Lectures 9

Statistics

Statistics is the mathematical analysis of data associated with uncertain (probabilistic) events. So statistics is all about *data analysis.*

Data

Data when first collected is called *raw data.* It is then processed into tables or charts. E.g. marks out of 5 for 10 students are

0,3,5,2,3,4,5,1,2,3.

This is raw data. First make an *ordered list.*

0,1,2,2,3,3,4,5,5

Then form a *frequency table.* (The frequency is the number of times a result occurs.)

Mark	Frequency	Relative Frequency
0	1	0.1
1	1	0.1
2	2	0.2
3	3	0.3
4	1	0.1
5	2	0.2

The *relative frequency* $= \dfrac{\text{frequency}}{\text{total frequency}}$

Bar charts, pie charts, histograms

In a *bar chart* data is represented by rectangles or *bars* which may be horizontal or vertical. The *length* corresponds to the frequency.

In a *pie chart* a circular "pie" is divided into a number of portions, each *sector* representing a different category. E.g. in a survey 360 students were asked which subject they liked the most. 270 said "maths", 60 said "don't know" and 30 did not respond. The pie chart would then be

A *histogram* illustrates *continuous data* which has been grouped into classes. The range along the x-axis now corresponds to the size of the class. E.g.

Example: A group of students obtains the following marks:
0,3,2,1,4,3,12,7,5,9,10,4

Draw a bar chart showing the number of students gainingmarks in the ranges 0–4, 5–8, 9–12. What is the relative frequency of 3? Ordered list:

0,1,2,3,3,4,4,5,7,9,10,12

Frequency of 3 = 2.

Total number of students (i.e. total frequency) = 12.

Relative frequency of 3 = $\frac{2}{12}$ = 0.167.

Mean, mode and median

These are different kinds of average for a set of data.

The *mean* (or *arithmetic mean* is defined by

$$\text{Mean} = \frac{\text{Sum of values}}{\text{Total number of values}}.$$

Example: 10 students sit an exam with marks out of 5:

0,1,2,2,3,3,4,4,5,5

$$\begin{aligned} \text{Sum} =& 29 \\ \text{Total number of values} =& 10 \\ \Rightarrow \text{Mean} =& \frac{29}{10} = 2.9 \end{aligned}$$

More generally, suppose we have n values, $x_1, x_2, \ldots x_n$. The mean is denoted $\overline{x}$. Then

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

E.g. in the example, $x_1 = 0$, $x_2 = 1$, $x_3 = 2\ldots$, and $n = 10$.

The *mode* is the value that occurs most often. E.g. find the mode of the set of numbers
1 1 4 4 5 6 8 8 8 9
8 occurs most often so the mode is 8.

The *median* of a set of numbers is found by forming an ordered list of the numbers (in ascending order) and then selecting the value that lies half-way along the list.

N.B. for an even set of numbers, the median is $\frac{1}{2}$ the sum of the two numbers "either side" of the halfway point in the list.

E.g. find the median of the numbers
1 2 6 7 9 11 11 11 14 (odd set of numbers)
Median is 9.
But if list is
1 1 2 6 7 9 11 11 11 14 (even set of numbers)

$$\text{Median} = \frac{7+9}{2} = 8.$$

Standard Deviation

Standard deviation quantifies the variation of a set of data values. E.g. consider the two sets of data 4,4,4 and 2,3,7. Their respective means are $\frac{4+4+4}{3} = 4$ and $\frac{2+3+7}{3} = 4$. They have the same mean, but the data in the second set are widely spread while the data in the second set are all the same. These will have a bigger standard deviation.

Definition Suppose we have a set of n values $x_1, x_2 \ldots, x_n$. The mean value is $\overline{x}$. Then the standard deviation is

$$\text{Standard Deviation} = \sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}}.$$

So the standard deviation is a measure of the *spread* (or *dispersion*) of a set of numbers about the mean. The standard deviation is often written as σ.

Example: For 2,3,7 $\overline{x} = 4$, $n = 3$. $x_1 = 2$, $x_2 = 3$, $x_3 = 7$.

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = (2-4)^2 + (3-4)^2 + (7-4)^2 = (-2)^2 + (-1)^2 + (3)^2 = 14.$$

$$\Rightarrow \sigma = \sqrt{\frac{14}{3}} = 2.16.$$

For 4,4,4, $x_1 = x_2 = x_3 = 4$.

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 = 0 + 0 + 0 = 0 \Rightarrow \sigma = \sqrt{\frac{0}{3}} = 0.$$

Example: Find the mean and standard deviation of

(i) 0,0,0,0

(ii) 1,-1,1,-1.

(i) Mean=0, standard deviation=0.
(ii) Mean=0,

$$\sigma = \sqrt{\frac{1^2 + (-1)^2 + 1^2 + (-1)^2}{4}} = 1.$$

Example: Which of the following sets of data, both of which have the same mean, has the largest spread of values about the mean?
(i) 1,2,3,4,5,7,9,9
(ii) 1,2,2,4,5,8,9,9
$n = 8$, $\overline{x} = 5$.
(i)

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \overline{x})^2 =&(-4)^2 + (-3)^2 + (-2)^2 + (-1)^2 + 0^2 + 2^2 + 4^2 + 4^2\\ =&66\\ \Rightarrow \sigma =&\sqrt{\frac{66}{8}} = 2.87.\end{aligned}$$

(ii)

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 =(-4)^2 + (-3)^2 + (-3)^2 + (-1)^2 + 0^2 + 3^2 + 4^2 + 4^2$$

$$\Rightarrow \sigma = \sqrt{\frac{76}{8}} = 3.08.$$

So (ii) has the greater spread (could see by inspection of the list). For larger n it's more difficult to see so we need to compute σ.