Statistics

Numerical Data Types

In statistics, numerical data from a population varies from one member of the population to another. The property of a member, which provides the data, is referred to as a **statistical variable**. A variable is either **discrete** or **continuous**. Discrete data is usually made up of integer values and an example of a discrete variable is the number of customers entering a shop in a 10-minute period. An example of a continuous variable is the lifetime of a light bulb.

With continuous data, it is extremely unlikely that 2 data values are equal whereas it is clear that it is reasonable to expect a repetition of a certain value of a discrete variable. For this reason, continuous data is usually classified into groups (e.g., time intervals) and discrete data is classified into groups only if there are very many possible individual values

Practical Numerical Methods

Data Grouping/Classification

Let us consider an attempt to obtain information on the age distribution in the population of Sheffield. Here age is defined as 'age at last birthday' and so it is a discrete variable.

Suppose that we have relevant data from a random sample of 2000 people. Then we have to devise a way to present the information clearly and concisely. It would not be misleading to simply list all 2000 values of the data but such a large amount of data could not be understood easily. Instead we would group the data in to specified age groups and then present this data although, in doing so, there is some **loss of information**. The following is a smaller scale example where the sample is of size 100.

66	70	10	54	62	13	11	15	69	26	49	11	3
67	10	54	42	32	56	39	60	79	33	12	47	24
19	47	63	32	7	70	55	46	11	20	15	39	37
28	72	46	64	61	51	56	53	61	11	80	53	28
76	6	5	39	58	29	52	54	47	60	62	51	72
41	57	32	12	33	17	40	20	10	27	47	71	68
44	7	23	17	81	23	12	33	16	46	71	48	58
79	80	43	31	72	68	36	41	11				

We can group these data in a **frequency table** as follows, where x represents the age.

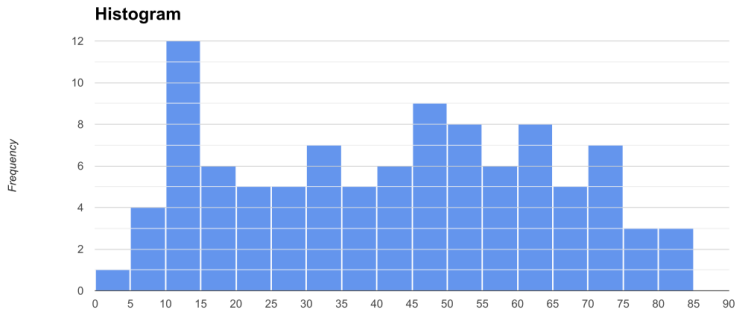
Group Number	**Age Range**	**Tally**	**Number in Group (Frequency)**
1	$0 \leq x < 5$		1
2	$5 \leq x < 10$		4
3	$10 \leq x < 15$		12
4	$15 \leq x < 20$		6
5	$20 \leq x < 25$		5
6	$25 \leq x < 30$		5
7	$30 \leq x < 35$		7
8	$35 \leq x < 40$		5
9	$40 \leq x < 45$		6
10	$45 \leq x < 50$		9
11	$50 \leq x < 55$		8
12	$55 \leq x < 60$		6
13	$60 \leq x < 65$		8
14	$65 \leq x < 70$		5
15	$70 \leq x < 75$		7
16	$75 \leq x < 80$		3
17	$80 \leq x < 85$		3
18	$85 \leq x < 90$		0
19	$90 \leq x < 95$		0
20	$95 \leq x < 100$		0
21	$100 \leq x < 105$		0
	Total:		

In this table we have **class widths** of 5 and **class boundaries** of 0, 5, 10, 15, ..., 100, 105. Each class boundary could be a **class upper bound** or a **class lower bound**. **Note:** The class $5 \leq x < 10$ contains 5.0000, 5.04, 5.2, 6.04, ..., 8.975, 9.9, 9.9999, , but *not* 10.0000. Other sources online or in textbooks may define bins so that they exclude the lower bound but include the upper bound. There appears to be no standard way of defining this.

