IAQS | INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: MSc

Subject : Probability and Statistics

Chapter: Unit 1 Chapter 2

Chapter Name: Descriptive Statistics

Topics to be covered

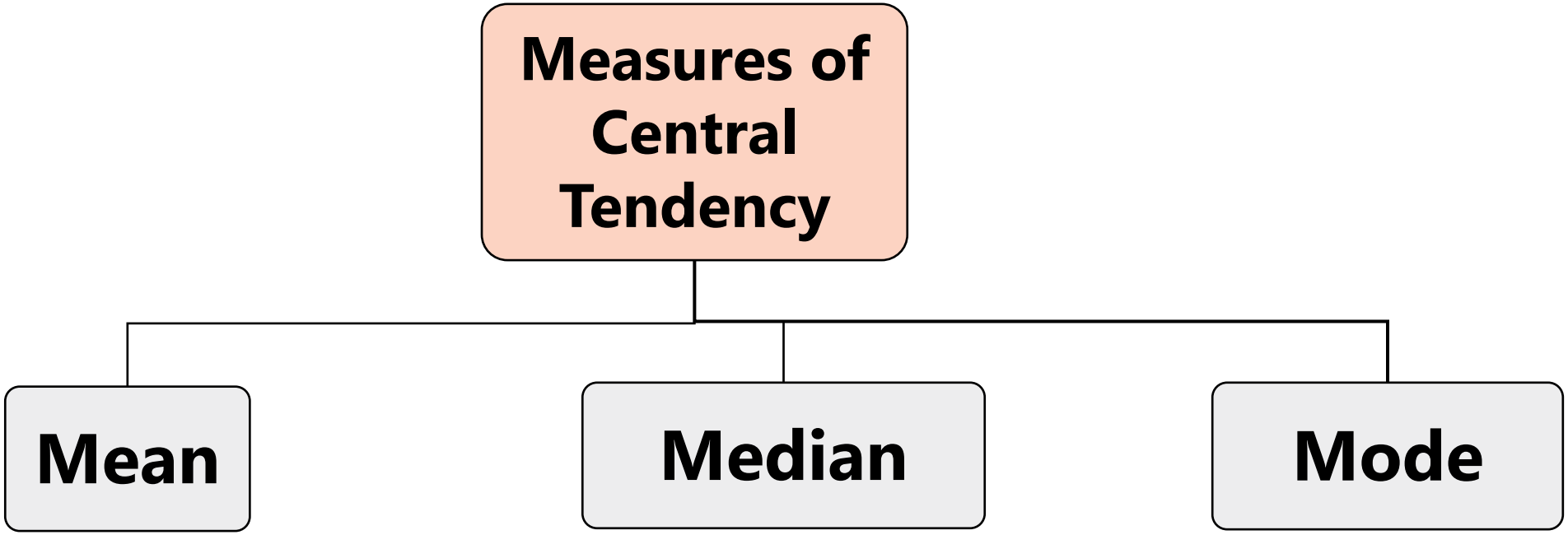
1. Measures of Central Tendency
2. The Mean
3. The Median
4. The Mode
5. Mean, Median & Mode
6. Measures of Dispersion
7. Range
8. The problem with outliers
9. Quartiles
10. Variance and Standard Deviation
11. Moments
12. Skewness
13. Skewness and Central Tendencies

1 Measures of Central Tendency

A measure of central tendency (also referred to as measures of centre or central location) is a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution.

- There are three main measures of central tendency: the mode, the median and the mean. Each of these measures describes a different indication of the typical or central value in the distribution.

1 Measures of Central Tendency



2 The mean

The mean is the sum of the value of each observation in a dataset divided by the number of observations. This is also known as the arithmetic average.

- The mean is one of the most commonly used statistics around, and statisticians use it so frequently that they've given it a symbol all of its own: . This is the Greek letter mu ($\boldsymbol{\mu}$) (pronounced "mew"). Remember, it's just a quick way of representing the population mean. The sample mean is represented by $\bar{x}$.
- The mean is calculated as, $\boldsymbol{\mu} = \frac{\sum \boldsymbol{x}}{\boldsymbol{n}}$

Advantage of the mean:

- The mean can be used for both continuous and discrete numeric data.

Limitations of the mean:

- The mean cannot be calculated for categorical data, as the values cannot be summed.
- As the mean includes every value in the distribution the mean is influenced by outliers and skewed distributions.

3 The median

The median is the middle value in distribution when the values are arranged in ascending or descending order.

- The median divides the distribution in half (there are 50% of observations on either side of the median value). In a distribution with an odd number of observations, the median value is the middle value.

Advantage of the median:

- The median is less affected by outliers and skewed data than the mean, and is usually the preferred measure of central tendency when the distribution is not symmetrical.

