Quantitative Methods

Topics

- Sources of Data
- Summary Data
- Correlation and Bivariate Linear Regression
- Index Numbers
- Simple and Compound Interest
- Time value of Money
- IRR and NPV

Sources of Data

Sources of Data

Learning outcomes……..

- **Identify and Distinguish** between different sources and types of data
- **Distinguish** between population and sample
- **Explain** the key sampling methods
- **Distinguish** between continuous and discrete data
- **Define** categorical data and **Explain** how it can be converted into ordinal data
- **Interpret** a frequency and relative frequency distribution

Sources of Data

Learning outcomes…….. continued

- **Explain** the use of following in the presentation of Discrete and Continuous data
  - Pie chart
  - Bar chart
  - Histogram
  - Scatter plots
  - Line graph

Sources of Data

- We are all surrounded by huge volumes and types of data.
- An investment analyst should be aware of right sources of data that can be gathered for analysis
- An analyst should also be able to present the data conveniently.
- Let us discuss now from where the data can be gathered i.e. what are the sources of data.

Sources of Data

- **Primary Vs Secondary data**

- **Primary data** is generated or collected with a specific project or task in mind.

- For example, market research agencies conduct surveys among the consumers to understand their behavior or their likes and dislikes etc. This is a primary data that is generated by conducting consumer surveys. Or scientists / medial practitioners may conduct laboratory experiments to test some hypothesis. This is also an example of primary data.

- **Secondary data** is Economic, Financial and Social data collected by government bodies, international bodies, banks, companies which is presented in a convenient form.

Sources of Data

- **Official sources of Economic, Social and Financial information**
- Office for National Statistics (ONS): Wide variety of economic, social and financial data including labor market data is published.
- BOE's Quarterly Bulletin, US Federal Reserve Bulletin, IMF's International Financial Statistics; all provide international financial data and analysis.
- International data related to Banking is provided by Bank for International Settlements (BIS).
- World Bank and OECD also provide international data.

↳ Org. for Eco. Cooperation & Develop

↳ developed countries

Sources of Data

- **Commercial databases** → charge for data

- Investment analysts should be aware that huge amount of economic, financial, social, industry-specific, company-specific data is provided by several companies. These commercial databases are offered online and provide excellent course of secondary data.

- Best known in UK are Datastream Professional and Bloomberg Terminal. Analysts can obtain data such as GDP growth and inflation for several countries over last many years or P/E ratios or Operating Margins for all listed UK companies.

- Investment analysts typically use secondary data which is more appropriate and cheapest source of data. Sometimes, to support the analysis, even primary data can be collected by using questionnaires or conducting interviews.

Sources of Data

- **Cross Section and Time Series data**
- Cross section data are collected for a particular point or a period of time for many entities / markets etc. For example, if we would like to know the equity returns for various markets post-covid, we might collect the return data for say year 2021. This is data collected for year 2021 for stock markets of various countries.

- Time Series data, on the other hand, is collected for a particular variable over many intervals of time. For example, gas production in Qatar has changed over last 10 years. So, we may collect data for each of the last 10 years to analyse the trend. This is time series data.

- Both types of data are useful and extensively utilized.

Samples, Populations and Sampling Methods

- **Population and Sample**

- Population represents all members of a well-defined group.
  - Could be finite (countable): For example, it could be all FIFA World Cup 2022 attendees in Qatar
  - Could also be infinite (not countable): For example, grains of sand on a popular beach in Doha

- Sample: A subset of a population.
  - For example, it could be FIFA World Cup 2022 Quarterfinals attendees in Qatar

Samples, Populations and Sampling Methods

- **Population and Sample**

- The context of the research or problem will guide in deciding whether we are targeting population or sample.
  - For example, if an automobile analyst is studying demand for a particular model of Lamborghini, say Countach, in entire gulf region then, analysing trend in sales of this car in Qatar will represent a sample.
  - But if the analyst wants to project demand for this car in Qatar itself, then sales data for this car in Qatar represents a population.

Samples, Populations and Sampling Methods

- **Population and Sample**

- In economic or financial world, we quite often rely on samples.
  - Why do we do that? Why not study the entire population?
  - Because it is less costly and less time consuming to study samples
  - For example, if researcher wants to know whether prices of goods have increased over last 1 year in Qatar, a Consumer Price Index (CPI) can be looked into.
  - CPI consists of a basket of goods (sample) and not all good (population).