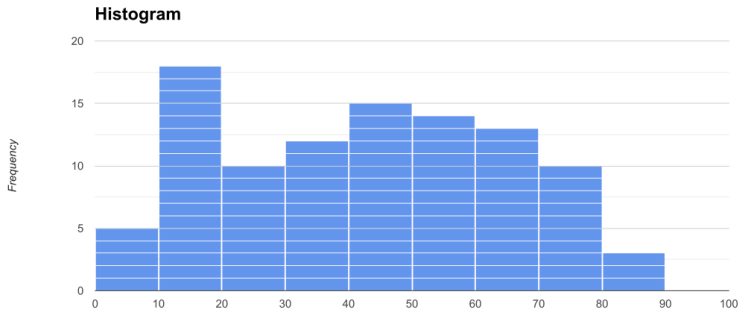
This **frequency table** is arguably much more informative than the original list of 100 values. If we had 2000 original data values then the frequency table could be of the same physical size with the same groupings but with generally higher frequency values.

For example, the table indicates that there appears to be a greater number of people in the 10-15 year age group than in any other single group. This impression may or may not be true. In addition, the table indicates that the random sample did not produce evidence of anyone older than 85 years of age. This doesn't necessarily mean that there is no one over the age of 85 living in Sheffield.

We can represent the data in a **histogram** with labelling directly related to the data. This is more appropriate than just producing a bar chart.



The third bar from the left indicates that, in our sample, there are 12 people whose age is at least 10 but less than 14. We could draw an alternative histogram where the bin sizes are of width 10 instead of 5, to produce the plot below.



More intricate histograms can be made in which bins of lower frequency are pooled together.

Data Summary

When we wish to compare two, or sometimes more, samples, it would be easier if we could extract a small number of numerical characteristics (each one a 'statistic') from the data in each sample. These numerical characteristics can then be compared. Here we look at just such a few numerical characteristics.

Measures of Central Tendency

These are also known as averages and measures of location.

Arithmetic Mean

Here we simply add up all of the data values and then divide by the number of data values.

Example 1 In a sample of five values (6, 9, 2, 4, 3) the mean value is:

$$\frac{6+9+2+4+3}{5} = \frac{24}{5} = 4.8.$$

Example 2 The one hundred values from the previous histogram example add up to 4165 and so the mean value is $\frac{4165}{100} = 41.65$.

The formula for the **arithmetic mean** (or just **mean**) is quite simple. Suppose that we have n data values x_1, x_2, x_3, ..., x_n. The mean, often written as $\bar{x}$ is calculated as:

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}.$$

If we had grouped data (k groups) with frequencies f_1, f_2, f_3, ..., f_k, corresponding to the **class mid-points** x_1, x_2, x_3, ..., x_k, then we would use the formula:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + x_3 f_3 + \ldots x_k f_k}{n}$$

where $n = f_1 + f_2 + f_3 + \ldots + f_k$. An example of this is given for the grouped data from the histogram example, where the last few rows have been ignored because of having frequency zero.

Age Range	**Class Mid-Point**	**Number in Group (Frequency)**	$x_i f_i$
$0 \le x < 5$	2.5	1	2.5
$5 \le x < 10$	7.5	4	30
$10 \le x < 15$	12.5	12	150
$15 \le x < 20$	17.5	6	105
$20 \le x < 25$	22.5	5	112.5
$25 \le x < 30$	27.5	5	137.5
$30 \le x < 35$	32.5	7	227.5
$35 \le x < 40$	37.5	5	187.5
$40 \le x < 45$	42.5	6	255
$45 \le x < 50$	47.5	9	427.5
$50 \le x < 55$	52.5	8	420
$55 \le x < 60$	57.5	6	345
$60 \le x < 65$	62.5	8	500
$65 \le x < 70$	67.5	5	337.5
$70 \le x < 75$	72.5	7	507.5
$75 \le x < 80$	77.5	3	232.5
$80 \le x < 85$	82.5	3	247.5
	Total:	100	4225

Hence according to this information the mean age is 42.25. Notice that the grouping of the data modifies the mean value obtained. This is because information is lost when the data are grouped. The mean value so calculated, is really only an estimate.

