Lesson Three : **Descriptive Statistics:** Presentation of quantitative data

Presentation and display of quantitative data: graphs,tables, scattergrams, bar charts. Distributions: normal and skewed distributions;characteristics of skewed distributions.
On this spread, we continue our discussion of descriptivestatistics and look at some of the ways in which data can be presented and how it may be distributed.

Key terms

Scattergram – A type of graph that represents the strength and direction of a relationship between co-variables in a correlational analysis.
Bar chart A graph used to represent the frequency of data; the categories on the *x*-axis have no fi xed order and there is no true zero.
Histogram Type of frequency distribution in which the number of scores in each category of continuous data are represented by vertical columns.There is a true zero and no spaces between the bars.
Negative skewed distribution Most of the scores are bunched towards the right. The mode is to the right of the mean because the mean is affected by the extreme scores tailing off to the left.
Normal distribution A symmetrical bell-shaped frequency distribution.
This distribution occurs when certain variables are measured, such as IQ or the life of a light bulb. Such 'events' are distributed in such a way that most of the scores are clustered close to the mid-point; the mean,median and mode are at the mid-point.
Pie chart A circular graph divided into sections or 'slices' each representing the proportion of the total.
Positive skewed distribution Most of the scores are bunched towards the left. The mode is to the left of the mean because the mean is affected by the extreme scores tailing off to the right.
Skewed distribution A distribution is skewed if one tail is longer than another, signifying that there are a number of extreme values to one side or the other of the mid-score.

DISPLAY OF QUANTITATIVE DATA

Graphs and tables should be simple so they can be read easily.
• They should clearly show the findings from a study.
• There should be a short but informative title.
• In a graph both axes should be clearly labelled. The *x*-axis usually goes across the page. In the case of a bar chart or histogram, it is usually the independent variable. The *y*-axis is usually frequency.
• Always use squared paper if you are hand-drawing graphs.

Tables

The measurements collected in a research study are called *raw data*—these are simply numbers that haven't been processed or analyzed yet. We can organize this data in tables and also summarize it using measures like averages (central tendency) and spread (dispersion). These summary tables make it much easier to understand and interpret the results.

Summary Statistics:

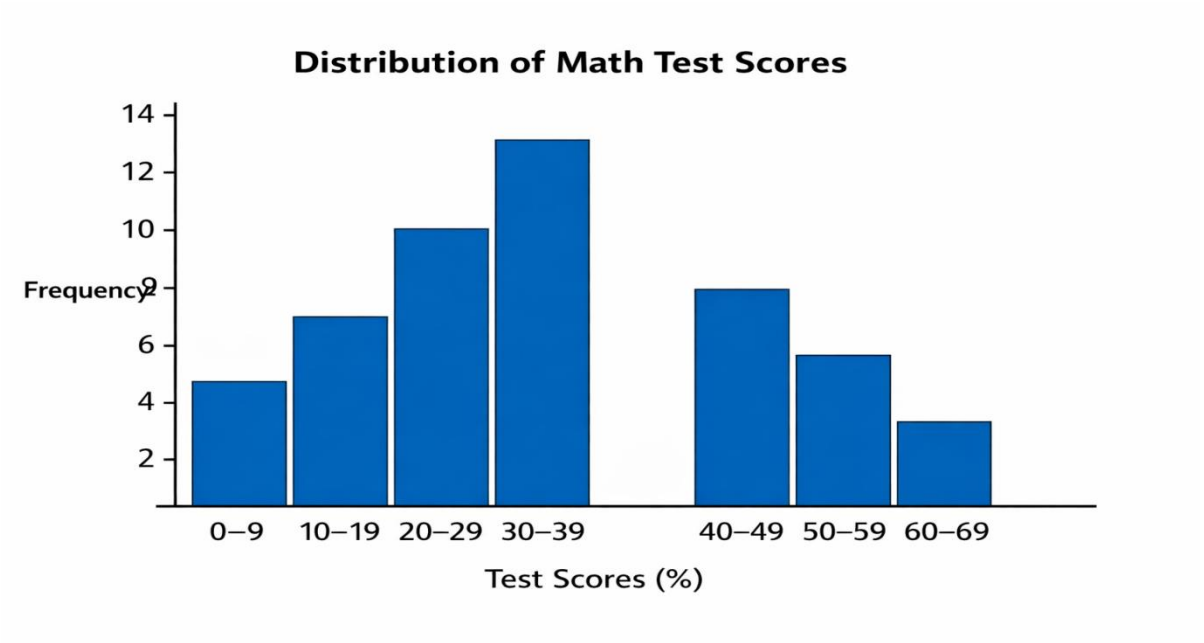
Condition	Mean	Median	Range	Standard Deviation
A	12.8	12	8–20	4.6
B	15.4	15	10–22	4.7