- Hence how the sample is selected is very important
  - The sample should have no bias i.e. it should be unbiased.
  - It will be unbiased, if
    - It is truly representative of population i.e. members of population are randomly selected, and sample should be sufficiently large

Samples, Populations and Sampling Methods

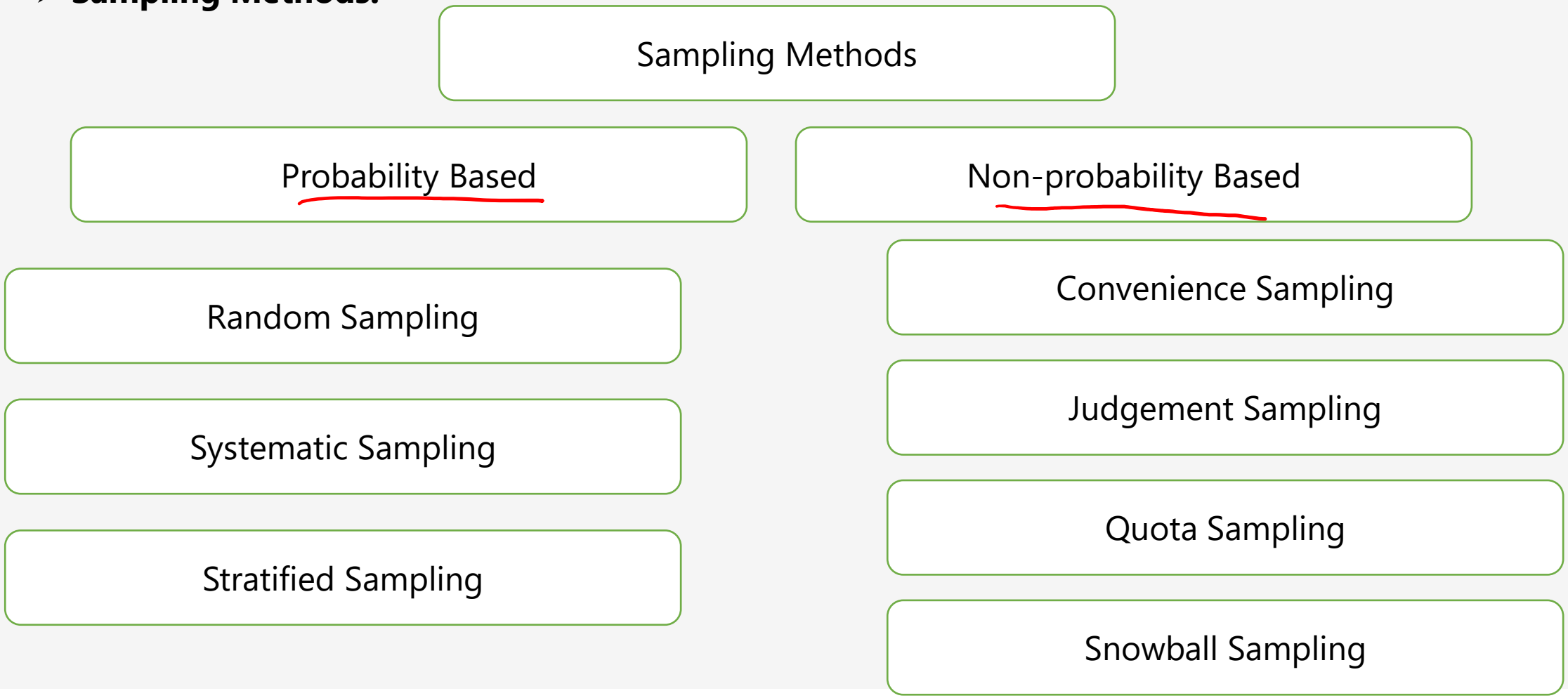
- **Random Vs. Non-random Sampling**

- In random sampling, each member of population has equal probability of getting selected. Hence it becomes unbiased sampling.

- In non-random sampling, some judgment is used in selecting a sample.

Samples, Populations and Sampling Methods

- **Sampling Methods:**



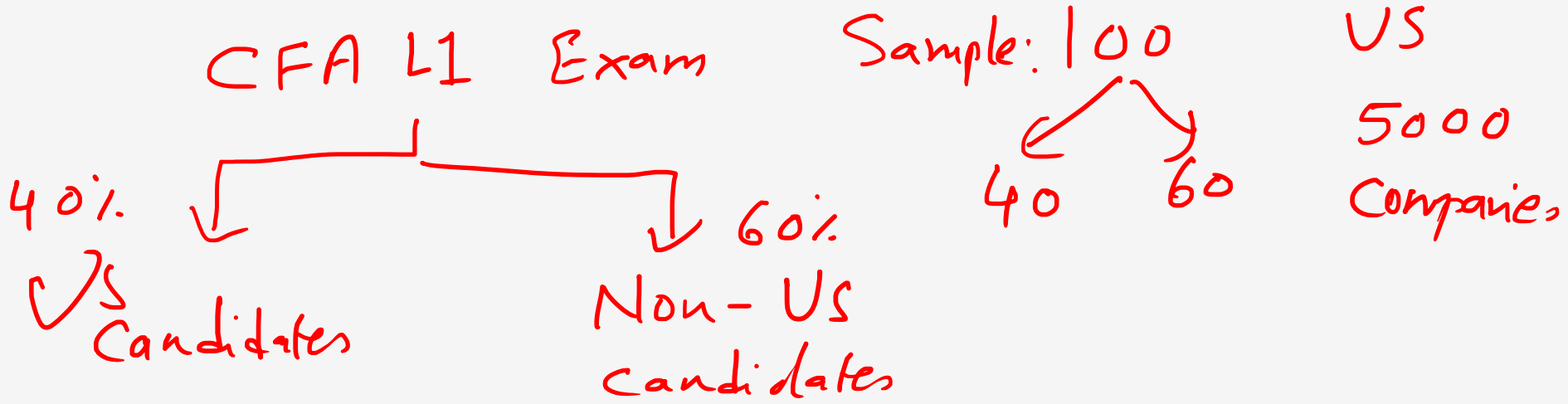
Samples, Populations and Sampling Methods

