Queen Mary
University of London

Maths & Stats Pre-Sessional

Sampling, Measures of Central Tendency, and Measures of Variability and Skewness

Lecturer: Claudio Vallar
School of Economics and Finance

Teaching Team

<u>Lectures</u>

Lecturer: Claudio Vallar

Email: c.vallar@qmul.ac.uk

Office: GC317 (third floor - Graduate Centre)

<u>Tutorials</u>

Lecturer: Ivi Theodoulou

Email: i.theodoulou@qmul.ac.uk

Office: GC317 (third floor - Graduate Centre)

Aims of Maths & Stats Pre-sessional Course

This course provides you with the opportunity to:

- Review key mathematical and statistical concepts and tools.
- Show examples of how these tools are used in Economics and Finance.
- Ensure a solid foundation for your study in the MSc programme.

Topics Overview

The Key Topics we will cover are:

- **Topic 1:** Sampling, central tendency, and other moments
- **Topic 2:** Probability distributions, covariances and correlations
- **Topic 3:** Estimation and hypothesis testing
- **Topic 4:** Simple functions and basics of present value
- **Topic 5:** Basics of derivatives
- **Topic 6:** Basics of regression analysis and matrix algebra (This topic is for students enrolled in MSc Banking and Finance)

End of Pre-sessional Test

- The end-of-presessional test is ONLY for the students enrolled in **MSc Banking and Finance (conversion)**

- **Location:** online.

Midterm test will be available on QMPlus in the page "SEF PGT Welcome Week and Pre-Sessional - January 2025" (here the link).

End of Pre-sessional Test

- **Date and time:** The test will be available on QMPlus from Thursday 30th January at 11am to Friday 31st January 2025 at 1pm.

- **Duration:** 1 hour. You can start whenever you want during the available time window. However, once you start, you have one hour. When time expires , all open attempts will be automatically submitted.

End of Pre-sessional Test

- **Structure:** the test is an online quiz on QMPlus.

  There are 15 questions. You have to answer ALL of them

- Multiple Choice Questions

- **Pass Mark:** you have to answer correctly at least to 10 questions

- **Examinable Material:** all the topics covered during lectures and tutorials (topics 1-2-3-4).

End of Pre-sessional Test

- Mock exam will be available on QMPlus on Friday 14th January 2025.
- A Q&A sessions will be held during on Friday 17th January 2025.
- Additional office hours will be done on Monday 27th January and Wednesday 29th January. More details will be provided.

End of Pre-sessional Test - RESIT

- For those of you who failed the test in January or could not do the first sit, they can **resit the test** in the week starting 24th February.
- Exact date and time will be provided in the due course

Sampling, Measures of Central Tendency, and Measures of Variability and Skewness

In this session:

- We will review some basic statistical concepts
- Statistics is a tool for collecting, analysing, presenting, and interpreting data
- Many decisions are based on incomplete information
- Statistics can be used to enable a more informed decision

For more extensive reading, refer to Chapter 1 and 2 of Newbold, P., Carlson, W., and Thorne, B. (2010). Statistics for Business and Economics, Pearson, 7th Edition

Statistics

- What data is needed?
- How should the data be collected?
- How should the data be presented?
- How should the data be analysed?
- What inference can be made from the data?

Statistics is the science concerned with development and studying methods for collecting, analysing, interpreting and presenting empirical data

Descriptive vs Inferential Statistics

Descriptive Statistics

Use for summarising and presenting the data.

