GCSE STATISTICS

You should know:

1) How to draw a **frequency diagram**:

e.g.	NUMBER	TALLY	FREQUENCY
	1		3
	2		5

2) How to draw a **bar chart**, a **pictogram**, and a **pie chart**.

3) How to use **averages**:
   a) **Mean** - *add up all the numbers and divide by how many there are.*
   b) **Median** - *Put all the numbers in order and find the middle one. If there are two, then find the mean of them.*
   c) **Mode** - *The number that appears most often (there can be more than one).*

4) How to calculate the **range**:
   The difference between the biggest number and the smallest number.

5) How to use a frequency table to find averages:

NUMBER	TALLY	FREQUENCY	NUMBER x FREQUENCY
1		3	3
2		5	10
		Total - 8	*Total* - 13

   MEAN - *13 divided by 8 gives 1.625*
   MEDIAN - *working down the table gives 2 and 2, the mean of which is 2.*
   MODE - *it is clear that 2 is the mode.*
   RANGE - *'2 - 1 = 1'*

6) How to use **multiple** bar charts and **component** bar charts.

7) How to use **grouped data**.

8) How to draw a **histogram**, and a **frequency polygon** (lines to join the middle of each bar together).

9) How to calculate averages for grouped data:

GROUP	MID-POINT	FREQUENCY	MID-POINT x FREQUENCY
> 0 <= 50	25	6	150
> 50 <= 100	75	8	600
> 100 <= 150	125	11	1,375
		Total - 25	*Total* - 2125

MEAN - *estimated at 2125 / 25 = 85*
MODAL GROUP - *'> 100 <= 150'*
MEDIAN - *the number will be in the '> 50 <= 100' group. The median is the 13th number, and there are the 7th up to the 14th numbers in this group. There are 8 numbers, so the 13th number will be 7/8 into the group. It covers 50, so 7/8 of 50 = 43.75. This is the estimated median.*

10) How to write out questionnaires.

11) That a **sample** can be used in a survey when you cannot ask everyone. It should be selected at **random**, for if groups of people are not included then it will be **biased**. The bigger a sample is, the better. If it is not possible to use **random sampling**, then **systematic sampling** should be used. This is when the names are in a list, and you choose, for instance, every 5th person on the list.

12) That the **expected frequency** is the amount of times you would expect something to happen, e.g. if you tossed a coin 100 times, you would expect 50 heads and 50 tails.

13) That the **relative frequency** is the relative number of times you would expect something to happen, e.g. you would expect a coin to show heads ½ of the time and tails ½ of the time when it is tossed. This is the **probability** of an event happening, and the **sample space** is the list of events that could happen (i.e. heads or tails).

14) That the **absolute error** of a measurement is the maximum error that can be made, e.g. a measurement to the nearest millimetre has an absolute error of ½ a millimetre. The **relative error** is the absolute error divided by the measurement taken, and the **percentage error** is the relative error times 100.

15) That data can be either **quantitative** (when it is numbers) or **qualitative** (when it is something else, like a description).

16) **Quantitative** data can be either **discrete** (when it is only certain numbers and can't be in between them) or **continuous** (when it can be absolutely any number and to any degree of accuracy, like a measurement for instance).

17) That you must use the **class boundaries** of data to find the mid-point of each group.

18) That to estimate the **mode** with **grouped data**, you need to find the modal group on the **histogram**, and draw a line from the top left hand corner of the bar to the top left

hand corner of the bar on its right, and do the same for the top right hand corner and the bar on the left. You can then draw a line straight down to the x-axis, and take the reading as an estimate of the mode.

19) How to draw a **cumulative frequency curve**.

20) That to find the **median** on a **cumulative frequency curve**, you find the middle value on the y-axis, draw a horizontal line to the curve, then a vertical line straight down to the x-axis to take a reading.

21) That the **lower quartile** has a y value of the total number of values divided by 4, and the **upper quartile** y value is the total number of values minus the lower quartile y value. You can draw these on the graph in the same way as the median. The **interquartile range** is the upper quartile minus the lower quartile, and it shows how spread out the data is.

22) That the **semi-interquartile** range is half the interquartile range. Sometimes data is split into **deciles** (10 equal parts) or **percentiles** (100 equal parts) instead.

23) How to draw a **scatter graph** and **line of best fit**.

24) That the line of best fit can show the **correlation** of the data. If it is sloping upwards away from the origin, then it is **positively correlated**. If it is sloping the other way, then it is **negatively correlated**. If you cannot draw a line of best fit, then the data is **uncorrelated**.

25) That if you want to find a value beyond the line of best fit, then it is called **extrapolating**. Finding a value between the known values is called **interpolating**.

26) That a **stratified random sample** is when you split people for a survey into **groups**, and choose people at **random** from each group. The number chosen though, depends on the size of the group.

27) That **quota sampling** is when you choose certain **types** of people to ask. It is not random though, so it can be **biased**.

28) That an **event** is anything that happens when you are calculating probabilities. Several events make up an **outcome**.

29) That events are **independent** when they don't affect each other. Events are **dependant** if they affect each other.

30) How to draw a **tree diagram**.

31) That the probability of something happening is the **number of your outcomes** over the **total number of outcomes**.