Limitation of the median:

- The median cannot be identified for categorical nominal data, as it cannot be logically ordered.

3 The median

Finding the median

Line your numbers up in order, from smallest to largest.

If you have an odd number of values, the median is the one in the middle. If you have n numbers, the middle number is at position: (n+1)/2

If you have an even number of values, get the median by adding the two middle ones together and dividing by 2 You can find the midpoint by calculating (n + 1) / 2. The two middle numbers are on either side of this point.

4 The mode

The mode is the most commonly occurring value in a distribution.

Advantage of the mode:

- The mode has an advantage over the median and the mean as it can be found for both numerical and categorical (non-numerical) data.

Limitations of the mode:

- The are some limitations to using the mode. In some distributions, the mode may not reflect the centre of the distribution very well.
- It is also possible for there to be more than one mode for the same distribution of data, (bi-modal, or multi-modal). The presence of more than one mode can limit the ability of the mode in describing the centre or typical value of the distribution because a single value to describe the centre cannot be identified.
- In some cases, particularly where the data are continuous, the distribution may have no mode at all (i.e. if all values are different).
- In cases such as these, it may be better to consider using the median or mean, or group the data in to appropriate intervals, and find the modal class.

4 The mode

In addition to the mean and median, there's a third type of average called the ***mode***. The mode of a set of data is the most popular value, the value with the highest frequency. Unlike the mean and median, the mode absolutely *has* to be a value in the data set, and it's the most frequent value.

Sometimes data can have more than one mode. If there is more than one value with the highest frequency, then each one of these values is a mode. If the data looks as though it's representing more than one trend or set of data, then we can give a mode for each set. If a set of data has two modes, then we call the data ***bimodal***.

It even works with categorical data

5 Mean, Median & Mode

Mean	Median	Mode
• The mean is the most widely spread measure of central tendency. It is the simple average of the dataset. • **Note**: easily affected by outliers	• The median is the midpoint of the ordered dataset. It is not as popular as the mean, but is often used in academia and data science. That is since it is not affected by outliers.	• The mode is the value that occurs most often. A dataset can have 0 modes, 1 mode or multiple modes.

5 Mean, Median & Mode

Mean

- The formula to calculate the mean is:
- $\sum_{i=1}^{n} x_i \,/N$ or
- $\frac{x_1+ x_2+x_3+\cdots+x_N}{N}$

Median

- In an ordered dataset, the median is the number at position $\frac{n+1}{2}$
- If this position is not a whole number, it, the median is the simple average of the two numbers at positions closest to the calculated value.

Mode

- The mode is calculated simply by finding the value with the highest frequency.

5 Mean, Median & Mode

Average	How to calculate	When to use it
Mean (μ)	Use Either $\frac{\sum x}{n}$ or $\frac{\sum fx}{\sum x}$	When the data is fairly symmetric and shows just the one trend.
Median	• Line up all the values in ascending order. • If there are an odd number of values, the median is the one in the middle. • If there are an even # of values, add the two middle ones together, and divide by two.	When the data is skewed because of outliers.
Mode	• Choose the value(s) with the highest frequency. • If the data is showing two clusters of data, report a mode for each group.	• When you're working with categorical data. • When the data shows two or more clusters of data.

Concept Checker 1

The following table shows a simple frequency distribution of the retirement age data. Calculate it's mean, median and mode.

Age	Frequency
54	3
55	1
56	1
57	2
58	2
60	2

Concept Checker 1

Solution:

The mean:

- The mean is calculated by adding together all the values (54+54+54+55+56+57+57+58+58+60+60 = 623) and dividing by the number of observations (11) which equals 56.6 years.

The median:

- Looking at the retirement age distribution (which has 11 observations), the median is the middle value, which is 57 years. When the distribution has an even number of observations, the median value is the mean of the two middle values. In the following distribution, the two middle values are 56 and 57, therefore the median equals 56.5 years.

The mode:

- The most commonly occurring value is 54, therefore the mode of this distribution is 54 years.

6 Measures of Dispersion

Dispersion in statistics is a way of describing how spread out a set of data is.

- The measures of central tendency are not adequate to describe data. Two data sets can have the same mean but they can be entirely different. Thus to describe data, one needs to know the extent of variability.
- This is given by the measures of dispersion. Range, interquartile range, and standard deviation are the three commonly used measures of dispersion.

7 Range

The range is the difference between the smallest value and the largest value in a dataset.

So far we've looked at calculating averages for sets of data, but quite often, the average only gives part of the picture

Averages give us a way of determining where the center of a set of data is, but they don't tell us how the data varies

Each player has the same average score, but there are clear differences between each data set.