Describe a sample

Inferential Statistics

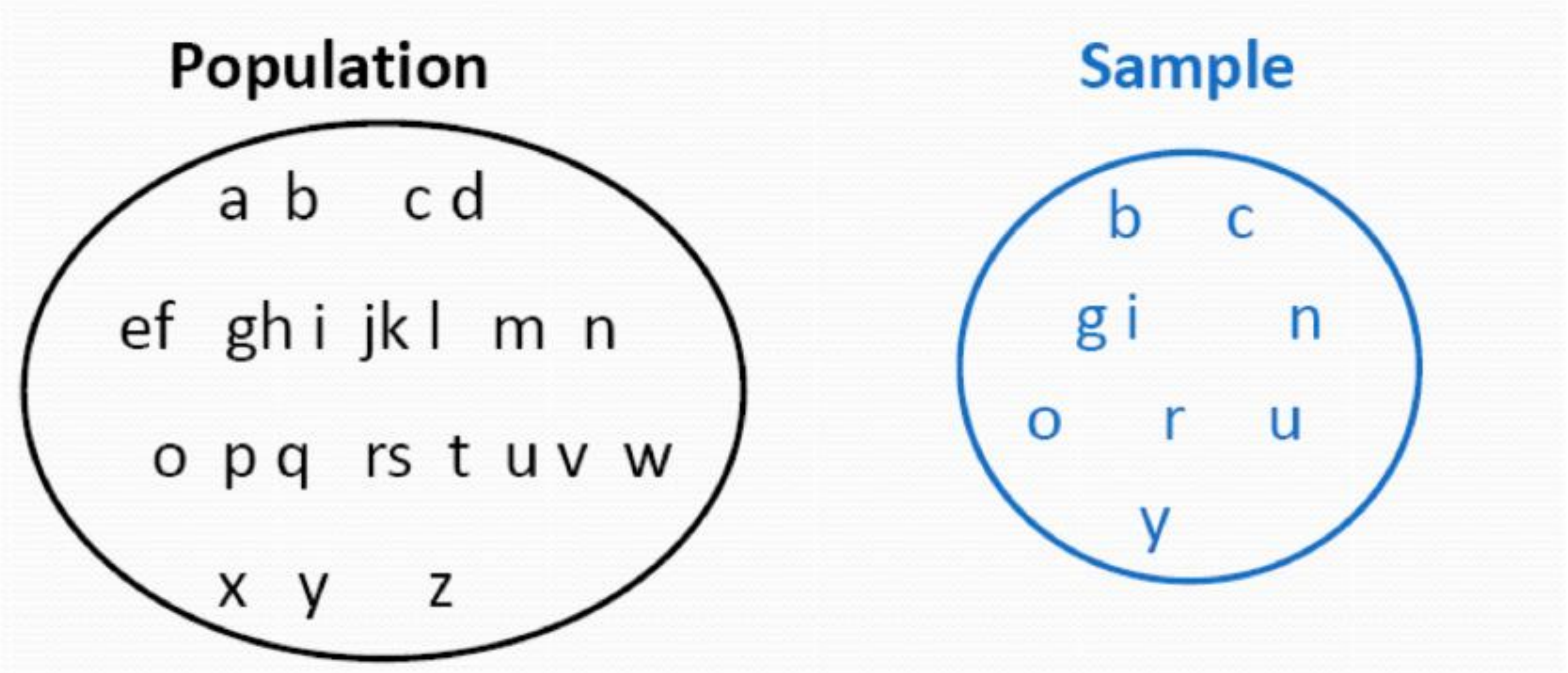
Make inference about the population using a sample.

The use of a data sample to make predictions, forecasts and estimates about the population.

Key Definition

- **Population:** a collection of all elements of interest (N = population size)
- **Sample:** an observed subset of the population (n = sample size)
- **Parameter:** characteristics of a population
- **Statistic:** characteristics of a sample

Population vs Sample



Values calculated using population data are called **parameters**.	Values calculated from sample data are called **statistics**.

Random Sampling

Random sampling requires that:

- Each member of the population has the same probability of being selected
- Each member of the population is chosen strictly by chance

Two types of random sampling used are:

- Simple random sampling
- Stratified random sampling

Type of Data

- Primary versus secondary
- Numeric versus non-numeric
- Discrete versus continuous
- Categorical versus ordinal

Descriptive Statistics

- Frequency Distribution and charts
- Measures of central tendency
    - Arithmetic mean
    - Median
    - Mode
- Measures of variability
    - Range
    - Variance
    - Standard Deviation

Frequency Distributions: a definition

Definition:

- Frequency distributions are visual displays that organise and present the number of observations within a given interval so that the information can be interpreted more easily.

- Frequency distributions can show absolute frequencies or relative frequencies, such as proportions or percentages.

How to show a frequency distribution:

A frequency distribution of data can be shown in a table or graph. Some common methods of showing frequency distributions include frequency tables, histograms or bar charts.

Frequency Distributions

- **Absolute frequency**

Summarize data by dividing it into classes or intervals and showing the number of observations in each class

- **Relative frequency**

The proportion of observations belonging to a class

$$relative\ frequency\ of\ a\ class = \frac{frequency\ of\ the\ class}{total\ number\ of\ observations}$$

- **Cumulative frequency**

For each class, the total number of observations in all classes up to and including that class

Frequency Distributions

How to find the frequency distribution of a discrete variable?

1. Determine the number of intervals (*k*)
2. Intervals should be the same width (*w*)

$$w = \frac{Largest\ data\ value\ - smallest\ data\ value}{Number\ of\ intervals}$$

3. Intervals must be inclusive and non-overlapping: each observation must belong to one and only one interval

Frequency Distribution: an example

Example: Student grades (n=120)

70	73	30	16	69	84	76	85	40	50
65	89	83	70	78	35	36	65	75	65
60	73	70	30	86	48	80	60	71	87
81	45	82	50	76	50	73	88	50	84
80	70	75	30	59	88	65	60	50	74
50	55	59	40	55	35	70	66	77	50
73	76	70	59	60	35	65	60	87	35
40	60	55	50	60	77	57	55	73	55
50	56	75	48	45	49	40	70	63	72
70	16	71	66	40	55	33	35	31	81
55	43	60	73	89	69	50	50	85	35
69	68	80	70	88	42	35	70	65	95