32) That events are **mutually exclusive** if they have no points in common, and events that have points in common are **non-mutually exclusive**.

33) That you can use special **symbols** to represent numbers in statistics:
x *- represents all the data collected.*
n *- stands for the number of items collected.*
å *- (sigma) means to add all the numbers up.*
x *- (x bar) represents the mean.*

$$\text{e.g. } \bar{x} = \frac{\sum x}{n}$$

34) That the **deviation** from the mean is the number minus the mean (x - x). The **variance** of a set of numbers shows how spread out they are:

$$\text{variance} = \frac{\sum (x - \bar{x})^2}{n}$$

35) That the **standard deviation** can be used to measure the spread of data. It is useful because it takes all the numbers into account:

$$\text{standard deviation} = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

36) That the **mean deviation** can also be used, but it doesn't take into account the sign, and doesn't do any squaring. It uses the **modulus** (| |) to show that it doesn't matter what the sign is:

$$\text{mean deviation} = \frac{\sum |x - \bar{x}|}{n}$$

37) That to compare two sets of data which aren't from the same experiment, you need to **scale** one or both of the sets of data. You first need to **standardise** the data, then you scale it:

$$\text{standardise - } s = \frac{x - \bar{x}}{s}$$

$$\text{scale - } \bar{x} + (ss)$$

38) That the **frequency distribution** shows all the possible results, and the frequency of each result.

39) That the **shape** of the frequency distribution can be described:

SYMMETRICAL - *has the mode at the centre.*
BIMODAL - *has two peaks.*
SKEW - *not symmetrical.*
POSITIVELY SKEWED - *skewed towards the y-axis.*
NEGATIVELY SKEWED - *skewed away from the y-axis.*
NORMAL DISTRIBUTION - *a special curve with few high and low values, but with the highest frequencies around the middle.*

40) How to draw a **box plot**, and a **stem and leaf diagram**.

41) That to find the **equation** of a line of best fit, you need the **gradient** of the line, and the **y-intercept** of the line. The gradient can be found by drawing a line horizontally across from the y-axis, then going vertically up or down again, to reach the x-axis. The length of the vertical line divided by the length of the horizontal line gives the gradient. The y-intercept is the y-coordinate of the point the line crosses the y-axis (it may need to be extended). This will give the following equation:

$$y = (\text{gradient})x + (\text{y-intercept}) \quad \textbf{OR} \quad y = mx + c$$

42) That the gradient is called the **regression coefficient**. The higher the regression coefficient, the **steeper** the line.

43) How to use **Spearman's coefficient of Rank Correlation** (*d* is the difference between ranks, and *n* is the number of things placed in order):

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

44) That diagrams can be **misleading** by changing scales, and by increasing the width of bars / pictures, as well as the height.

45) That 5! means 5 **factorial**, or 5 x 4 x 3 x 2 x 1. The same applies for any number (e.g. 3! = 3 x 2 x 1).

46) That to calculate the **number of ways** of **choosing several items** out of a group of items (e.g. 5 out of 24), you use the following formula (*n* is the number to choose from, and *r* is the number to choose):

$$^{n}C_{r} = \frac{n!}{r!(n-r)!}$$

47) To find the **probabilities** of several events (**all the same**) happening one after another in an experiment, you use the following formula:

$$\text{Probability } (r \text{ successes}) = {}^{n}C_{r}\, p^{r} \left(1-p\right)^{n-r}$$

48) That a **time series** is when you collect data over a series of time. If a graph goes up and down quite regularly, it is called **variation**. There are three types of variation:

SEASONAL - *a regular change that can happen over any period of time (weeks, months, years etc.).*
CYCLICAL - *a change that keeps on happening but is not regular.*
RANDOM - *a change that isn't seasonal or cyclical.*

49) That **moving averages** are used to show the general **trend** by evening out seasonal variation, and how to use moving averages.

50) That you can **extend** the trend line on a time series graph to **predict** what could happen in the future. After extending the trend line, you need to add the seasonal variation. To do this you look at the previous times when seasonal variation takes place, and calculate how far above or below the trend line you should be. You then find the mean, and add it to (or subtract it from) the *x* value of the trend line at that point.

51) That the **birth rate** is the number of births for every 1000 in the population in one year. It can be calculated using the following formula:

$$\frac{\text{total number of births in the year}}{\text{total population at the middle of the year}} \times 1000$$

52) How to **standardise** the **birth rate** and **death rate** for different populations.

53) That to calculate the **weighted mean** you use the following formula:

$$\text{weighted mean} = \frac{w_1 \times score_1 + w_2 \times score_2 + w_3 \times score_3 + \ldots\ldots}{w_1 + w_2 + w_3 + \ldots\ldots}$$

54) That an **index number** is a percentage used to compare prices. A **base price** is used to compare other prices with, and is given the index number of 100. The formula for calculating other index numbers is as follows:

$$\text{index number} = \frac{\text{price}}{} \times 100$$

base price

55) That a **retail price index** is an **index number** for lots of different items. It can be used to compare the cost of living from one year to another. The formula to calculate a retail price index is as follows (weighting = w, and index number = d):

$$\text{retail price index} = \frac{\sum(w \times d)}{\sum w}$$