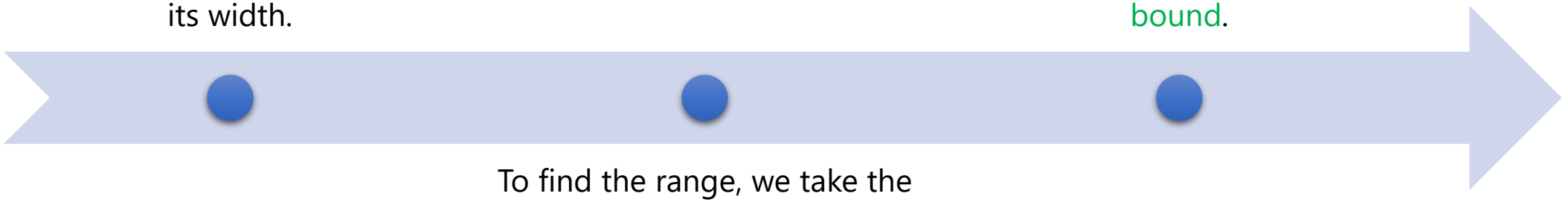
We need some other way of measuring these differences. This can be done by calculating the range.

7 Range

Calculating the Range

The range tells us over how many numbers the data extends, a bit like measuring its width.

The smallest value is called the lower bound, and the largest value is the upper bound.



To find the range, we take the largest number in the data set, and then subtract the smallest.

$$\text{Range} = \text{Upper Bound} - \text{Lower Bound}$$

Concept Checker 2

For the given volleyball score data, calculate the lower bound, upper bound and range for both Table and Table B.

Table A					
Score	8	9	10	11	12
Frequency	1	2	3	2	1

Table B					
Score	8	9	10	11	12
Frequency	1	0	8	0	1

Concept Checker 2

Solution:

For Table A:

- Lower Bound = 8
- Upper Bound = 12
- Range = 4

For Table B:

- Lower Bound = 8
- Upper Bound = 12
- Range = 4

The results appear same even though the data's different.

8 The problem with outliers

Both data sets above have the same range, but the values are distributed differently. We wonder if the range really gives us the full story about measuring spread?

The range only describes the width of the data, not how it's dispersed between the bounds.

Both sets of data above have the same range, but the second set has outliers—extreme high and low values. It looks like the range can measure how far the values are spread out, but it's difficult to get a real picture of how the data is distributed.

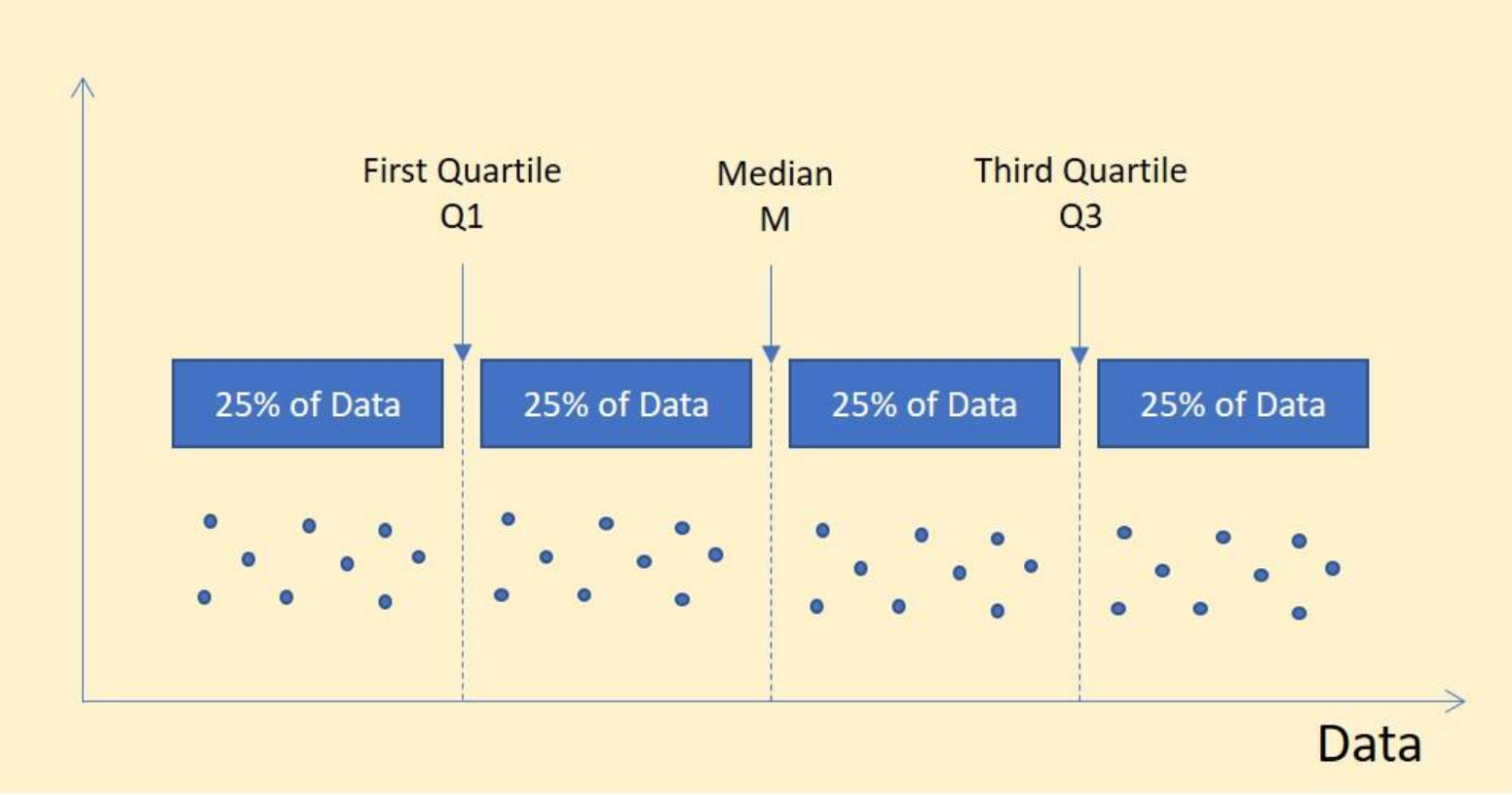
The range is a simple way of saying what the spread of a set of data is, but it's often not the best way of measuring how the data is distributed within that range.

