Statistics: Introduction and recap*

Free High School Science Texts Project

1 Introduction

Information in the form of numbers, graphs and tables is all around us; on television, on the radio or in the newspaper. We are exposed to crime rates, sports results, rainfall, government spending, rate of HIV/AIDS infection, population growth and economic growth.

This chapter demonstrates how Mathematics can be used to manipulate data, to represent or misrepresent trends and patterns and to provide solutions that are directly applicable to the world around us.

Skills relating to the collection, organisation, display, analysis and interpretation of information that were introduced in earlier grades are developed further.

2 Recap of Earlier Work

The collection of data has been introduced in earlier grades as a method of obtaining answers to questions about the world around us. This work will be briefly reviewed.

2.1 Data and Data Collection

2.1.1 Data

Definition 1: Data
Data refers to the pieces of information that have been observed and recorded, from an experiment or a survey. There are two types of data: primary and secondary. The word "data" is the plural of the word "datum", and therefore one should say, "the data are" and not "the data is".

Data can be classified as *primary* or *secondary*, and primary or secondary data can be classified as *qualitative* or *quantitative*. Figure 1 summarises the classifications of data.

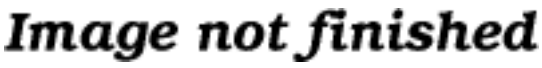

Figure 1: Classes of data.

*Version 1.1: Aug 2, 2011 2:32 am +0000
†http://creativecommons.org/licenses/by/3.0/

Primary data: describes the original data that have been collected. This type of data is also known as *raw* data. Often the primary data set is very large and is therefore summarised or processed to extract meaningful information.

Qualitative data: is information that cannot be written as numbers, for example, if you were collecting data from people on how they feel or what their favourite colour is.

Quantitative data: is information that can be written as numbers, for example, if you were collecting data from people on their height or weight.

Secondary data: is primary data that has been summarised or processed, for example, the set of colours that people gave as favourite colours would be secondary data because it is a summary of responses.

Transforming primary data into secondary data through analysis, grouping or organisation into secondary data is the process of generating information.

2.1.2 Purpose of Collecting Primary Data

Data is collected to provide answers that help with understanding a particular situation. Here are examples to illustrate some real world data collections scenarios in the categories of qualitative and quantitative data.

2.1.3 Qualitative Data

- The local government might want to know how many residents have electricity and might ask the question: "Does your home have a safe, independent supply of electricity?"
- A supermarket manager might ask the question: “What flavours of soft drink should be stocked in my supermarket?" The question asked of customers might be “What is your favourite soft drink?” Based on the customers’ responses (i.e. which flavours are chosen), the manager can make an informed decision as to what soft drinks to stock.
- A company manufacturing medicines might ask “How effective is our pill at relieving a headache?” The question asked of people using the pill for a headache might be: “Does taking the pill relieve your headache?” Based on responses, the company learns how effective their product is.
- A motor car company might want to improve their customer service, and might ask their customers: “How can we improve our customer service?”

2.1.4 Quantitative Data

- A cell phone manufacturing company might collect data about how often people buy new cell phones and what factors affect their choice, so that the cell phone company can focus on those features that would make their product more attractive to buyers.
- A town councillor might want to know how many accidents have occurred at a particular intersection, to decide whether a robot should be installed. The councillor would visit the local police station to research their records to collect the appropriate data.
- A supermarket manager might ask the question: “What flavours of soft drink should be stocked in my supermarket?" The question asked of customers might be “What is your favourite soft drink?” Based on the customers’ responses (i.e. the number of customers who liked soft drink A), the manager can make an informed decision as to what soft drinks to stock.

However, it is important to note that different questions reveal different features of a situation, and that this affects the ability to understand the situation. For example, if the first question in the list was re-phrased to be: "Does your home have electricity?" then if you answered yes, but you were getting your electricity from a neighbour, then this would give the wrong impression that you did not need an independent supply of electricity.