Frequency Table
Frequency data shows **how often each item occurs**. For example, a researcher might ask people to name their favorite pet. The responses can then be summarized in a frequency table, which makes it easy to see which options are most or least common.

Example: Favorite Pet Survey (n = 20)

Pet	Frequency	Percentage (%)
Deer	8	40
Cat	6	30
Fish	3	15
Bird	2	10
Other	1	5
Total	20	100

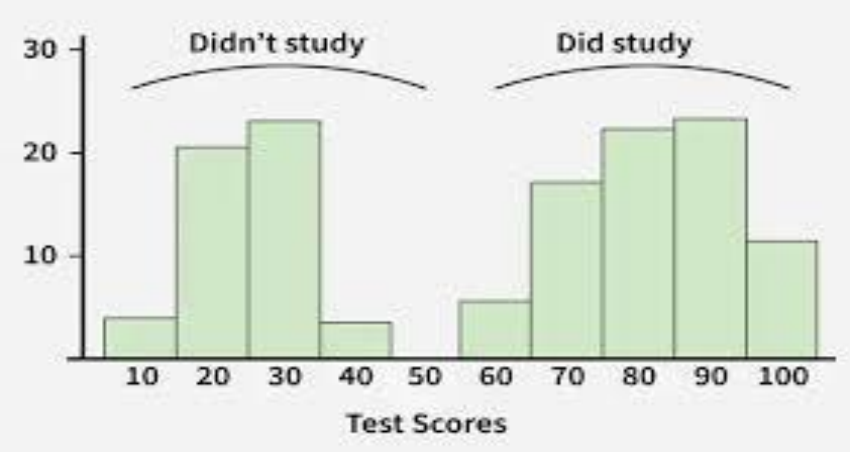
Bar chart
Frequency data can be represented in a **bar chart** (see Graph A). The height of each bar represents the frequency of each item. Bar charts are suitable for data that is not continuous.In a bar chart a space is left between each bar to indicate the lack of continuity.



Graph A – a bar chart showing the data in the table on the right on favourite pets

Histogram

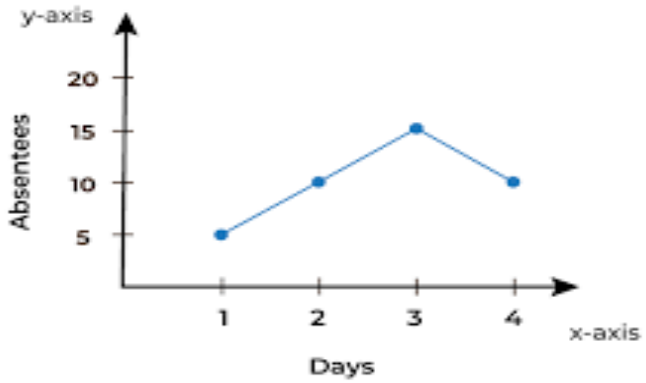
A **histogram** is similar to a bar chart except that the area within the bars must be proportional to the frequencies represented (see Graph C). In practice this means that the vertical axis (frequency) must start at zero. In addition the horizontal axis must be continuous (therefore you can't draw a histogram with data in categories). Finally, there should be no gaps between the bars.



