# Introduction to Statistics

## These are the learning objectives for today's class:

- Be able to define the following terms:
    1. Statistics
    2. Data
    3. Discrete Data and Continuous Data
    4. Frequency Distributions
    5. Grouped Distributions

- Be able to distinguish between the various types of statistical diagrams (i.e. Bar Chart; Pie Chart; Histogram; Frequency polygon) and know when it is appropriate to use them.

- Have a clear understanding of the three different types of **Statistical Averages** (Mode, Mean and Median) and what they are used for. This also includes drawing **Cumulative Frequency curves).**

- Being comfortable enough to handle any statistical question at the CXC level.

---

***Statistics*** is the branch of mathematics that deals with the collection, organization, analysis, and interpretation of numerical data. Statistics is especially useful in drawing general conclusions about a set of data from a sample of the data.

***Data*** is a collection of facts, such as values or measurements. It can be numbers, words, measurements, observations or even just descriptions of things.

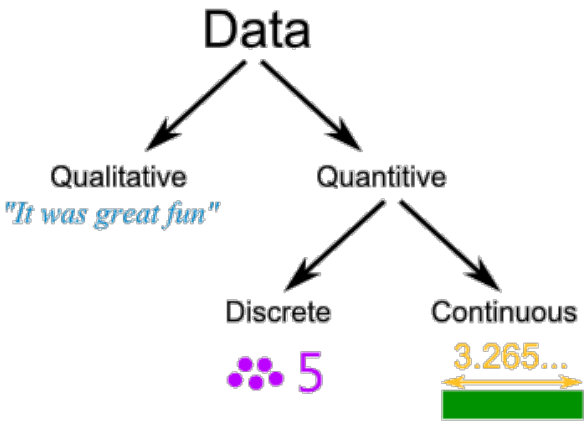
Data can be **qualitative** or **quantitative**.

- **Qualitative data** is descriptive information (it *describes* something)
- **Quantitative data** is numerical information (numbers).

**NB**. It should be noted that in mathematical statistics we mainly deal with quantitative data because this is the type that can be sorted and we can perform calculations on.

**NB**. Only when data is organized in a meaningful way we call this **information**.

Kevin Small
www.cxctutor.org
www.cxctutor.com




And **Quantitative data** can also be Discrete or Continuous:

- **Discrete data** can only take certain values (like whole numbers)
- **Continuous data** can take any value (within a range)

Put simply: **Discrete data** is counted, **Continuous data** is measured

Example: What do we know about Arrow the Dog?

**Qualitative**:

- He is brown and black
- He has long hair
- He has lots of energy

**Quantitative**:

- Discrete:
  - He has 4 legs
  - He has 2 brothers
- Continuous:
  - He weighs 25.5 kg
  - He is 565 mm tall

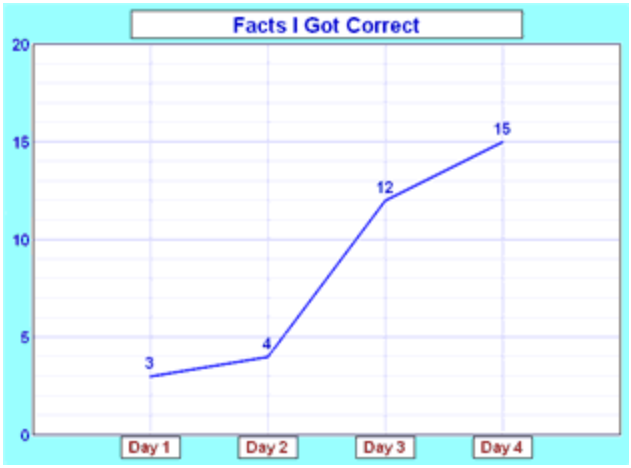
**Statistical Diagrams:** There are various ways of representing data such as:

- **Line Graph** - A graph that shows information that is connected in some way (such as change over time)

For example: You are learning/revising math each day, and at the end of each day you do a short test to see how many math questions you can get correct in 30 minutes. These are the results:

| *Table: Questions I got Correct* | | | | |
|---|---|---|---|---|
| **X-axis** | Day 1 | Day 2 | Day 3 | Day 4 |
| **Y-axis** | 3 | 4 | 12 | 15 |

And here is the same data as a Line Graph:




From this graph both you and your parents can see that you are improving significantly!