2.2 Methods of Data Collection

The method of collecting the data must be appropriate to the question being asked. Some examples of data collecting methods are:

1. Questionnaires, surveys and interviews
2. Experiments
3. Other sources (friends, family, newspapers, books, magazines and the Internet)

The most important aspect of each method of data collecting is to clearly formulate the question that is to be answered. The details of the data collection should therefore be structured to take your question into account.

For example, questionnaires, interviews or surveys would be most appropriate for the list of questions in "Purpose of Collecting Primary Data" (Section 2.1.2: Purpose of Collecting Primary Data).

2.3 Samples and Populations

Before the data collecting starts, it is important to decide how much data is needed to make sure that the results give an accurate reflection to the required answers. Ideally, the study should be designed to maximise the amount of information collected while minimising the effort. The concepts of *populations* and *samples* is vital to minimising effort.

The following terms should be familiar:

Population: describes the entire group under consideration in a study. For example, if you wanted to know how many learners in your school got the flu each winter, then your population would be all the learners in your school.

Sample: describes a group chosen to represent the population under consideration in a study. For example, for the survey on winter flu, you might select a sample of learners, maybe one from each class.

Random sample: describes a sample chosen from a population in such a way that each member of the population has an equal chance of being chosen.

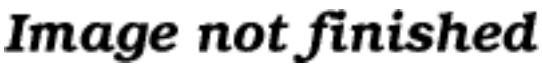

Figure 2

Choosing a representative sample is crucial to obtaining results that are unbiased. For example, if we wanted to determine whether peer pressure affects the decision to start smoking, then the results would be different if only boys were interviewed, compared to if only girls were interviewed, compared to both boys and girls being interviewed.

Therefore questions like: "How many interviews are needed?" and "How do I select the candidates for the interviews?" must be asked during the design stage of the sampling process.

The most accurate results are obtained if the entire population is sampled for the survey, but this is expensive and time-consuming. The next best method is to *randomly* select a sample of subjects for the interviews. This means that whatever the method used to select subjects for the interviews, each subject has an equal chance of being selected. There are various methods of doing this for example, names can be picked out of a hat or can be selected by using a random number generator. Most modern scientific calculators have a random number generator or you can find one on a spreadsheet program on a computer.

So, if you had a total population of 1 000 learners in your school and you randomly selected 100, then that would be the sample that is used to conduct your survey.

3 Example Data Sets

The remainder of this chapter deals with the mathematical details that are required to analyse the data collected.

The following are some example sets of data which can be used to apply the methods that are being explained.

3.1 Data Set 1: Tossing a Coin

A fair coin was tossed 100 times and the values on the top face were recorded. The data are recorded in "Data Set 1: Tossing a coin" (Table 1).

H	T	T	H	H	T	H	H	H	H
H	H	H	H	T	H	H	T	T	T
T	T	H	T	T	H	T	H	T	H
H	H	T	T	H	T	T	H	T	T
T	H	H	H	T	T	H	T	T	H
H	T	T	T	T	H	T	T	H	H
T	T	H	T	T	H	T	T	H	T
H	T	T	H	T	T	T	T	H	T
T	H	T	T	H	H	H	T	H	T
T	T	T	H	H	T	T	T	H	T

Table 1: Results of 100 tosses of a fair coin. H means that the coin landed heads-up and T means that the coin landed tails-up.

3.2 Data Set 2: Casting a die

A fair die was cast 100 times and the values on the top face were recorded. The data are recorded in "Data Set 2: Casting a die" (Section 3.2: Data Set 2: Casting a die).