9 Quartiles

Quartiles divide an ordered dataset into four equal parts, and refer to the values of the point between the quarters.

- Just as the median divides a set of data into two halves, the quartiles divide a set of data into four quarters. They are denoted by Q1, Q2 and Q3.
- The lower quartile (Q1) is the point between the lowest 25% of values and the highest 75% of values. It is also called the 25th percentile.
- The second quartile (Q2) is the middle of the data set. It is also called the 50th percentile, or the median.
- The upper quartile (Q3) is the point between the lowest 75% and highest 25% of values. It is also called the 75th percentile.

9 Quartiles


First Quartile
Q1
Median
M
Third Quartile
Q3
25% of Data
25% of Data
25% of Data
25% of Data
Data

9 Quartiles

Interquartile Range

- The interquartile range gives us a standard, repeatable way of measuring how values are dispersed.

Interquartile Range = Upper quartile – Lower quartile
Interquartile Range = Q3 – Q1

- The good thing about the interquartile range is that it's a lot less sensitive to outliers than the range is.
- The interquartile range includes the middle part of the data.

10 Variance & Standard Deviation

The variance and the standard deviation are measures of the spread of the data around the mean. They summarise how close each observed data value is to the mean value.

- In datasets with a small spread all values are very close to the mean, resulting in a small variance and standard deviation. Where a dataset is more dispersed, values are spread further away from the mean, leading to a larger variance and standard deviation.
- The smaller the variance and standard deviation, the more the mean value is indicative of the whole dataset. Therefore, if all values of a dataset are the same, the standard deviation and variance are zero.
- The standard deviation of a normal distribution enables us to calculate confidence intervals. In a normal distribution, about 68% of the values are within one standard deviation either side of the mean and about 95% of the scores are within two standard deviations of the mean.

- The standard deviation is the square root of the variance. The standard deviation for a population is represented by σ, and the standard deviation for a sample is represented by s.

10 Variance & Standard Deviation

Population Variance:

The population Variance σ^2 (pronounced sigma squared) of a discrete set of numbers is expressed by the following formula:

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

where:

- X_i represents the i^{th} unit, starting from the first observation to the last
- μ represents the population mean
- N represents the number of units in the population

Sample Variance:

The Variance of a sample s2 (pronounced s squared) is expressed by a slightly different formula:

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

where:

- x_i represents the i^{th} unit, starting from the first observation to the last
- $\bar{x}$ represents the sample mean
- n represents the number of units in the sample

Concept Checker 3

For Dataset A & B, calculate measures of central tendency & dispersion.

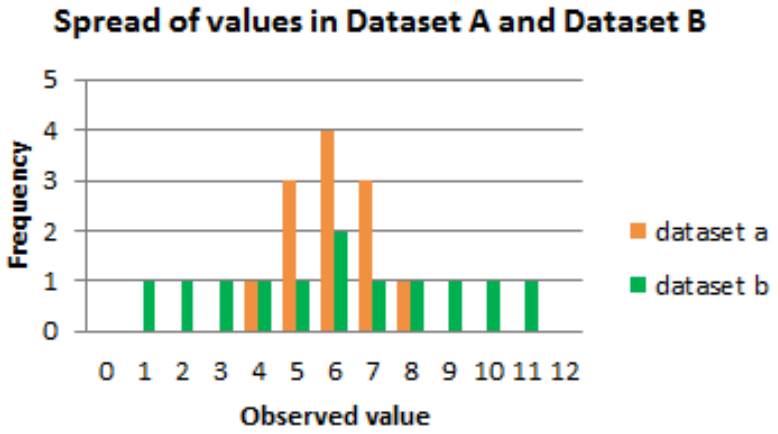
Dataset A											
4	5	5	5	6	6	6	6	7	7	7	8