- **Probability based sampling**
- **Random sampling:**
  - As already stated, in random sampling, each member of population has equal probability of getting selected. Hence it becomes unbiased sampling.

- **Systematic Sampling:**
  - It involves selecting every nth value or member (say every 10^{th} or every 25^{th}) of the population.
  - It is as good as random sampling if the values in the data set do not follow a particular order.
  - Could be considered as a simpler version of random sampling.

Samples, Populations and Sampling Methods

- **Probability based sampling**
- **Stratified Sampling:**
  - In this case, characteristics of population are identified (for example, proportion of non-US students writing CFA L1 exam), and then we select random sample to represent those characteristics in the correct proportion.
  - It is considered better than simple random sampling it reduces sampling error or possibility of selecting unrepresentative sample.



Samples, Populations and Sampling Methods

- **Non-probability based sampling**
- **Convenience sampling:**
  - Sample is selected based on convenience of researcher, not incurring cost or time required to create a random sample.

- **Judgement Sampling:** → apply some judgement
  - It involves selecting every nth value or member (say every 10^{th} or every 25^{th}) of the population.
  - It is as good as random sampling if the values in the data set do not follow a particular order.
  - Could be considered as a simpler version of random sampling.

Samples, Populations and Sampling Methods

- **Non-probability based sampling**

- **Quota Sampling:**
  - In this case, characteristics of population are identified (for example, proportion of non-US students writing CFA L1 exam), and then we select random sample to represent those characteristics in the correct proportion.
  - It is considered better than simple random sampling it reduces sampling error or possibility of selecting unrepresentative sample.

- **Snowball Sampling:**
  - This is used if the desired sample characteristics is very rare.
  - Hence it may be very costly to locate respondents with rare characteristics.
  - Additional sample information is generated by relying on referrals from initial respondents.

Samples, Populations and Sampling Methods

- **Data Types**
- **Continuous data**
  - This data can take any value from negative infinity to positive infinity.
  - For example, rate of return on investment.
  - In simple words, the values are not countable.
- **Discrete data:**
  - This data can take only finite number of values.
  - For example, number of companies that pay dividends.
- **Categorical data:**
  - This data is presented in categories. Generally, it may not be possible to apply descriptive statistics to categorical data.
  - If we rank the categorical data, we get **ordinal** data. For ordinal data, some descriptive statistics such as median may be used.

Samples, Populations and Sampling Methods

- **Data Presentation**

- **Frequency distribution**
  - When raw data is grouped into intervals, we get a frequency distribution.
  - Each interval has an associated frequency which means the number of observations or data points in that interval.
  - The width of the interval and number of intervals depends upon the purpose of data presentation and the volume of data.

Samples, Populations and Sampling Methods

- **Data Presentation**
- **Frequency distribution**
  - When raw data is grouped into intervals, we get a frequency distribution.
  - Each interval has an associated frequency which means the number of observations or data points in that interval.
  - The width of the interval and number of intervals depends upon the purpose of data presentation and the volume of data.

THE FREQUENCY DISTRIBUTION OF STORES ACCORDING TO THEIR WEEKLY SALES FIGURES

Weekly sales (£ in thousands)	Number of stores
30 or less	7
31 to 40	9
41 to 50	20
51 to 60	37
61 to 70	50
71 to 80	38
81 to 90	20
91 to 100	15
101 or more	4
Total	200

Samples, Populations and Sampling Methods

- **Data Presentation**
- **Relative frequency distribution**
  - This is derived from frequency distribution, in which values in each interval are expressed as percentage of total values
- **Cumulative frequency distribution:**
  - In this case, we keep adding relative frequencies of intervals till last interval is reached.
  - The cumulative relative frequency for last interval will be 100%.