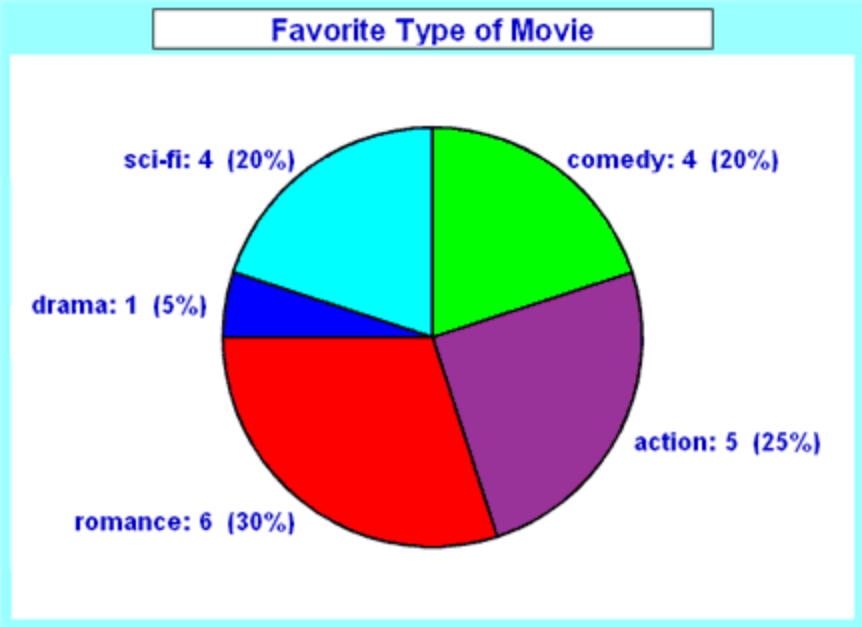
**NB.** I put in the column with the X-axis and Y-axis for your convenience but you should know that we put data that we can control on the X-axis we call this (independent or controllable variable) the data that we have little control over but depends on the independent variable (i.e. you get more questions correct the more you study) goes on the Y-axis and this is called the (dependent variable).

- **Pie Chart** - A special chart that uses "pie slices" to show relative sizes of data.

Imagine you just did a survey of your friends to find which kind of movie they liked best. Here are the results:

| *Table: Favorite Type of Movie* | | | | |
|---|---|---|---|---|
| **Comedy** | **Action** | **Romance** | **Drama** | **SciFi** |
| 4 | 5 | 6 | 1 | 4 |

You could show this information by means of a pie chart:




It is a really good way to show relative sizes: it is easy to see which movie types are most liked, and which are least liked, at a glance.

### Let's see how to construct a pie chart:

First, put your data into a table (like above), then add up all the values to get a total:

| Comedy | Action | Romance | Drama | SciFi | TOTAL |
|---|---|---|---|---|---|
| 4 | 5 | 6 | 1 | 4 | **20** |

Next, divide each value by the total and multiply by 100 to get a percent:

| Comedy | Action | Romance | Drama | SciFi | TOTAL |
|---|---|---|---|---|---|
| 4 | 5 | 6 | 1 | 4 | **20** |
| 4/20 = **20%** | 5/20 = **25%** | 6/20 = **30%** | 1/20 = **5%** | 4/20 = **20%** | **100%** |

Now you need to figure out how many degrees for each "pie slice" (correctly called a sector).

A Full Circle has 360 degrees, so we do this calculation:

| Comedy | Action | Romance | Drama | SciFi | TOTAL |
|---|---|---|---|---|---|
| 4 | 5 | 6 | 1 | 4 | **20** |
| 4/20 = **20%** | 5/20 = **25%** | 6/20 = **30%** | 1/20 = **5%** | 4/20 = **20%** | **100%** |
| 4/20 × 360° = **72°** | 5/20 × 360° = **90°** | 6/20 × 360° = **108°** | 1/20 × 360° = **18°** | 4/20 × 60° = **72°** | **360°** |

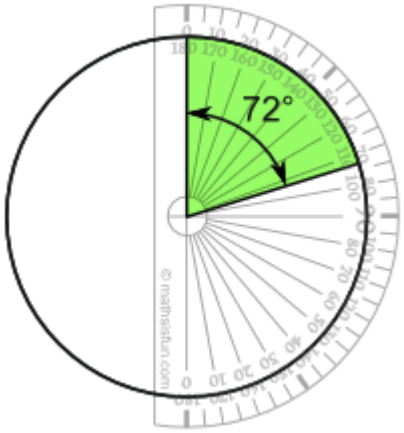
Now you are ready to start drawing!

Draw a circle.

Then use your protractor to measure the degrees of each sector.

Here I show the first sector ...

... You can do the rest!




- **Bar Graph** (also called Bar Chart) is another graphical display of data using bars of different heights.

