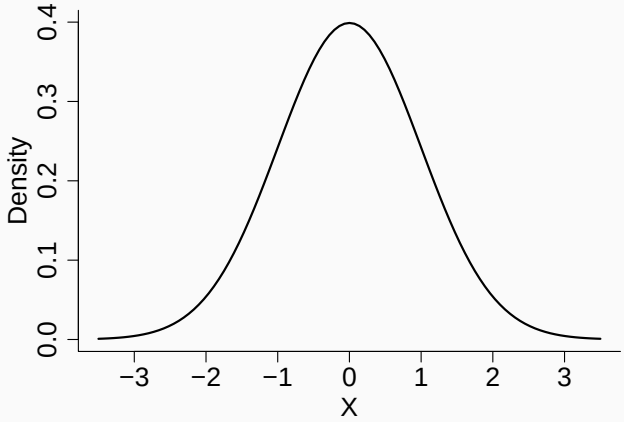
Shape of Distribution

- **Shape of distribution**: describes the overall form of the data, including symmetry, number of peaks, and tail behaviour.
- **Skewness**: indicates the direction and degree of asymmetry in a distribution.
- **Kurtosis**: reflects the peakedness of the distribution and the heaviness of its tails.
- These characteristics help summarize how data are distributed beyond measures of central tendency.
- They guide the choice of appropriate statistical methods and models.

Skewness

- **Skewness** measures the asymmetry of a distribution around its center.
- A distribution is **symmetric** if the left and right sides are mirror images.
- **Positive (right) skewness**: longer tail on the right side.
- **Negative (left) skewness**: longer tail on the left side.
- Skewness affects the relative positions of the mean, median, and mode.

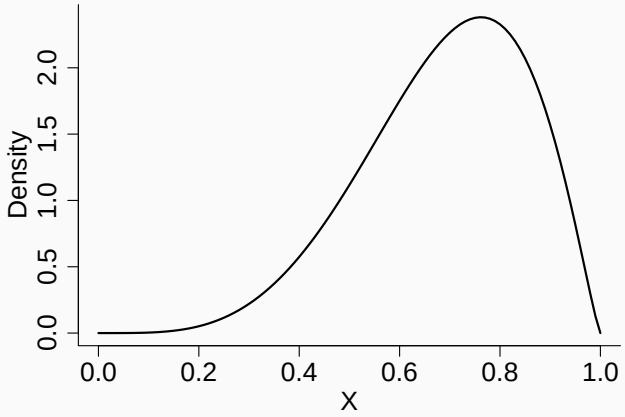
Symmetric Distribution



Above is a symmetric (bell shaped) distribution.

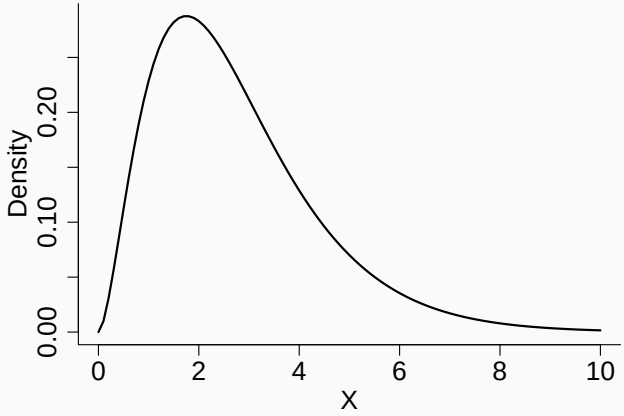
In such cases: mean = median = mode.

Left Skewed Distribution



Here: mode > median > mean.

Right Skewed Distribution



Here: mean > median > mode.

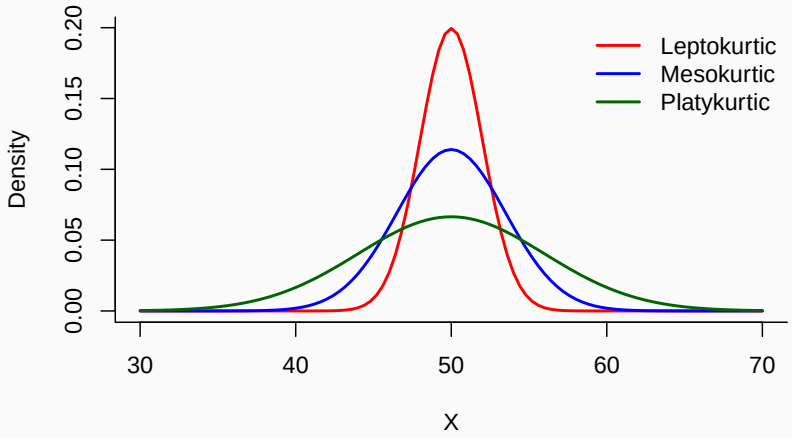
Caluclating Skewness

- Pearson's Coefficient of Skewness $= \frac{3(\text{mean} - \text{median})}{\text{Standard deviation}}$
- Bowley's coefficient of skewness $= \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_1)}$
- If the coefficient is positive, then the distribution is right skewed
- Left skewed if the coefficient is negative
- Symmetric if the coefficient is zero

Kurtosis

- **Kurtosis** describes the peakedness and tail heaviness of a distribution
- It indicates how concentrated the data are around the mean
- **Leptokurtic**: sharper peak with heavier tails
- **Mesokurtic**: peak not too sharp and tails neither heavy nor light
- **Platykurtic**: flatter peak with lighter tails