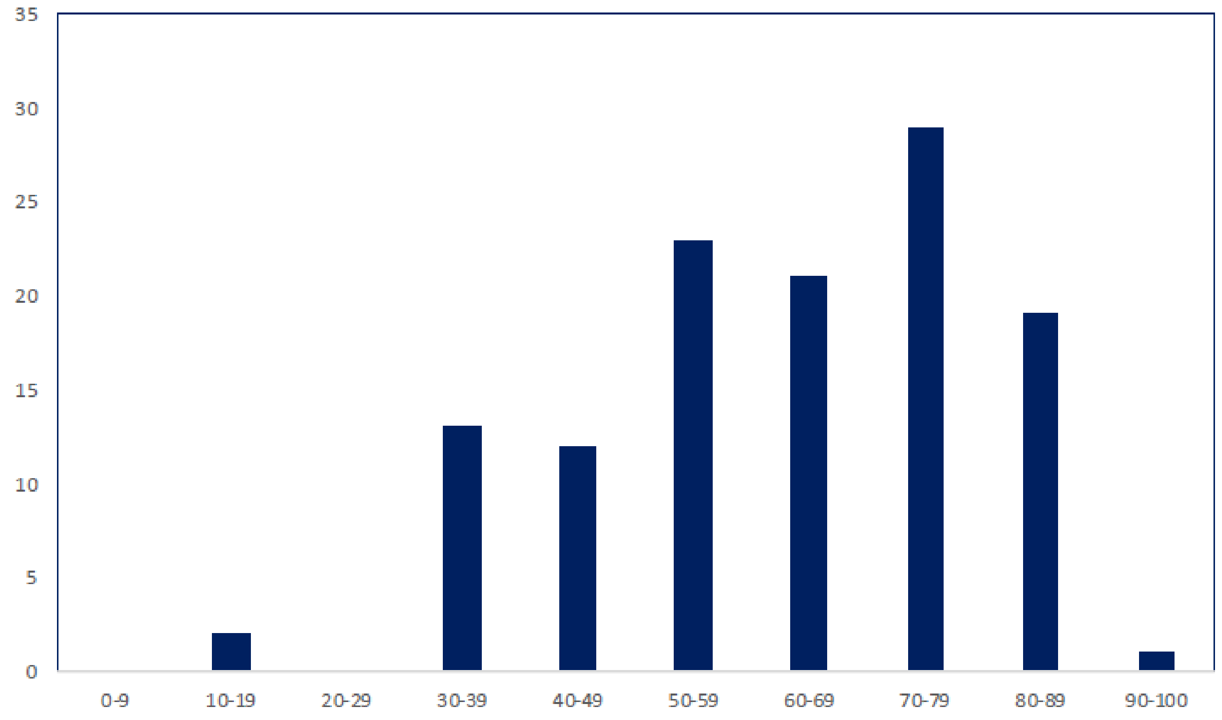
Frequency Distribution: an example

The frequency distribution in a table

Interval	Absolute Frequency	Relative Frequency	Cumulative Absolute Frequency	Cumulative Relative Frequency
0-9	0	0.0%	0	0.0%
10-19	2	1.7%	2	1.7%
20-29	0	0.0%	2	1.7%
30-39	13	10.8%	15	12.5%
40-49	12	10.0%	27	22.5%
50-59	23	19.2%	50	41.7%
60-69	21	17.5%	71	59.2%
70-79	29	24.2%	100	83.3%
80-89	19	15.8%	119	99.2%
90-100	1	0.8%	120	100.0%

Frequency Distribution: an example

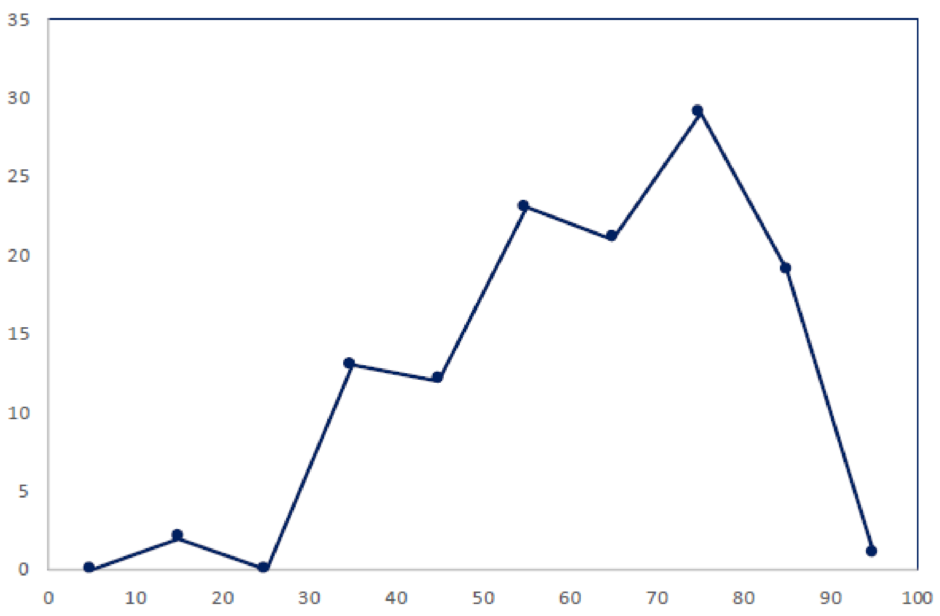
The frequency distribution visualized in a chart (a histogram)