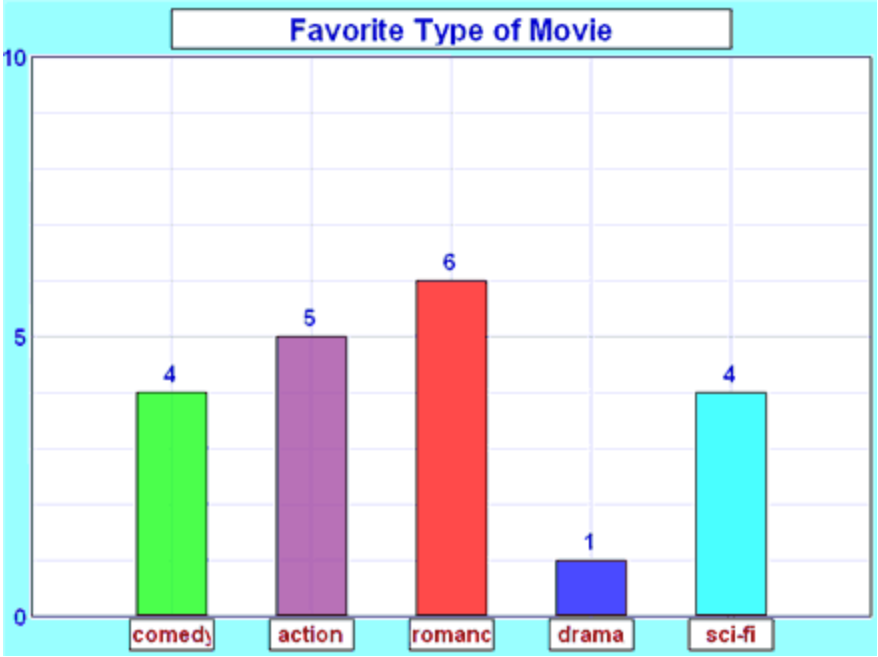
Imagine you just did a survey of your friends to find which kind of movie they liked best.

Here are the results:

| ***Table: Favorite Type of Movie*** | | | | |
|---|---|---|---|---|
| **Comedy** | **Action** | **Romance** | **Drama** | **SciFi** |
| 4 | 5 | 6 | 1 | 4 |

You could show that on a bar graph like this:

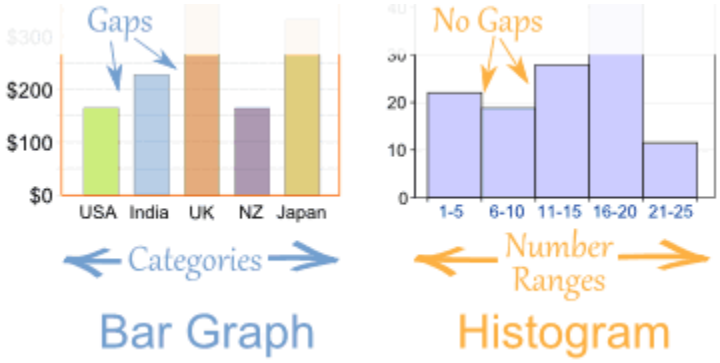
Kevin Small
www.cxctutor.org
www.cxctutor.com




## Histograms vs. Bar Graphs

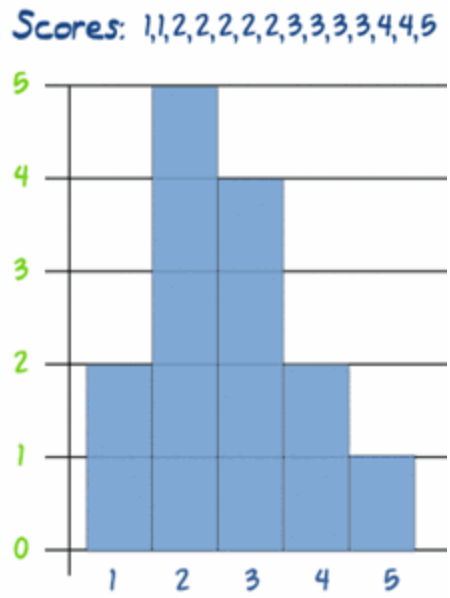
Bar Graphs are good when your data is in **categories**(such as "Comedy", "Drama", etc).

But when you have **continuous data**(such as a person's height) then use a **Histogram**.




## Frequency Histogram

A Frequency Histogram is a special histogram that uses vertical columns to show frequencies (how many times each score occurs):




Here I have added up how often 1 occurs (2 times), how often 2 occurs (5 times), etc, and shown them as a histogram.

In summary:

1. **Bar Chart** shows comparison of figures and numbers of items.
2. **Line Graph** shows trends over a period of time.
3. **Pie Chart** shows percentage of fractions of a whole.

---

Statistical Average

As we learnt before raw data doesn't mean much until it is arranges into a frequency distribution or until it is represented as a histogram. A second way of making the data more understandable is to try to find a single value which can best represent all the values in a distribution. This single representative value is called an **average**.