RELATIVE FREQUENCY DISTRIBUTION FOR THE DATA IN TABLE 7.1

Weekly sales (£ in thousands)	Number of stores	Percentage of stores (%)	Cumulative percentage (%)
30 or less	7	3.5	3.5
31 to 40	9	4.5	8.0
41 to 50	20	10.0	18.0
51 to 60	37	18.5	36.5
61 to 70	50	25.0	61.5
71 to 80	38	19.0	80.5
81 to 90	20	10.0	90.5
91 to 100	15	7.5	98.0
101 or more	4	2.0	100.0
Total	200	100.0	100.0

Samples, Populations and Sampling Methods

Candidate scores in the exam

Groups Scores	No. of Student
Less than 35%.	10
36 to 55%.	25
56 to 75%.	30
76 to ~~100%.~~ 95%.	10
95% & above	5
	80

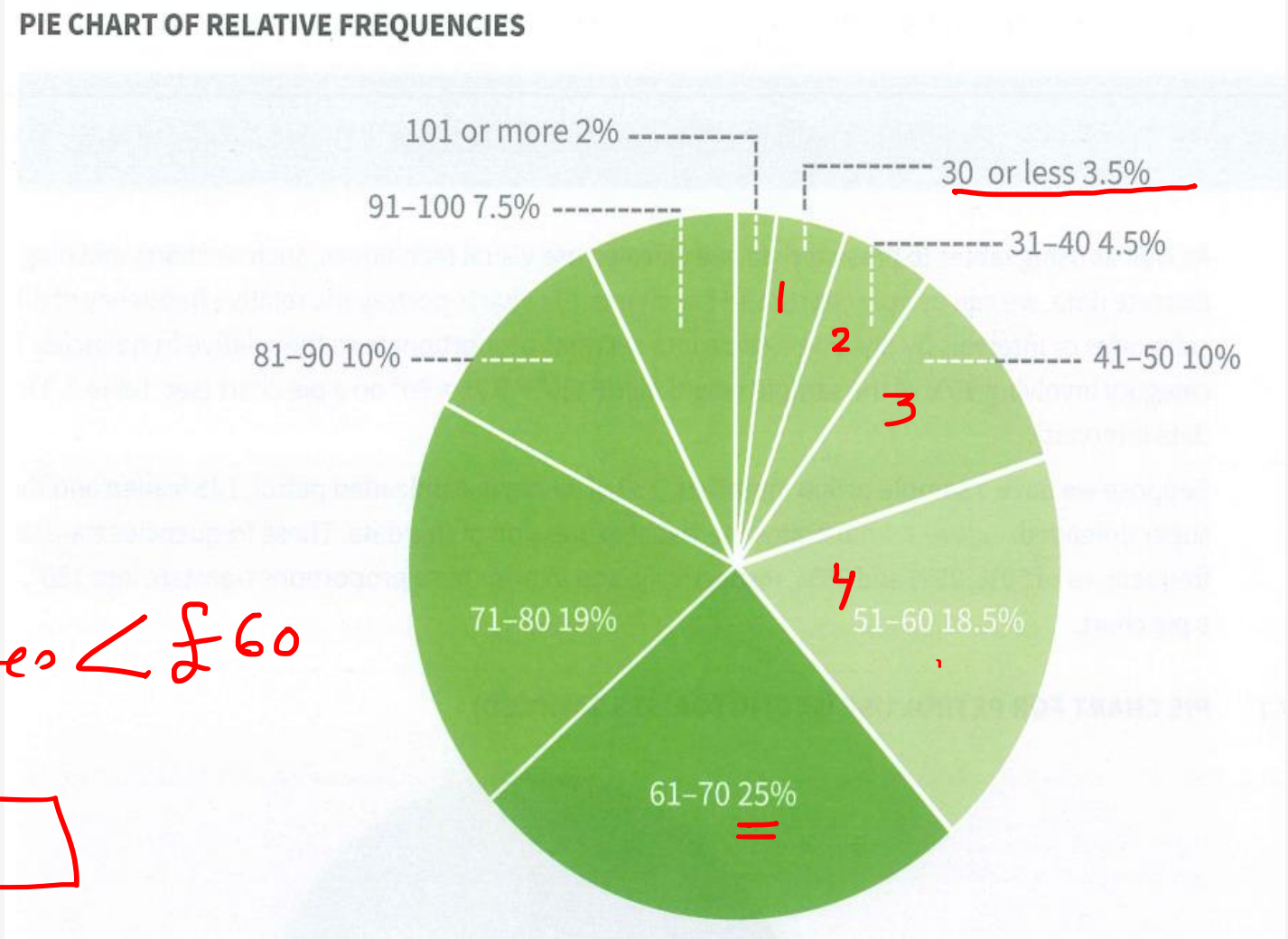
i) What is the relative frequency of interval 36 to 55%.
OR 25/80 = 31.25%

What % of student have scored between 36 to 55%.

ii) What is cum. freq. of interval 36 to 55%.
OR how many % of student scored below 55%. ?
35/80

Samples, Populations and Sampling Methods

- **Visual Presentation for Discrete Data**
- **Pie Chart**
  - Presenting information for different categories by dividing the circle into sections **proportionate to relative frequencies**.