Graph C – a histogram showing Distribution of Test Scores: Comparing Students Who Studied vs. Didn't Study"

Line Graph

A line graph is similar to a histogram in that it shows **continuous data on the x-axis**. Instead of bars, we mark the value at each point with a **dot**, and then **connect the dots with a line**. This makes it easy to see trends or changes in the data over time or across different conditions (see Graph D).

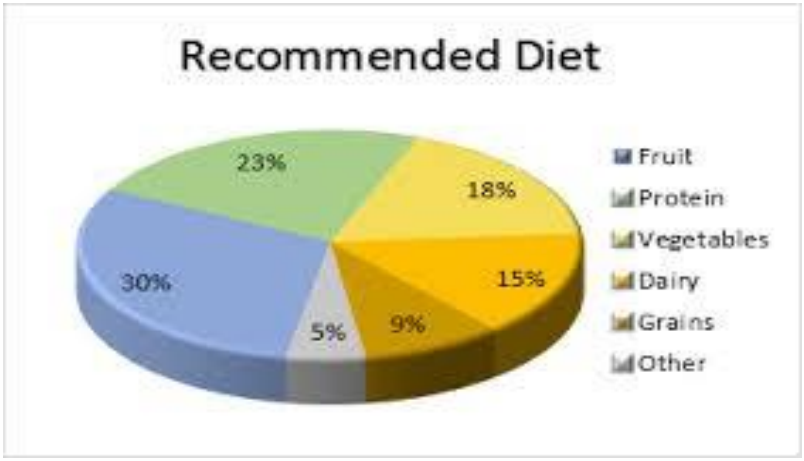


Graph D – a line graph showing Number of Absentees Over Four Days

https://www.youtube.com/watch?v=bmY28wdGL9M&t=210s

Pie chart

Pie charts are another way to represent frequency data or can be used to represent any proportion. Each slice of the 'pie' represents the proportion (or fraction) of the total. The size of each slice is calculated by working out the appropriate proportion out of 360 (since there are 360 degrees in a circle). (see Graph E).



Graph E – a Pie chart showing Fruit-Based Diets

DATA DISTRIBUTIONS

When we plot frequency data the *y*-axis represents frequency and the *x*-axis is the item of interest as in a histogram (see facing page). When doing this for large data sets we can see an overall pattern of the data called a distribution.