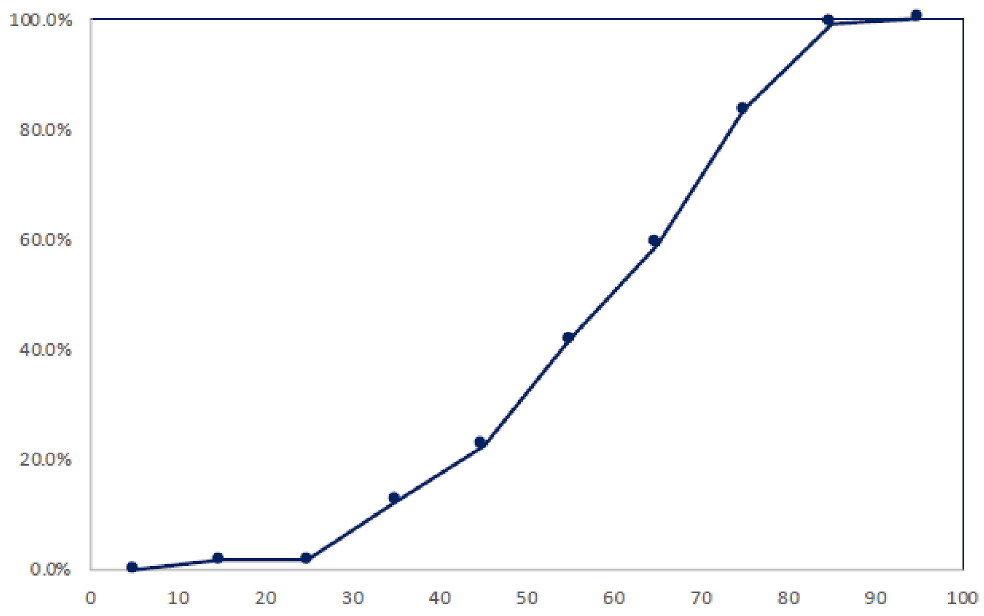


Frequency Distribution: an example

More graphic representations of the frequency distribution.

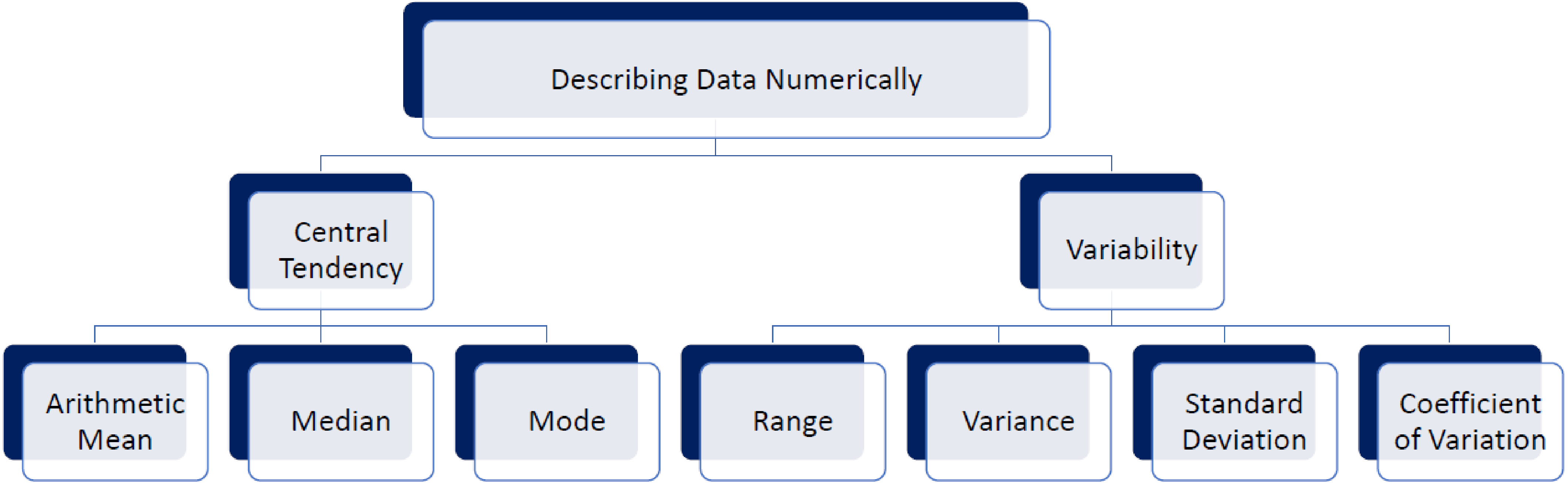
Frequency polygon	OGIVE (Cumulative frequency polygon)