Dataset B											
1	2	3	4	5	6	6	7	8	9	10	11

Concept Checker 3

Solution:

- The mode (most frequent value), median (middle value*) and mean (arithmetic average) of both datasets is 6. (*note, the median of an even numbered data set is calculated by taking the mean of the middle two observations).
- If we just looked at the measures of central tendency, we may assume that the datasets are the same.
- However, if we look at the spread of the values in the following graph, we can see that Dataset B is more dispersed than Dataset A. Used together, the measures of central tendency and measures of spread help us to better understand the data



Concept Checker 3

Solution: Dataset A

- The range is 4, the difference between the highest value (8) and the lowest value (4).
- As the quartile point falls between two values, the mean (average) of those values is the quartile value:
  - Q1 = (5+5) / 2 = 5
  - Q2 = (6+6) / 2 = 6
  - Q3 = (7+7) / 2 = 7
- The IQR for Dataset A is = 2
  IQR = Q3 - Q1 = 7 – 5 = 2

Dataset B

- The range is 10, the difference between the highest value (11) and the lowest value (1).
- As the quartile point falls between two values, the mean (average) of those values is the quartile value:
  - Q1 = (3+4) / 2 = 3.5
  - Q2 = (6+6) / 2 = 6
  - Q3 = (8+9) / 2 = 8.5
- The IQR for Dataset B is = 5
  IQR = Q3 - Q1 = 8.5 - 3.5 = 5

Concept Checker 3

Solution: Dataset A

- Calculate the population mean (μ) of Dataset A.
  (4+5+5+5+6+6+6+6+7+7+7+8) / 12
  mean (μ) = 6

- Calculate the deviation of the individual values from the mean by subtracting the mean from each value in the dataset
  = -2, -1, -1, -1, 0, 0, 0, 0, 1, 1, 1, 2
- Square each individual deviation value
  = 4, 1, 1, 1, 0, 0, 0, 0, 1,1,1, 4

- Calculate the mean of the squared deviation values = (4+1+1+1+0+0+0+0+1+1+1+4) / 12

- Variance σ^2 = 1.17
- Calculate the square root of the variance
- Standard deviation σ = 1.08

Dataset B

- Calculate the population mean (μ) of Dataset B.
  (1+2+3+4+5+6+6+7+8+9+10+11) / 12
  mean (μ) = 6

- Calculate the deviation of the individual values from the mean by subtracting the mean from each value in the dataset
  = -5, -4, -3, -2, -1, 0, 0, 1, 2, 3, 4, 5,
- Square each individual deviation value
  = 25, 16, 9, 4, 1, 0, 0, 1, 4, 9, 16, 25

- Calculate the mean of the squared deviation values=(25+16+9+4+1+0+0+1+4+9+16+25)/ 12

- Variance σ^2 = 9.17
- Calculate the square root of the variance
- Standard deviation σ = 3.03

The larger Variance and Standard Deviation in Dataset B further demonstrates that Dataset B is more dispersed than Dataset A.

11 Moments

Moments are set of statistical parameters used to describe a distribution.

- The mean and variance are special cases of a set of summary measures called the *moments* of a set of data.
- In general the k^{th} order moment about the value a is defined by:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \alpha)^k$$

Here α is any fixed number.

- So the mean is the first order moment about the origin, and the variance is essentially the second order moment about the mean with a divisor of n – 1 rather than n .