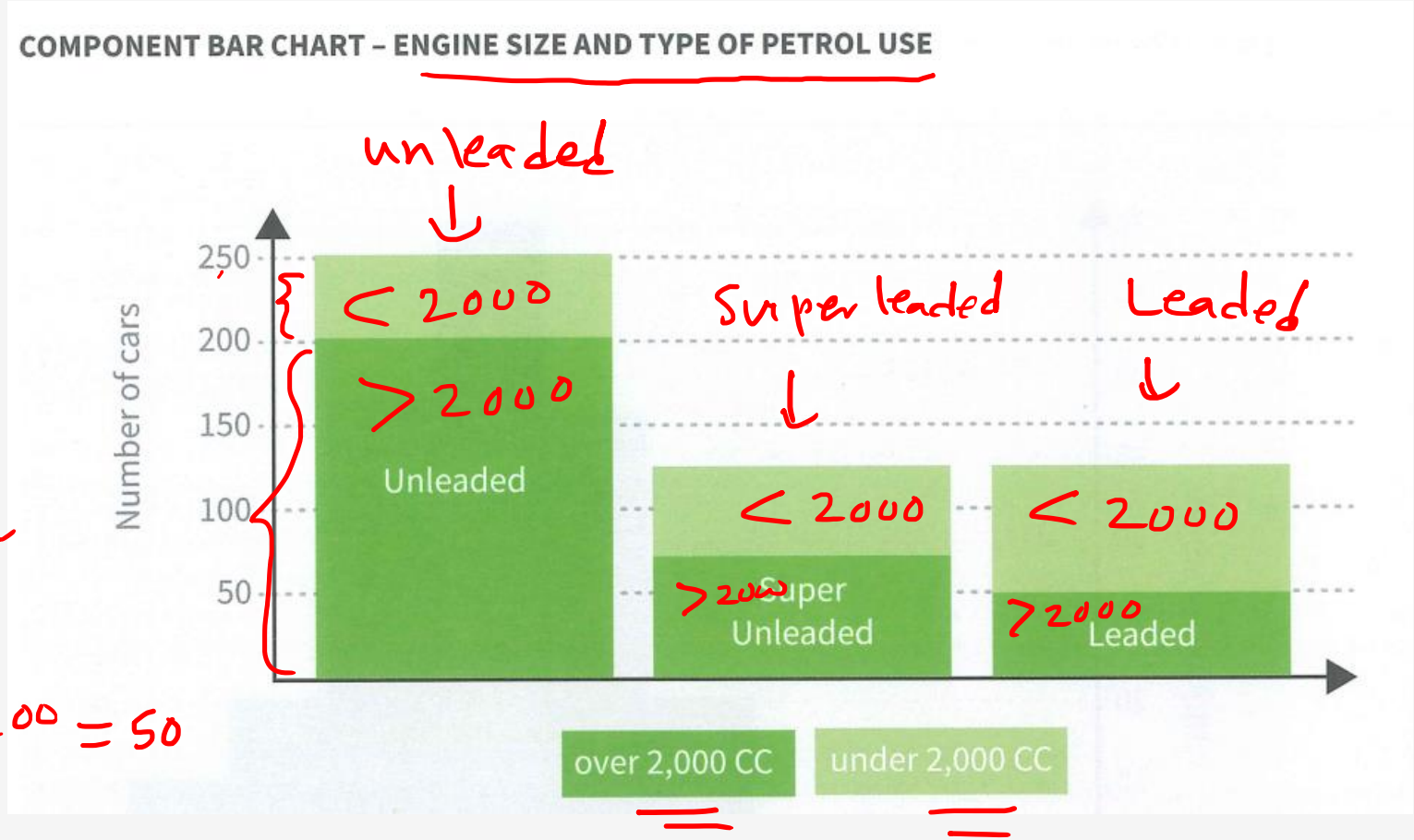
what % of stores have sales < £60

1 + 2 + 3 + 4

3.5 + 4.5 + 10 + 18.5 = ☐

Samples, Populations and Sampling Methods

- **Visual Presentation for Discrete Data**
- **Bar Chart**
  - Presenting information for different categories through bars.