Numerical Measures to Describe Data

- Data distributions can be summarized by measures of central tendency and measures of variability



Measures of Central Tendency: Arithmetic Mean

- Measures of central tendency provide values around which the data is distributed.
- The **arithmetic mean**, also known as the "mean," is the most common measure of central tendency

Definitions. The arithmetic mean for a population $\{x_1, x_2, x_3, \ldots, x_N\}$ is denoted by the **population mean**, μ,

$$\mu = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}.$$

If the data set is from a sample, then the **sample mean**, $\bar{x}$, is a statistic given by

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n},$$

where n = sample size.

Measures of Central Tendency: Arithmetic Mean

Example:

What is the arithmetic mean of the sample: 9, 56, 82, 14, 62, 92, 45, 28, 31, 71?

$$\bar{x} = \frac{9 + 56 + 82 + 14 + 62 + 92 + 45 + 28 + 31 + 71}{10} = 49$$

Commonly used but not suitable for data with outliers

Measures of Central Tendency: Arithmetic Mean

<u>Example:</u>

The following data is the weekly wages (in £) of a set of employees in a small workshop:

158, 138, 141, 148, 148, 146, 157, 252 (where the latter wage being that for the workshop manager)

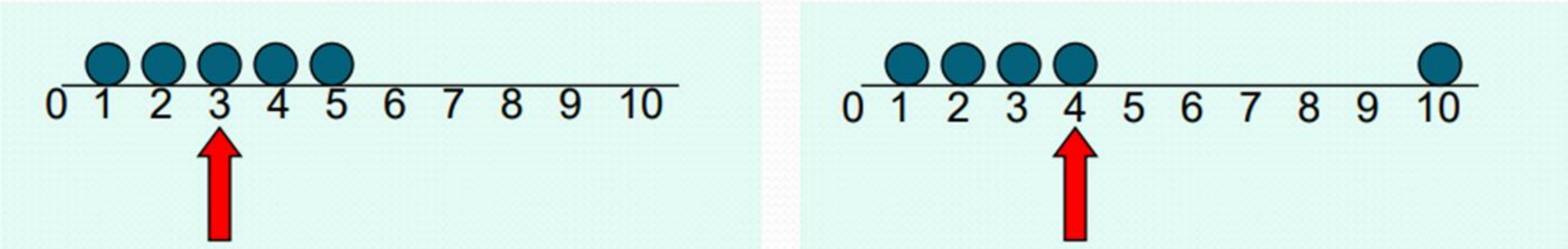
The mean of the wages is:

$$\bar{w} = \frac{158 + 138 + 141 + 148 + 148 + 146 + 157 + 252}{8} = 161$$

But the mean, 161, represents neither the workshop manager's wage nor the workshop members' wages

Measures of Central Tendency: Mean

- The arithmetic mean is the most common measure of central tendency. The main problem is that it is affected by outliers



$$\bar{x} = \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$$

$$\bar{x} = \frac{1+2+3+4+10}{5} = \frac{20}{5} = 4$$

Measures of Central Tendency: Median and Mode

- The **median** measures the central value of the ranked distribution
  - If *n* is odd: median is the middle observation
  - If *n* is even: median is the average of the two middle observations

- The **mode** measures the most frequently occurring value
  - Problems: does not always exist and is not always unique

Measures of Central Tendency: an example

<u>Example:</u>

Calculate the mean, median, and mode for the following sample:

88, 51, 63, 85, 79, 65, 79, 70, 73, 77, 65, 79

$$mean = \frac{88 + 51 + 63 + 86 + 79 + 65 + 79 + 70 + 73 + 77 + 65 + 79}{12} = 72.83$$

Order the numbers from small to large values:

$$51, 63, 65, 65, 70, 73, 77, 79, 79, 79, 85, 88$$

$$median = \frac{73 + 77}{2} = 75$$

$$mode = 79$$

Mean (average) in Finance

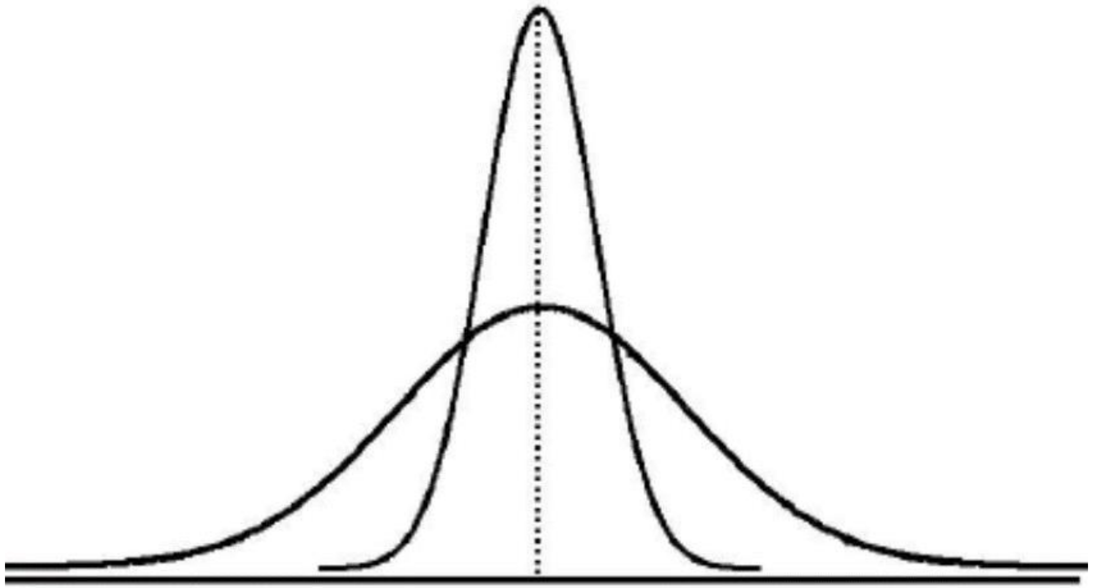
- **Stock Returns:** Investors often calculate the mean (average) of historical returns for a particular stock or portfolio. For example, if you have the monthly returns of a stock for the past year, you can calculate the mean return to get an idea of its average performance during that period.

- **Bond Yields:** In fixed-income securities like bonds, the mean yield can help investors understand the average income generated by the bond over its life.

- **Portfolio Performance:** When evaluating the performance of a diversified investment portfolio, the mean return of the entire portfolio is calculated to assess its overall profitability.

Measures of Variability

Same mean but different variation of data:



Measures of Variability: the Range

- The simplest measure of variability is the **range**.

Definition. **Range** is defined as the numerical difference between the smallest and largest values of the items in a set or distribution:

$$range = x_{max} - x_{min}.$$

- Range measures the total spread of the data.
- The greater the spread of data from the centre of the distribution, the larger the range.

Measures of Variability: the Range

Example:

What is the range of the numbers of defective products of each of two machines over 14 days?

Which machine is more reliable?

Machine 1	4, 7, 1, 2, 2, 6, 2, 3, 0, 4, 5, 3, 7, 4
Machine 2	3, 2, 2, 3, 3, 2, 4, 1, 1, 3, 2, 4, 2, 2

Range of Machine 1: 7 − 0 = 7

Range of Machine 2: 4 − 1 = 3

Measures of Variability

- Variance and standard deviation measure the dispersion around the mean.
- Coefficient of variation is standard deviation normalized by the mean.

Definition. The **population variance** from a population with size N, $\boldsymbol{\sigma^2}$, is

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

The **sample variance** from a sample with size n, s^2, is

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

The **population standard deviation** and the **sample standard deviation** are then σ and s.

The **population coefficient of variation** is

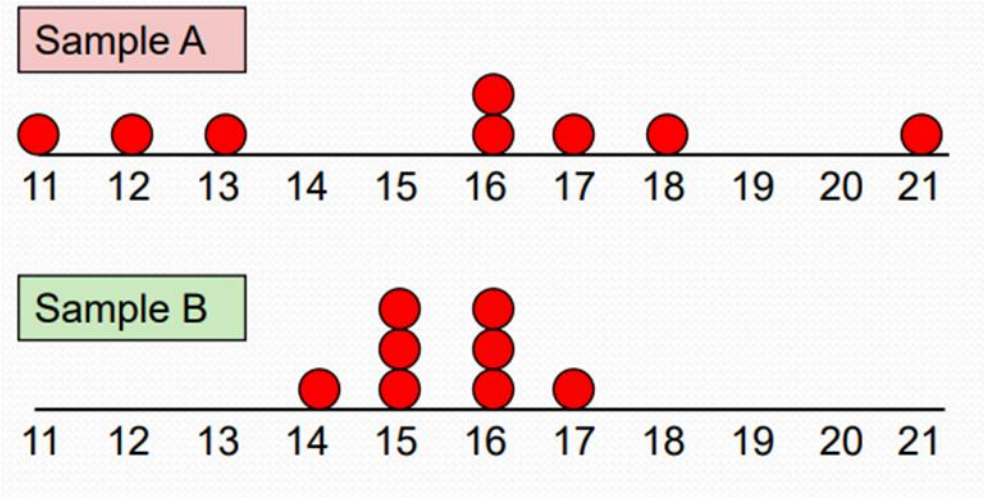
$$CV = \frac{\sigma}{\mu}\times 100\%, \text{ if } \mu > 0.$$

The **sample coefficient of variation** is

$$CV = \frac{s}{\bar{x}}\times 100\%, \text{ if } \bar{x} > 0.$$

Measures of Dispersion: Variance and Standard Deviation

Example:



Samples A and B have the same mean but different variance/standard deviation.