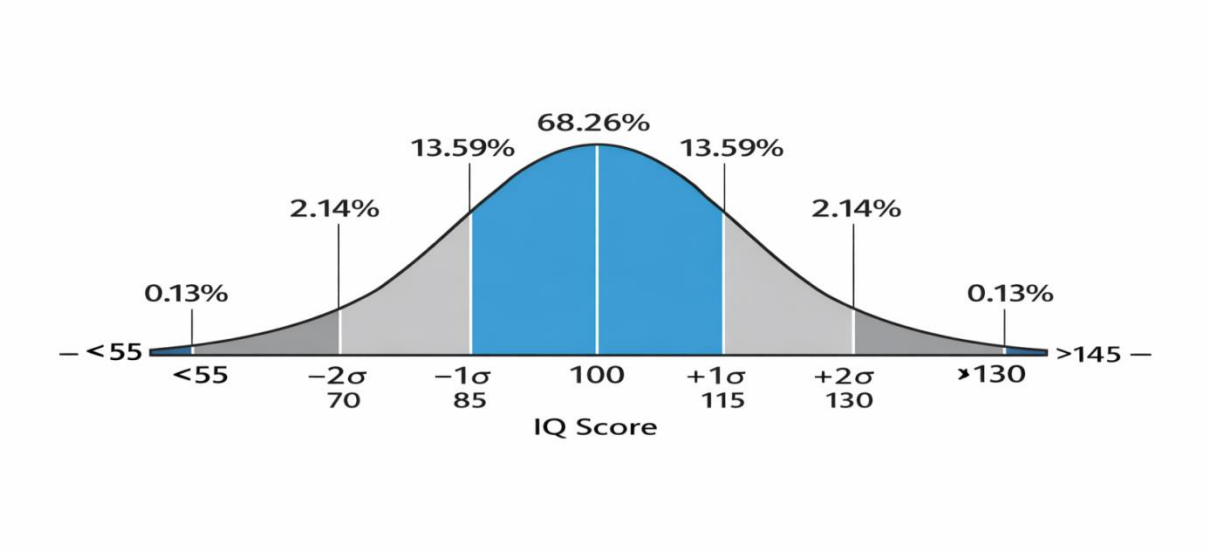
Normal distribution

The **normal distribution** has a classic bell-shaped curve. It is the predicted distribution when considering an equally-likely set of results. For example, if a light bulb has a **mean** life time of 100 hours we would expect some light bulbs to last a little less than this and some to last a little more. If we plot the lifetime of 1000 light bulbs we would get a normal distribution. Many human characteristics are normally distributed, such as shoe sizes or intelligence. A normal distribution has certain defi ning features:

• The mean, median and mode are all in the exact mid-point.
• The distribution is symmetrical around this mid-point.
• The dispersion of scores or measurements either side of the mid-point is consistent and can be expressed in **standard deviations**.

For any set of data that is normally distributed, 34.13% of the people will lie within one standard deviation below the mean and 34.13% of the people will lie within one standard deviation above the mean. A total of 68.26% lie one standard deviation above or below the mean. Therefore a total of 95.44% of people lie two standard deviations above or below the

mean, which means that only 4.56% lie in the area beyond this; 2.28% are at less than two standard deviations below the mean.



The percentage of people in each part of the normal distribution curve is shown above. Consider IQ. The mean score for IQ is 100 and one standard deviation is 15 points. If we tested a group of 1000 people this is what we should find:

• Score between 85 and 115 (one standard deviation either side of the mean) = 683 people (68.26% of 1,000)

• Score between 70 and 130 = 954 people (95.44% of 1,000)

• Score less than 55 = 1 (0.13% of 1,000)

Skewed Distributions

Definition:
In some populations, scores are not symmetrically distributed around the mean. When data are not symmetric, the distribution is called **skewed**.

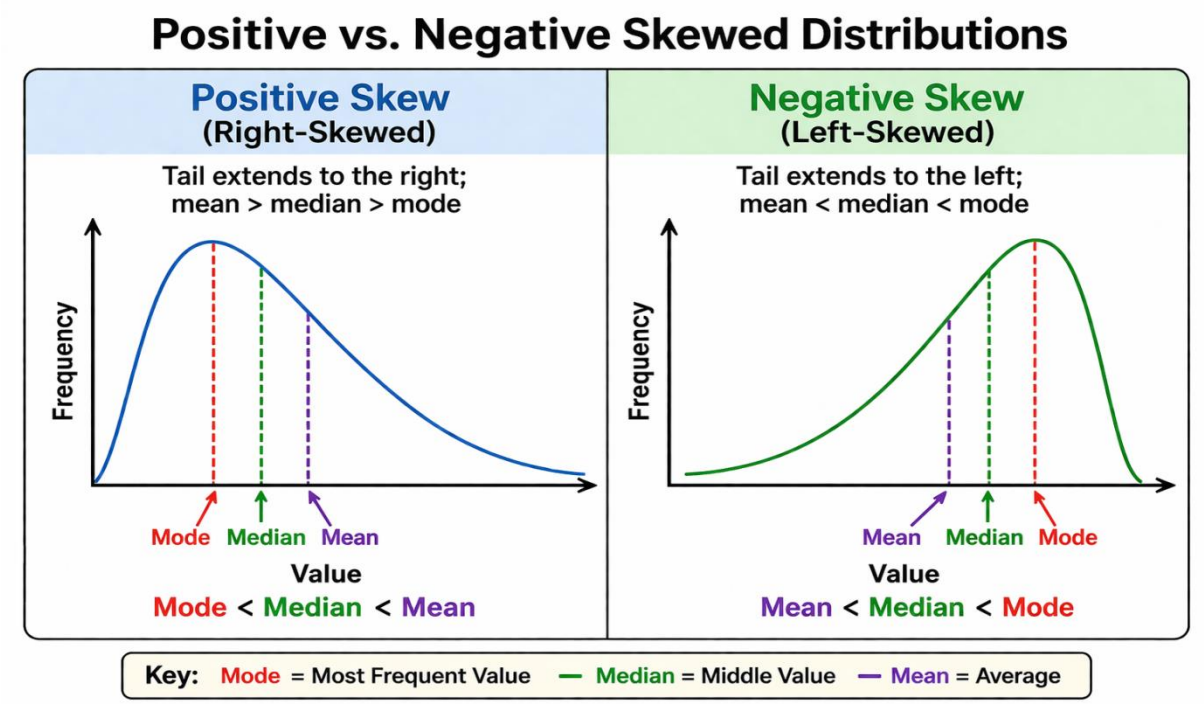
1. Positive Skew (Right-Skewed)