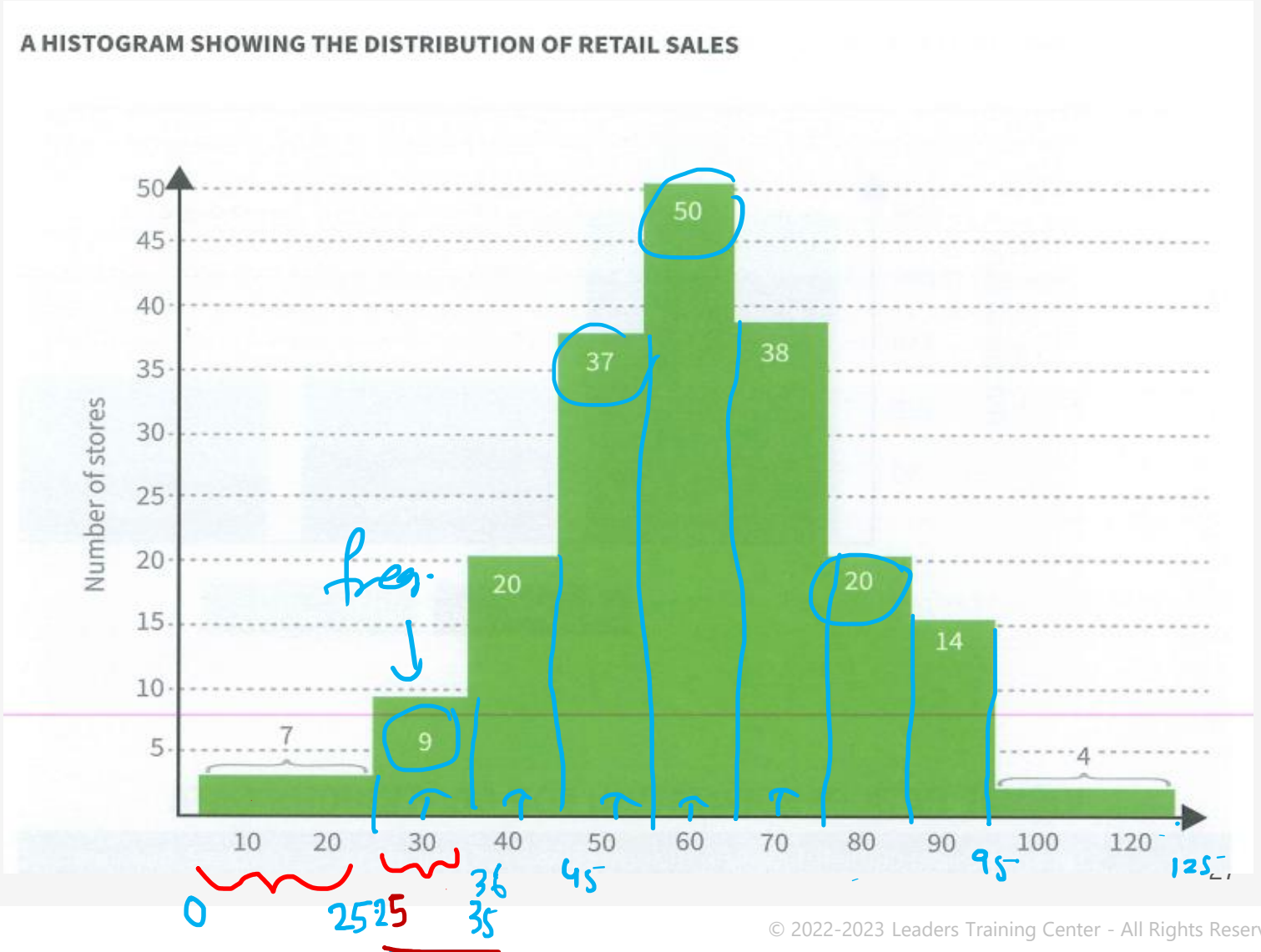
How many vehicles under 2000 cc are using unleaded ? → 250-200 = 50

Samples, Populations and Sampling Methods

- **Visual Presentation for Continuous Data**
  - → infinite
- **Histogram**
  - Hight of each bar indicates the frequency in that interval

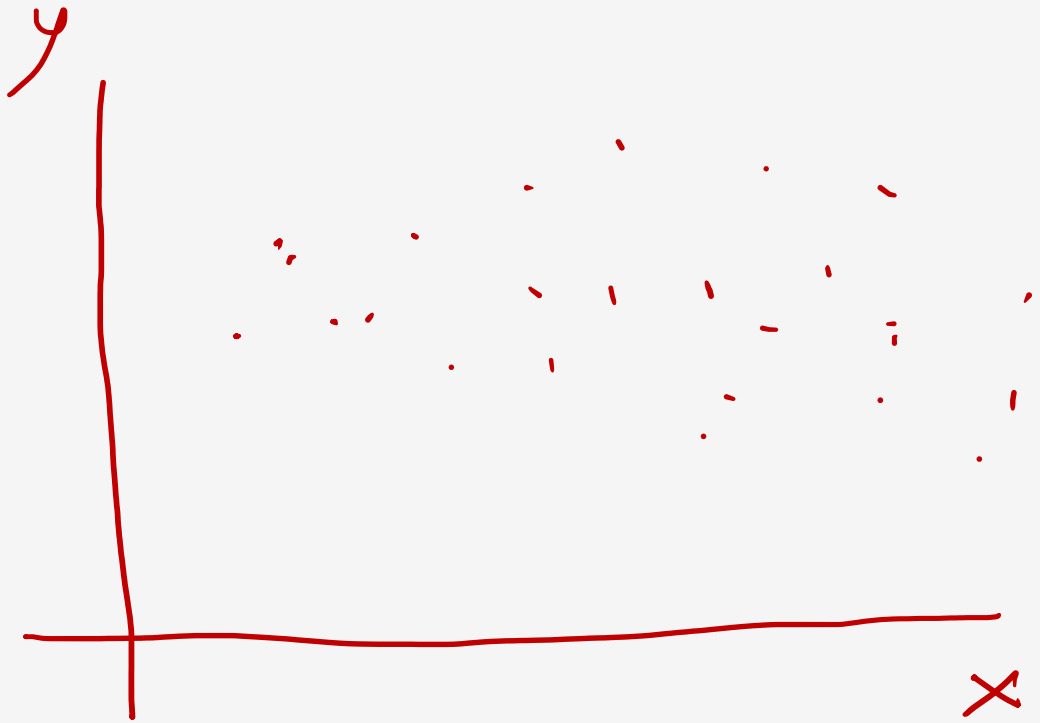
Freq. bet 36 to 45 OR fre. for bar with mid pt of 40