The standard deviation is higher for sample A meaning that there is a higher variability in sample A

	Sample A	Sample B
Mean	15.50	15.50
St. Deviation	3.34	0.93

Measures of Variability - Example

Example:

Take the previous example about the machines. What is the standard deviation and coefficient of variation for each machine over the 14 days?

Machine 1	4, 7, 1, 2, 2, 6, 2, 3, 0, 4, 5, 3, 7, 4
Machine 2	3, 2, 2, 3, 3, 2, 4, 1, 1, 3, 2, 4, 2, 2

Machine 1. The (population) mean is 3.57. The (population) standard deviation is $\sigma_1 = \sqrt{\frac{238}{14} - 3.57^2} = 2.06$

The coefficient of variation is $CV_1 = \frac{2.06}{3.57} = 57.7\%$

Machine 2. The mean is 2.43. The standard deviation is $\sigma_2 = \sqrt{\frac{94}{14} - 2.43^2} = 0.90$

The coefficient of variation is $CV_2 = \frac{0.90}{2.43} = 37.2\%$

Measures of Variability

Chebychev's Theorem. For any population with mean μ, standard deviation σ, and $k > 1$, the percentage of observations that lie within the interval $[\mu - k\sigma, \mu + k\sigma]$ is at least

$$100\left(1 - \frac{1}{k^2}\right)\%,$$

Where k is the number of standard deviations.

Example. Let the mean of exam grades be $\bar{x} = 72$ and the standard deviation be $s = 4$. Then, Chebychev's Theorem implies:

- 75% of the grades will lie in the interval [64, 80]
- 89% of the grades will lie in the interval [60,84].

Measure of Variability in Finance

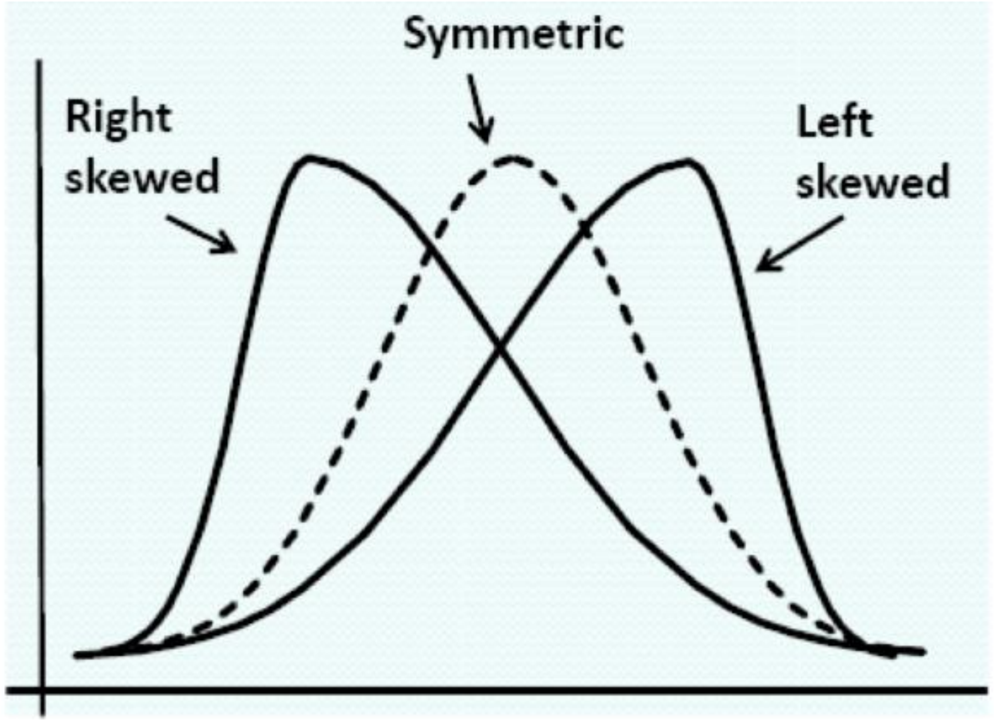
- **Risk Assessment:** Standard Deviation is commonly used in finance as a measure of risk. In the context of investment returns, it quantifies how much the returns fluctuate around the mean. A high variance indicates greater volatility and risk, while a low variance suggests stability.

- **Portfolio Diversification:** When constructing an investment portfolio, investors aim to reduce risk through diversification. Variance helps in assessing how the individual assets within a portfolio interact with each other. A portfolio with assets that have low or negatively correlated variances can result in lower overall portfolio variance and, consequently, reduced risk.

- **Option Pricing:** Variance is also used in the Black-Scholes option pricing model to estimate the potential future volatility of an underlying asset, which is a critical factor in determining the option's price.