- Most scores cluster **toward the low end** of the scale.
- A few extreme **high scores** pull the mean to the right.
- Relationship of central tendencies:
  Mode < Median < Mean

Example: Consider a test of depression where 0–50 represents normal behaviour and 50+ represents **clinical** depression. If we plotted the distribution of scores for 100 people we would expect most scores to be towards the low end rather than the high end of this score range. This produces a **positive skewed distribution** as illustrated below left.

2. Negative Skew (Left-Skewed)

- The fact that there are a few extreme high scores has a strong effect on the mean which is always higher than the median and mode in a positive skew.
  The alternative is a **negative skewed distribution** (below right)
- Relationship of central tendencies:
  Mean < Median < Mode
- **Example:** Very easy exam. Most students score very high, with few low scores.



- **Positive correlation** – both variables increase together.
- **Negative correlation** – One variable increases while the other decreases.
- **Zero correlation** – no relationship.

Correlations

Key terms

Correlation – A mathematical technique in which a researcher investigates an association between two variables, called co-variables.

Co-variables – The variables investigated within a correlation, for example height and weight. They are not referred to as the independent and dependent variables because a correlation investigates the association between the variables, rather than trying to show a cause and effect relationship.

Positive correlation – Both variables increase together..

Negative correlation – One variable increases while the other decreases.
Zero correlation — no relationship

Types of Correlation

Correlation measures the strength and direction of a relationship between two or more **co-variables** (the variables being observed). Correlations are typically visualized using a **scattergram**, where one variable is plotted along the x-axis and the other along the y-axis. Each point on the graph represents the paired values of the two variables.

Examples of Correlation

1. **Positive Correlation**
   Suppose we examine caffeine consumption and anxiety levels. Participants record the number of caffeinated drinks they consume over a week and rate their anxiety on a 20-point scale. If the data were plotted on a scattergram, we might see a **positive correlation**: as caffeine intake increases, anxiety levels also increase.
2. **Negative Correlation**
   Participants could also record how many hours of sleep they get each week. Since high caffeine intake may disrupt sleep, we might find a **negative correlation**: as caffeine consumption rises, hours of sleep decrease.
3. **Zero Correlation**
   Participants might also count the number of dogs they see in the street during the same period. There is likely no relationship between caffeine intake and the number of dogs seen, so we would expect a **zero correlation**.

Correlation vs. Experiment

In an **experiment**, the researcher manipulates an **independent variable (IV)** to observe its effect on a **dependent variable (DV)**. This allows for conclusions about cause and effect: changes in the DV can be attributed to the IV.

In contrast, a **correlation** does not involve manipulation of variables. While correlations can show that two variables are related, they **cannot establish causation**. For example, even if caffeine consumption and anxiety levels are strongly positively correlated, we cannot conclude that caffeine causes anxiety. Other factors, such as personality, stress, or personal problems, may influence anxiety levels. These are called **intervening variables**, and they must be considered when interpreting correlational data.