12 Skewness

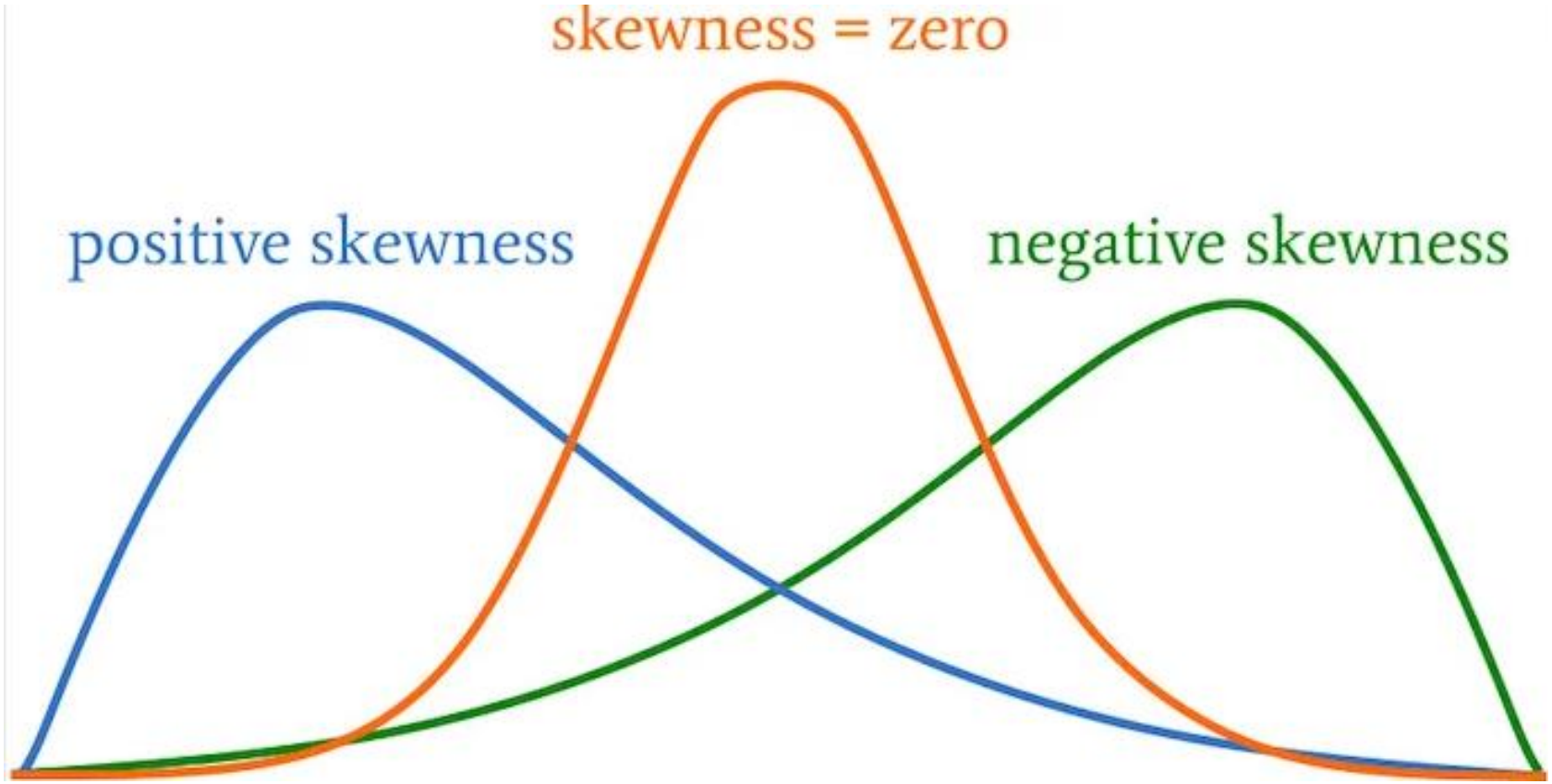
Skewness, also known as, third central moment tells us how symmetrical the data is around the mean.

A skewness of 0 indicates symmetrical data.

A positive skewness indicates that the peak of the distribution is to the left of the center.

A negative skewness indicates that the peak of the distribution is to the left of the center.

12 Skewness



12 Skewness

Measuring Skewness

- One particular measure of skewness is based on the third moment about the mean:

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^3$$

- The cubic power in this formula gives a positive or negative value depending on which side of the mean the value x_i is.

  positively skewed distributions with a long tail on the right give a positive value, and negatively skewed distribution with a long tail on the left give a negative value.

Concept Checker 4

The earlier table shows a simple frequency distribution of the retirement age data.
We'll see the effect of shape of distribution on measures of central tendency