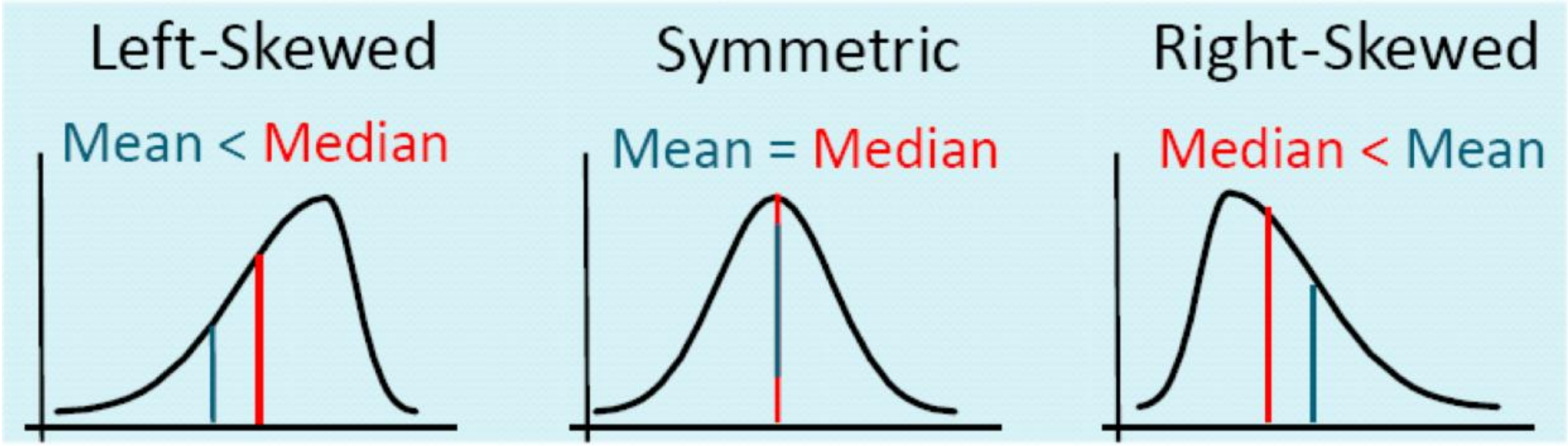
Skewness

Skewness is concerned with how non-symmetric or "lopsided" a frequency distribution is



Skewness

Skewness studies the relationship between the mean and the median



Covariance

- Covariance describes a linear relationship between two variables. A positive value indicates an increasing linear relationship and a negative value indicates a decreasing linear relationship.

Definitions. The **population covariance**, σ_{xy}, is

$$cov(x,y) = \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N},$$

Where x_i and y_i are observed values, μ_x and μ_y are the population means, and N is the population size.

The **sample covariance** from a sample with size n, s_{xy}, is

$$cov(x,y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1},$$

Where x_i and y_i are observed values, $\bar{x}$ and $\bar{y}$ are the sample means, and n is the sample size.

The **population correlation coefficient** is $\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$.

The **sample correlation coefficient** is $r = \frac{s_{xy}}{s_x s_y}$.

Covariance - Example

Example:

Take the previous example about the machines. What is the correlation between the number of defective products from Machine 1 and the number of defective products from Machine 2?

Machine 1	4, 7, 1, 2, 2, 6, 2, 3, 0, 4, 5, 3, 7, 4
Machine 2	3, 2, 2, 3, 3, 2, 4, 1, 1, 3, 2, 4, 2, 2

The covariance between the two machines:

$$cov(x, y) = \sigma_{xy} = \frac{\sum_{i=14}(x_i - 3.57)(y_i - 2.43)}{14} = -0.17$$

The correlation between of the two machines:

$$\rho = \frac{-0.17}{2.06 \times 0.90} = -0.09$$

Therefore, there is a slight negative correlation between defects from Machine 1 and defects from Machine 2.

Correlation Coefficient

The correlation coefficient is a measure of the strength of the relationship between or among variables.

- The **sample correlation coefficient** is calculated as :

$$r = \frac{s_{XY}}{s_X s_Y}$$

where:

s_{XY} is the sample covariance between X and Y

s_X is the sample standard deviation of X

s_Y is the sample standard deviation of Y

- The correlation coefficient is always between -1 and 1.

Correlation Coefficient

Interpretation of the correlation coefficient:

- $r = 0$ implies there is no correlation
- $r > 0$ implies a positive relationship between the two variables
- $r < 0$ implies a negative relationship between the two variables
- $r = 1$ implies a positive perfect linear relationship between the two variables
- $r = -1$ implies a negative perfect linear relationship between the two variables
- The closer the correlation coefficient is to –1 or +1, the stronger the linear relationship (negative or positive respectively)
- The closer it is to 0, the weaker the linear relationship.

Correlation Coefficient