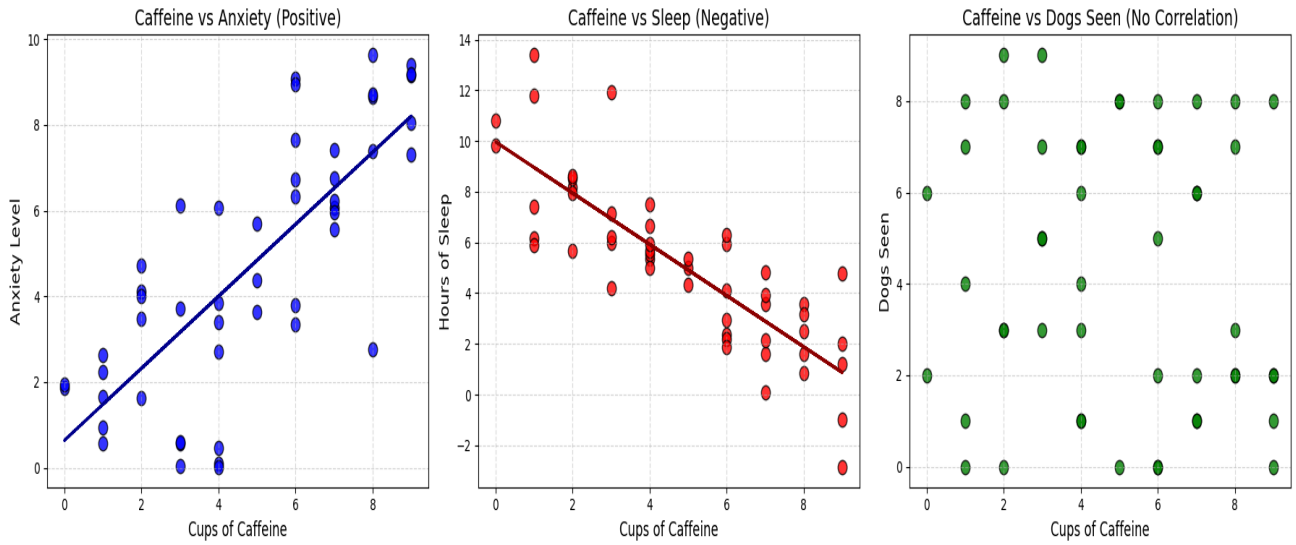
Apply it

Concepts: Positive and negative correlations

Are the following positive or negative correlations?

1. The more aggressive the parents, the more aggressive their children are.
2. The hotter the temperature, the fewer clothes people wear.
3. The fewer sweets eaten, the fewer llings needed.
4. The colder the weather, the higher people's fuel bills.
5. The more people exercise, the less their risk of heart disease.

6. More sociable people have more friends.
7. The fewer hours of daylight, the more depressed people there are.
8. The more lms you watch, the more interesting you are.



Scattergrams showing the relationships between number of caffeine drinks and anxiety level (left), number of caffeine drinks and hours sleep (middle) and number of caffeine drinks and number of dogs seen in the street (right).

Check Your Understanding

1. Calculate the mean, median, and mode for the following data sets:
    - SpeedUpp: 110, 59, 206, 89, 76, 141, 152, 98, 198, 57
    - Water: 122, 45, 135, 90, 42, 87, 131, 113, 129, 62
2. Calculate the range for the SpeedUpp dataset.
3. Explain the difference between mean and median. Which measure would you use if your dataset had extreme outliers?
4. Interpret the following standard deviations:
    - Condition A: Mean = 21.2, SD = 1.1
    - Condition B: Mean = 14.6, SD = 4.6
5. Classify these correlations as positive, negative, or zero:
    - More exercise → lower risk of heart disease
    - Number of hours studied → exam score
    - Number of dogs seen → caffeine intake
6. Identify whether these datasets are likely to be positively skewed, negatively skewed, or normally distributed:
    - Test scores: 100, 98, 95, 92, 91, 89, 88, 85, 80, 30
    - IQ scores in a large population

References

1. Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics* (5th ed.). Sage Publications.
2. Gravetter, F. J., & Wallnau, L. B. (2020). *Statistics for the Behavioral Sciences* (10th ed.). Cengage Learning.
3. Healey, J. F. (2019). *Statistics: A Tool for Social Research* (11th ed.). Cengage Learning.
4. YouTube: Descriptive Statistics – Link