3	5	3	6	2	6	6	5	5	6	6	4	2	1	5	3	2	4	5	4
1	4	3	2	6	6	4	6	2	6	5	1	5	1	2	4	4	2	4	4
4	2	6	4	5	4	3	5	5	4	6	1	1	4	6	6	4	5	3	5
2	6	3	2	4	5	3	2	2	6	3	4	3	2	6	4	5	2	1	5
5	4	1	3	1	3	5	1	3	6	5	3	4	3	4	5	1	2	1	2
1	3	2	3	6	3	1	6	3	6	6	1	4	5	2	2	6	3	5	3
1	1	6	4	5	1	6	5	3	2	6	2	3	2	5	6	3	5	5	6
2	6	6	3	5	4	1	4	5	1	4	1	3	4	3	6	2	4	3	6
6	1	1	2	4	5	2	5	3	4	3	4	5	3	3	3	1	1	4	3
5	2	1	4	2	5	2	2	1	5	4	5	1	5	3	2	2	5	1	1

Table 2: Results of 200 casts of a fair die.

3.3 Data Set 3: Mass of a Loaf of Bread

There are regulations in South Africa related to bread production to protect consumers. Here is an excerpt from a report about the legislation:

"The Trade Metrology Act requires that if a loaf of bread is not labelled, it must weigh 800g, with the leeway of five percent under or 10 percent over. However, an average of 10 loaves must be an exact match to the mass stipulated. - Sunday Tribune of 10 October 2004 on page 10"

We can use measurements to test if consumers getting value for money. An unlabelled loaf of bread should weigh 800g. The masses of 10 different loaves of bread were measured at a store for 1 week. The data are shown in Table 3.

Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
802.39	787.78	815.74	807.41	801.48	786.59	799.01
796.76	798.93	809.68	798.72	818.26	789.08	805.99
802.50	793.63	785.37	809.30	787.65	801.45	799.35
819.59	812.62	809.05	791.13	805.28	817.76	801.01
801.21	795.86	795.21	820.39	806.64	819.54	796.67
789.00	796.33	787.87	799.84	789.45	802.05	802.20
788.99	797.72	776.71	790.69	803.16	801.24	807.32
808.80	780.38	812.61	801.82	784.68	792.19	809.80
802.37	790.83	792.43	789.24	815.63	799.35	791.23
796.20	817.57	799.05	825.96	807.89	806.65	780.23

Table 3: Masses (in g) of 10 different loaves of bread, from the same manufacturer, measured at the same store over a period of 1 week.

3.4 Data Set 4: Global Temperature

The mean global temperature from 1861 to 1996 is listed in Table 4. The data, obtained from http://www.cgd.ucar.edu/stats/I , was converted to mean temperature in degrees Celsius.

Year	Temperature	Year	Temperature	Year	Temperature	Year	Temperature
1861	12.66	1901	12.871	1941	13.152	1981	13.228
1862	12.58	1902	12.726	1942	13.147	1982	13.145
1863	12.799	1903	12.647	1943	13.156	1983	13.332
1864	12.619	1904	12.601	1944	13.31	1984	13.107

continued on next page

[1] http://www.cgd.ucar.edu/stats/Data/Climate/

1865	12.825	1905	12.719	1945	13.153	1985	13.09
1866	12.881	1906	12.79	1946	13.015	1986	13.183
1867	12.781	1907	12.594	1947	13.006	1987	13.323
1868	12.853	1908	12.575	1948	13.015	1988	13.34
1869	12.787	1909	12.596	1949	13.005	1989	13.269
1870	12.752	1910	12.635	1950	12.898	1990	13.437
1871	12.733	1911	12.611	1951	13.044	1991	13.385
1872	12.857	1912	12.678	1952	13.113	1992	13.237
1873	12.802	1913	12.671	1953	13.192	1993	13.28
1874	12.68	1914	12.85	1954	12.944	1994	13.355
1875	12.669	1915	12.962	1955	12.935	1995	13.483
1876	12.687	1916	12.727	1956	12.836	1996	13.314
1877	12.957	1917	12.584	1957	13.139		
1878	13.092	1918	12.7	1958	13.208		
1879	12.796	1919	12.792	1959	13.133		
1880	12.811	1920	12.857	1960	13.094		
1881	12.845	1921	12.902	1961	13.124		
1882	12.864	1922	12.787	1962	13.129		
1883	12.783	1923	12.821	1963	13.16		
1884	12.73	1924	12.764	1964	12.868		
1885	12.754	1925	12.868	1965	12.935		
1886	12.826	1926	13.014	1966	13.035		
1887	12.723	1927	12.904	1967	13.031		
1888	12.783	1928	12.871	1968	13.004		
1889	12.922	1929	12.718	1969	13.117		
1890	12.703	1930	12.964	1970	13.064		
1891	12.767	1931	13.041	1971	12.903		
1892	12.671	1932	12.992	1972	13.031		