25.00 25.01



Samples, Populations and Sampling Methods

- **Graphs and Scatter diagrams (Self Read)**
- **Pages:** 13 to 15.



Summary Data

Summary Data

Learning Outcomes……..

- **Define, Explain and Calculate** arithmetic mean, geometric mean, median and mode using raw and interval data and **Calculate** geometric mean using a series of returns
- **Explain** the relationship between mean, median and mode for symmetric and skewed data
- **Define, Explain and Calculate** following measures of dispersion for both raw and interval data: standard deviation (population and sample), variance, range, quartiles and percentiles and interquartile range

Summary Data

Learning Outcomes……continued

- **Explain** the notion of probability distributions and **Identify** the properties of normal distribution

- **Explain and Apply** the concepts of null and alternative hypotheses and the role of statistical significance in rejecting / accepting null / alternative hypotheses in the context of investment decision making

Summary Data

Measures of Central Tendency

- Arithmetic Mean
- Median
- Mode
- Geometric Mean

give us an idea where the centre of data is

Summary Data

Measures of Central Tendency : Arithmetic Mean

- Most commonly used, known as average
- Calculated as: Sum of all observations divided by number of all observations
- Sample mean is denoted as $\bar{x}$
- For a frequency distribution, arithmetic mean can be calculated as follows:

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

what is A.M. for following data

10, 20, 30, 40 $\quad \frac{10+20+30+40}{4}$

$\rightarrow$ Numerator $\sum$ (Value x fre) $\quad 380$

$n \rightarrow 300$

Summary Data

Measures of Central Tendency : Arithmetic Mean

- Example: Calculate mean of printing errors for the following frequency distribution.

Number of errors per page (x)	Number of pages (frequency) (f)
0	100
1	80
2	70
3	40
4	10

300

$$= ((0 \times 100) + (1 \times 80) + (2 \times 70) + (3 \times 40) + (4 \times 10)) \div 300$$

$$= (0 + 80 + 140 + 120 + 40) \div 300$$

= 1.27 errors per page

$$\frac{(0\times100) + (1\times80) + (2\times70) + (3\times40) + (4\times10)}{(100+80+70+40+10)}$$

Summary Data

Measures of Central Tendency : Arithmetic Mean

- Example: Calculate mean of printing errors for the following frequency distribution. Data is given in less precise form.

THE DISTRIBUTION OF SHARE PRICES BY INTERVAL

Share price (pence)	Number of shares (f_i)	Midpoint (m_i)	$f_i m_i$
Less than 20p	80	10	800
20p or more, but less than 40p	120	30	3,600
40p or more, but less than 60p	390	50	19,500
60p or more, but less than 80p	210	70	14,700
80p or more, but less than 100p	200	90	18,000
Total	1,000		56,600

$$\bar{x} = \sum_{i=1}^{5} f_i m_i \div n = \frac{(80 \times 10) + (120 \times 30) + (390 \times 50) + (210 \times 70) + (200 \times 90)}{1{,}000} = \frac{56{,}600}{1{,}000}$$

$$= 56.6p$$

Disadv. of mean: Affected by outliers or extreme values

Adv. of Median → not affected by extreme value

Measures of Central Tendency : Median

- Arrange the data in ascending or descending order
- If the data has **odd** number of items: Median is exactly middle value
- If the data has **even** number of items: Median is the arithmetic average of middle two values

Marks of students: 80, 85, 78, 65, 42,

Arranging marks 42, 65, 78, 80, 85

$\frac{75+72}{2}$

42, 65, 72, 78, 80, 85

Summary Data

Measures of Central Tendency : Median

- Example: Calculating Median when frequency distribution is given

CUMULATED FREQUENCIES OF SHARE PRICES

Share price (pence)	Number of shares (f_i)	Cumulative number
Less than 20p	80	80
20p or more, but less than 40p	120	200
40p or more, but less than 60p	390	590
60p or more, but less than 80p	210	800
80p or more, but less than 100p	200	1,000
TOTAL	1,000	

- Formula for 'n'th observation:

$$\text{Median} = L + \frac{n - F}{f} W$$

Summary Data

Measures of Central Tendency : Median

- Example: Calculating Median when frequency distribution is given
- Formula for 'n'th observation:

$$\text{Median} = L + \frac{n - F}{f} W$$

The 500th item will therefore be:

40p + (300 ÷ 390 × 20p) = 40p + 15.38p = 55.38p

A similar calculation that would reveal that the 501st item is:

40p + (301 ÷ 390 × 20p) = 40p + 15.44p = 55.44p

An average of 55.38p and 55.44p gives the median as 55.41p.