Kurtosis Comparison



Calculating Kurtosis

$$\text{Kurtosis} = \frac{\frac{\sum(x_i - \bar{x})^4}{n}}{\left(\frac{\sum(x_i - \bar{x})^2}{n}\right)^2} - 3$$

- If kurtosis = 0, then the distribution is mesokurtic
- If it is positive, then leptokurtic
- If it is negative, then platykurtic

Outlier

Outlier

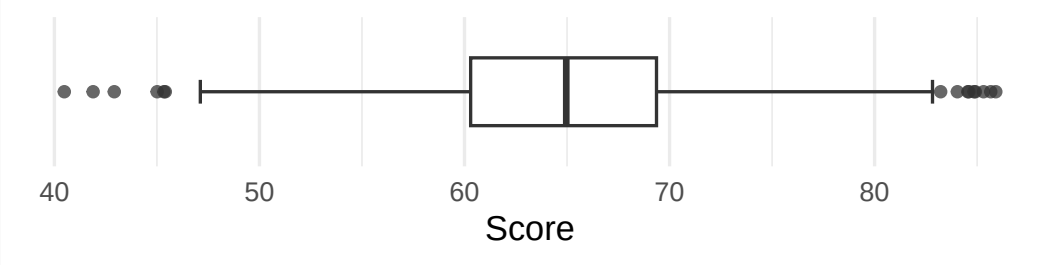
- Outliers are observations that are different from the rest of the data
- They may arise due to measurement errors, data entry errors, or genuine extreme values
- Outliers can strongly affect measures such as the mean, variance, and standard deviation
- Measures like the median and interquartile range are more resistant to outliers
- A common rule for identifying outliers is based on the interquartile range:

$$\text{Lower bound} = Q_1 - 1.5 \times \text{IQR}, \quad \text{Upper bound} = Q_3 + 1.5 \times \text{IQR}$$

- Values outside this range are considered outliers
- Outliers should be investigated carefully before being removed

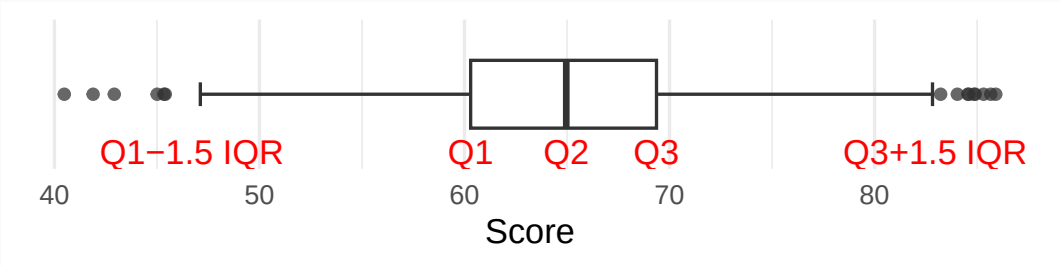
Boxplot

Boxplot



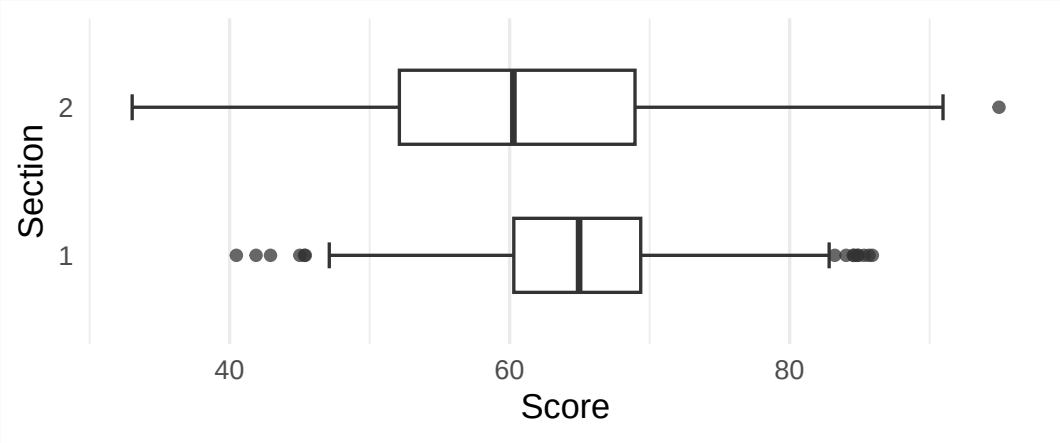
- A boxplot is a graphical method for summarizing the distribution of a dataset
- It is based on the five-number summary: a lower fence, Q_1, median (Q_2), Q_3, and an upper fence
- The box represents the interquartile range ($Q_3 - Q_1$)
- The line inside the box indicates the median

Boxplot (cont.)



- **Lower fence** $= Q_1 - 1.5 \times \text{IQR}$, **upper fence** $= Q_3 + 1.5 \times \text{IQR}$
- Whiskers extend to the smallest and largest non-outlier observations (lower fence and upper fence)
- Observations beyond the whiskers are plotted individually as outliers
- Thus boxplots can be used to detect outliers graphically

Comparing Groups with Boxplot



Section 2 has lower average (median) score, but its maximum score is higher than that of section 1. Further, section 2 has higher variability.

Questions?