continued on next page

1893	12.631	1933	12.857	1973	13.175		
1894	12.709	1934	12.982	1974	12.912		
1895	12.728	1935	12.943	1975	12.975		
1896	12.93	1936	12.993	1976	12.869		
1897	12.936	1937	13.092	1977	13.148		
1898	12.759	1938	13.187	1978	13.057		
1899	12.874	1939	13.111	1979	13.154		
1900	12.959	1940	13.055	1980	13.195		

Table 4: Global temperature changes over the past 135 years. There has been a lot of discussion regarding changing weather patterns and a possible link to pollution and greenhouse gasses.

3.5 Data Set 5: Price of Petrol

The price of petrol in South Africa from August 1998 to July 2000 is shown in Table 5.

Date	Price (R/l)
August 1998	R 2.37
September 1998	R 2.38
October 1998	R 2.35
November 1998	R 2.29
December 1998	R 2.31
January 1999	R 2.25
February 1999	R 2.22
March 1999	R 2.25
April 1999	R 2.31
May 1999	R 2.49
June 1999	R 2.61
July 1999	R 2.61
August 1999	R 2.62
September 1999	R 2.75
October 1999	R 2.81
November 1999	R 2.86
December 1999	R 2.85
January 2000	R 2.86
February 2000	R 2.81
March 2000	R 2.89
April 2000	R 3.03
May 2000	R 3.18
June 2000	R 3.22
July 2000	R 3.36

Table 5: Petrol prices

4 Grouping Data

One of the first steps to processing a large set of raw data is to arrange the data values together into a smaller number of groups, and then count how many of each data value there are in each group. The groups are usually based on some sort of interval of data values, so data values that fall into a specific interval, would be grouped together. The grouped data is often presented graphically or in a frequency table. (Frequency means "how many times")

Exercise 1: Grouping Data ***(Solution on p. 10.)***

Group the elements of Data Set 1 (Table 1) to determine how many times the coin landed heads-up and how many times the coin landed tails-up.

4.1 Exercises - Grouping Data

1. The height of 30 learners are given below. Fill in the grouped data below. (Tally is a convenient way to count in 5's. We use llll to indicate 5.)

142	163	169	132	139	140	152	168	139	150
161	132	162	172	146	152	150	132	157	133
141	170	156	155	169	138	142	160	164	168

Table 6

Group	Tally	Frequency
$130 \leq h < 140$		
$140 \leq h < 150$		
$150 \leq h < 160$		
$160 \leq h < 170$		
$170 \leq h < 180$		

Table 7

Click here for the solution[2]

2. An experiment was conducted in class and 50 learners were asked to guess the number of sweets in a jar. The following guesses were recorded.

56	49	40	11	33	33	37	29	30	59
21	16	38	44	38	52	22	24	30	34
42	15	48	33	51	44	33	17	19	44
47	23	27	47	13	25	53	57	28	23
36	35	40	23	45	39	32	58	22	40

Table 8

Draw up a grouped frequency table using intervals 11-20, 21-30, 31-40, etc.
Click here for the solution[3]

[2] http://www.fhsst.org/l4k
[3] http://www.fhsst.org/l40

Solutions to Exercises in this Module

Solution to Exercise (p. 8)

Step 1. There are two unique data values: H and T. Therefore there are two groups, one for the H-data values and one for the T-data values.

Step 2.

Data Value	Frequency
H	44
T	56

Table 9

Step 3. There are 100 data values and the total of the frequency column is 44+56=100.