Summary Data

$\frac{600}{2}$ 300th & 301st position

Measures of Central Tendency : Median

	Fre.	Cum fre.
Less than 10	45	45
10 – 20	120	165
20 – 30	310	475
30 – 40	95	570
40 – 50	30	600

Calculate Median

300th position

$L + \frac{(\frac{n}{2} - F)}{f} \times w$

$20 + \frac{(300 - 165)}{310} \times 10$

$= 24.35$

301st

$20 + \frac{(301 - 165)}{310} \times 10$

$=$

Summary Data

$\frac{600}{2}$ 300th & 301st position

Measures of Central Tendency : Median

	Fre.	Cum fre.
Less than 10	45	45
10 – 20	120	165
20 – 30	310	475
30 – 40	95	570
40 – 50	29 ~~30~~	~~600~~ 599

Calculate Median

300th position

$L + \frac{(n - F)}{f} \times w$

$20 + \frac{(300 - 165)}{310} \times 10$

$= 24.35$

301st

$20 + \frac{(301 - 165)}{310} \times 10$

$=$

$\frac{599+1}{2} = 300^{th}$

Summary Data

Measures of Central Tendency : Mode

- It is the most frequently occurring value in the distribution

- In frequency distribution, the model interval is one with highest frequency

Summary Data

Measures of Central Tendency : Geometric Mean and Geometric Mean Return

- Geo mean is 'n'th root of product of 'n' numbers.
- Given the following data:

AM > GM

6, 2, 3, 7, 1 (cm)

The arithmetic mean is:

$$\frac{6+2+3+7+1}{5} = 3.8 \text{ (cm)}$$

The geometric mean is:

$$(6 \times 2 \times 3 \times 7 \times 1)^{1/5} = (252)^{1/5} = 3.02 \text{ (cm)}$$

Summary Data

Measures of Central Tendency : Geometric Mean and Geometric Mean Return

- A M is always greater than G M
- If the data is more volatile, i.e. values are widely dispersed, the difference between A M and GM will be greater.
- G M is more relevant when the data is given in percentage returns.
- For example, what will be the geometric mean for the following stock market return data?
  - 2016: 7%; 2017: -4%; 2018: 12%; 2019: 16%; 2020: -14%, 2021: 6%

over a period of time : GM

$7\% \times -4\% \times 12\% \times 16\% \times -14\% \times 6\% \times$

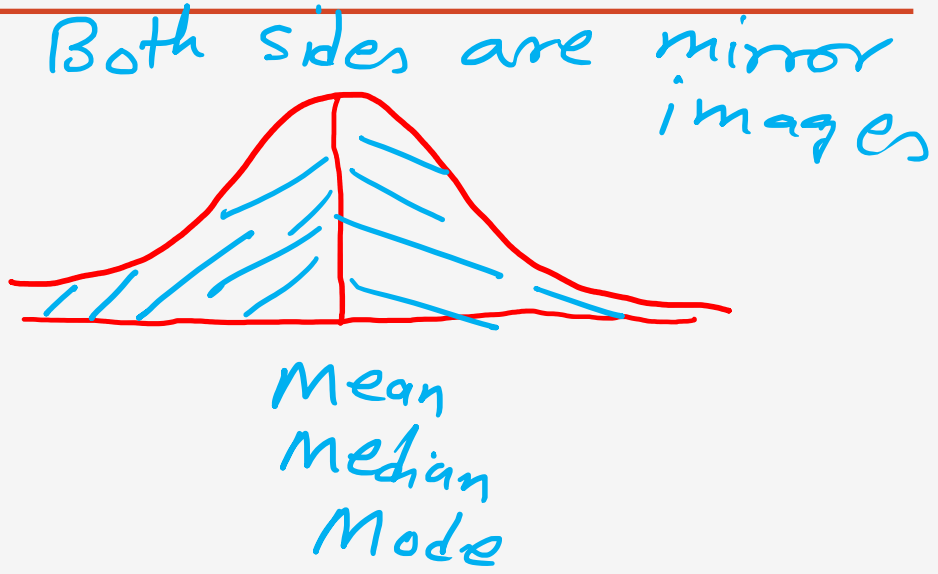
$$\left[(1+7\%)\times(1-4\%)\times(1+12\%)(1*16\%)(1-14\%)(1+6\%)\right]^{1/6}-1$$

$$=3.32\%$$

Summary Data

Relationship between Mean, Median and Mode

- For symmetric data: Mean = Median = Mode
- For positively skewed data: Mean > Median > Mode
- For negatively skewed data: Mean < Median < Mode



Right side

Left side

+ve Skew