Median

Ungrouped Data

Here we arrange the data values in ascending (or descending) order. Then the **median** is:

1. the middle item if there is an odd number of data items;
2. the average of the middle two items if there is an even number of data items.

If for example the ordered data is 2, 3, 4, 6, 9, then the middle item is 4 and so the median is 4.

From our histogram example, the sample of one hundred data values, the ordered data is: The middle two values are 43 and 44, hence the median is 43.5.

3	5	6	7	7	10	10	10	11	11	11	11	11
12	12	12	13	15	15	16	17	17	19	20	20	23
23	24	26	27	28	28	29	31	32	32	32	33	33
33	36	37	39	39	39	40	41	41	42	*43*	*44*	46
46	46	47	47	47	47	48	49	51	51	52	53	53
54	54	54	55	56	56	57	58	58	60	60	61	61
62	62	63	64	66	67	68	68	69	70	70	71	71
72	72	72	76	79	79	80	80	81				

Mode

This is simply the data value which occurs most often in the sample. For grouped data we have a modal group (or modal class) or we can specify the mid-point of the modal class as the mode if we wish to give a single value.

For the sample of one hundred data values above, the most commonly occurring value is 11, hence the modal age is 11. The corresponding grouped data gives the modal class as $10 \leq x < 15$. In each of these measures the value can be found exactly if the data is not grouped, otherwise the value obtained is an estimate and depends upon how the data is grouped.

Dispersion or Spread

Range

This is simply the difference between the maximum value and the minimum value. This is easy to determine if the data is ordered (as for the median). For example, using the ordered data (100 values), we can easily read off the maximum and the minimum values as 81 and 3, respectively. Hence the range is $81 - 3 = 78$.

The main advantage of using this measure of dispersion is that it is easy to calculate. The main disadvantage of using this measure of dispersion is that it ignores any concentration of data near the centre of the distribution and that it is easily affected by outlying values.

Standard Deviation

This attempts to take into account the deviation of each data value from the mean value. The mean value is $\bar{x}$ and is calculated as above and n is the number of data values. The deviation of each data point x_i from the mean is calculated by finding the difference $x_i - \bar{x}$. This could be positive or negative and so we use the square of this, $(x_i - \bar{x})^2$, to make all of the deviations positive. Now we calculate the average of these 'squares of deviations' to get the **variance**:

$$\frac{\sum(x_i - \bar{x})^2}{n} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}.$$

To get the **standard deviation**, we take the square root of this value:

$$\sqrt{\frac{\sum(x_i - \bar{x})^2}{n}} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_n - \bar{x})^2}{n}}.$$

(The symbol $\sum$ just means add up all of the values.)

If our data points are $x_1 = 6$, $x_2 = 9$, $x_3 = 2$, $x_4 = 4$, and $x_5 = 3$, then we have already calculated the mean value as $\bar{x} = 4.8$. The variance is then given by

$$\begin{aligned}\frac{(6-4.8)^2 + (9-4.8)^2 + (2-4.8)^2 + (4-4.8)^2 + (3-4.8)^2}{5} &= \frac{1.44 + 17.64 + 7.84 + 0.64 + 3.24}{5} \\ &= 6.16.\end{aligned}$$

For a finite population, we can also use another formula to find the variance:

$$\frac{1}{n}\sum x_i^2 - \bar{x}^2.$$

So in our example this would be

$$\frac{6^2 + 9^2 + 2^2 + 4^2 + 3^2}{5} - 4.8^2 = 6.16.$$

The standard deviation is then $s = \sqrt{6.16} = 2.48$ (2 d.p.).

When we wish to compare two samples, we can compare their means values to see if one sample has generally lower values than the other sample. We can also compare the sample standard deviations to see if one samples has values which are generally more spread out than the values in the other sample. There are further techniques which can be investigated in order to see if the differences in samples statistics are **significant**. This falls under the realm of **hypothesis testing**, which we will not investigate in this module.