In statistics several kind of averages are used. The more important are:

a) **Mean** ( the arithmetic mean i.e. sum /number of values)
b) **Median** (middle number after data is arranged in order)
c) **Mode** (most frequently occuring number)

Statistical Averages continued:

## The Arithmetic Mean

An arithmetic mean is a fancy term for what most people call an "average." When someone says the average of 10 and 20 is 15, they are referring to the arithmetic mean. The simplest definition of a mean is the following: Add up all the numbers you want to average, and then divide by the number of items you just added.

For example, if you want to average 10, 20, and 27, first add them together to get 10+20+27= 57. Then divide by 3 because we have three values, and we get an arithmetic mean (average) of 19.

Want a formal, mathematical expression of the arithmetic mean?

$$\text{arithmetic mean} = \frac{\sum_{n=1}^{k} x_n}{k}$$

That's just a fancy way to say "the sum of k different numbers divided by k."

## Negative Numbers

How do you handle negative numbers? Adding a negative number is the same as subtracting the number (without the negative). For example 3 + (-2) = 3-2 = 1.

Knowing this, let us try an example:

Example 3: Find the mean of these numbers:

3, -7, 5, 13, -2

- The sum of these numbers is **3 - 7 + 5 + 13 - 2 = 12**
- There are **5** numbers.
- The mean is equal to **12 ÷ 5 = 2.4**

The mean of the above numbers is 2.4

## The Median

Kevin Small
www.cxctutor.org
www.cxctutor.com

If a set of values is arranged in ascending or descending order of size the median is the value which lies half-way along the series. Thus the median of **3, 4, 4, 5, 6, 8, 8, 9, 10** is **6** because there are four numbers below this value and four numbers above it.

## Example 2

Look at these numbers:

3, 13, 7, 5, 21, 23, 39, 23, 40, 23, 14, 12, 56, 23, 29

If we put those numbers in order we have:

3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 39, 40, 56

There are **fifteen** numbers. Our middle number will be the **eighth** number:

3, 5, 7, 12, 13, 14, 21, **23**, 23, 23, 23, 29, 39, 40, 56

The median value of this set of numbers is **23**.

When there are and even number of values in the set the median is found by taking the average of the two middle values. In other words we need to find the **middle pair** of numbers, and then find the value that would be half way between them. This is easily done by adding them together and dividing by two.

An example will help:

3, 13, 7, 5, 21, 23, 23, 40, 23, 14, 12, 56, 23, 29

If we put those numbers in order we have:

3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 40, 56

There are now **fourteen** numbers and so we don't have just one middle number, we have a **pair of middle numbers**:

3, 5, 7, 12, 13, 14, **21**, **23**, 23, 23, 23, 29, 40, 56

In this example the middle numbers are **21 and 23**.

To find the value half-way between them, add them together and divide by 2:

**21 + 23 = 44**
**44 ÷ 2 = 22**

And, so, the **Median** in this example is **22**.

### The Mode

The mode of a set of values is the value which occurs most **frequently**. It is not necessarily unique (two different values could occur the same number of times).

To find the mode, or modal value, first put the numbers **in order**, then count how many of each number.

**Example:**

3, 7, 5, 13, 20, 23, 39, 23, 40, 23, 14, 12, 56, 23, 29

**In order** these numbers are:

3, 5, 7, 12, 13, 14, 20, **23, 23, 23, 23**, 29, 39, 40, 56

This makes it easy to see which numbers appear **most often**.

In this case the mode is **23**.

**Example:**

What is the mode of [-6, 2, 67, 4, 3, 9, -6, 5, 2, 0]?

**Solution:**

The mode is -6 and 2, because each of those values occurs twice and nothing else occurs more often. In other words, they are "tied" for the lead.

Having two modes is called **"bimodal"**.

Having more than two modes is called **"multimodal"**.

Now let's introduce the concept of frequency into statistics.

One way of organizing raw data into order is to arrange them in the form of a frequency distribution. Consider below the marks of 50 students obtained in a test:

4 3 5 4 3 5 5 4 3 6 5 4 5 3 4 4 5 5 7
4 3 4 3 4 5 4 3 6 1 3 6 3 2 6 6 3 5 2
7 5 7 1 7 6 5 8 6 4 3 5

The number of students obtaining 3 marks is found, the number obtaining 4 marks is found, and so on. A tally chart is the best way of doing this.

| Mark | Tally | Frequency |
|---|---|---|
| 1 | 1 1 | 2 |
| 2 | 1 1 | 2 |
| 3 | ~~1 1 1 1~~ ~~1 1 1 1~~ 1 | 11 |
| 4 | ~~1 1 1 1~~ ~~1 1 1 1~~ 1 | 11 |
| 5 | ~~1 1 1 1~~ ~~1 1 1 1~~ 1 1 | 12 |
| 6 | ~~1 1 1 1~~ 1 1 | 7 |
| 7 | 1 1 1 1 | 4 |
| 8 | 1 | 1 |
| | Total | 50 |

Kevin Small
www.cxctutor.org
www.cxctutor.com