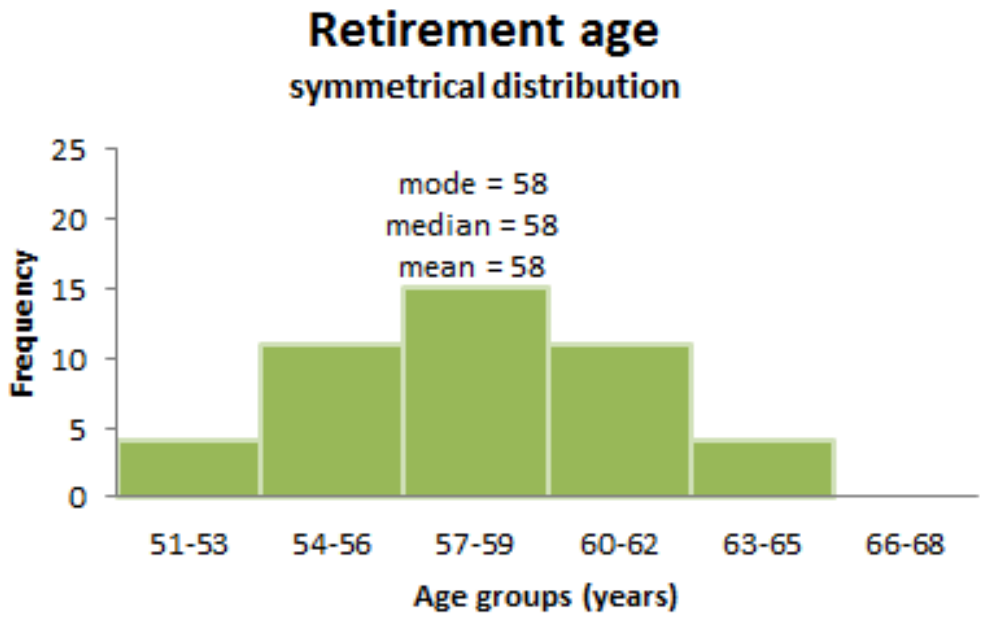
Age	Frequency
54	3
55	1
56	1
57	2
58	2
60	2

Concept Checker 4

Continued:

Symmetrical distributions:

When a distribution is symmetrical, the mode, median and mean are all in the middle of the distribution. The following graph shows a larger retirement age dataset with a distribution which is symmetrical. The mode, median and mean all equal 58 years.



Concept Checker 4

Continued:

Skewed distributions:

When a distribution is skewed the mode remains the most commonly occurring value, the median remains the middle value in the distribution, but the mean is generally ‘pulled’ in the direction of the tails. In a skewed distribution, the median is often a preferred measure of central tendency, as the mean is not usually in the middle of the distribution.

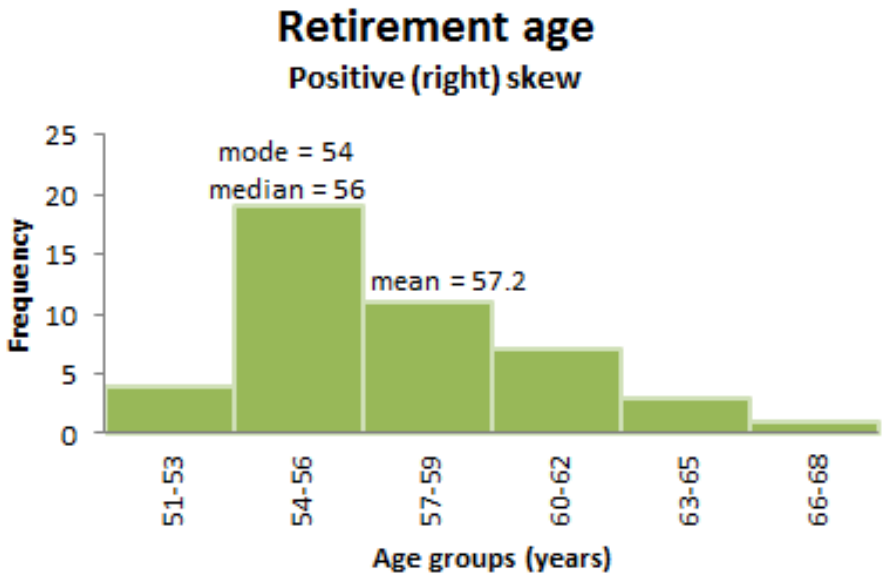
Concept Checker 4

Continued:

Positively skewed distribution:

A distribution is said to be positively or right skewed when the tail on the right side of the distribution is longer than the left side. In a positively skewed distribution it is common for the mean to be 'pulled' toward the right tail of the distribution. Although there are exceptions to this rule, generally, most of the values, including the median value, tend to be less than the mean value.

The following graph shows a larger retirement age data set with a distribution which is right skewed. The data has been grouped into classes, as the variable being measured (retirement age) is continuous. The mode is 54 years, the modal class is 54-56 years, the median is 56 years and the mean is 57.2 years.



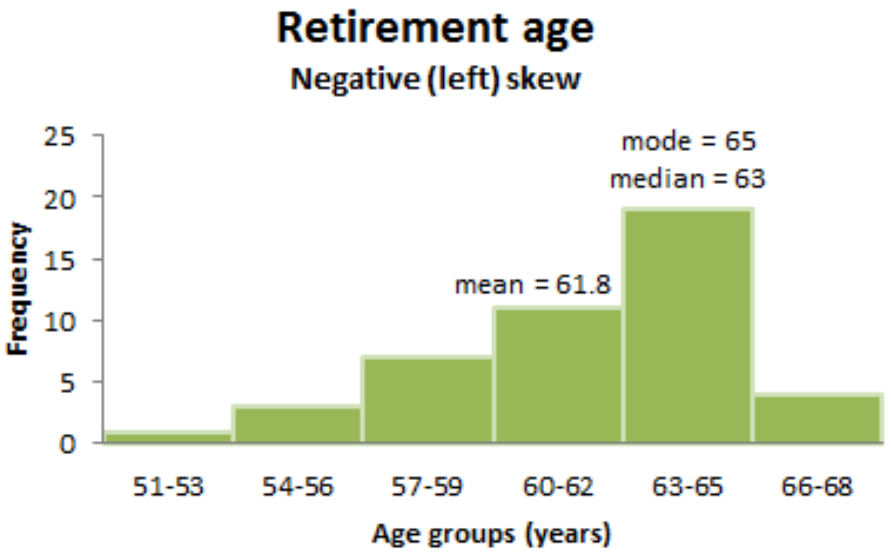
Concept Checker 4

Continued:

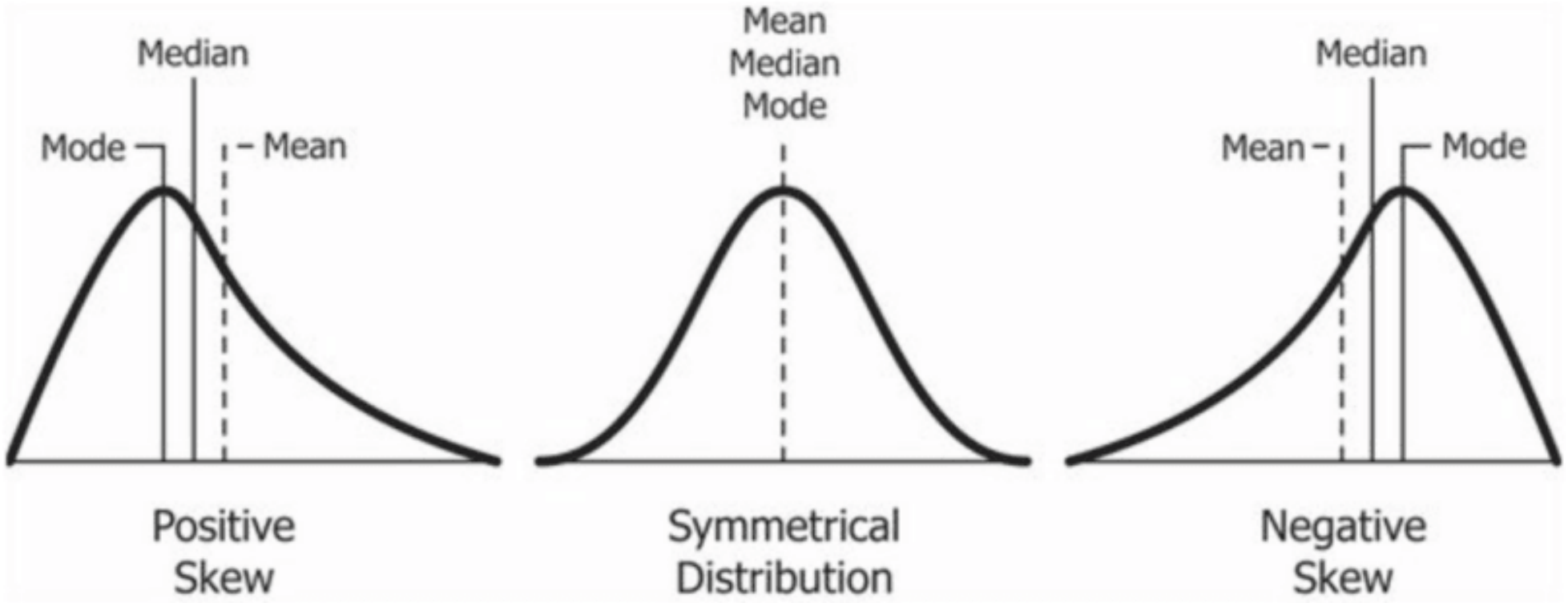
Negatively skewed distribution:

A distribution is said to be negatively or left skewed when the tail on the left side of the distribution is longer than the right side. In a negatively skewed distribution, it is common for the mean to be 'pulled' toward the left tail of the distribution. Although there are exceptions to this rule, generally, most of the values, including the median value, tend to be greater than the mean value.

The following graph shows a larger retirement age dataset with a distribution which left skewed. The mode is 65 years, the modal class is 63-65 years, the median is 63 years and the mean is 61.8 years.



13 Skewness & Central Tendencies



Summary

- The *location* of a data set can be summarized using the mean, the median or the mode.
- The *spread* of a data set can be summarized using the standard deviation, the range or the interquartile range.
- The variance measures the spread squared
- Third moments can be used to summarize the *skewness* (*ie* the degree of asymmetry) of a data set.