

Chapter 6 Statistics

6.1 Description of data

One of the common uses of statistics is to give a concise description of data. Take for example the collection of exam scores in a mathematics course: 93, 83, 43, 74, 74, 59, 81, 82, 84, 50, 74, 70, 81, 80, 95, 60, 66, 87, 69, 22, 106, 107, 77, 60, 37, 83, 77, 89, 80, 73, 80, 75, 69, 74, 77, 73, 66, 108, 53, 80, 72.

The most common description of a collection of such scores is to give the “average” or “mean” of these scores. Symbolically, if we let

$$\begin{aligned} x_1 &= \text{the 1st number (93)} \\ x_2 &= \text{the 2nd number (83)} \\ x_3 &= \text{the 3rd number (43)} \\ x_4 &= \text{the 4th number (74)} \\ &\dots \\ &\dots \\ &\dots \end{aligned}$$

$$x_{41} = \text{the 41st number (72)}$$

the mean \bar{x} of the scores is

$$\bar{x} = \frac{\sum_{k=1}^{41} x_k}{41} = \frac{3043}{41} \approx 74.22$$

The symbol $\sum_{k=1}^{41} x_k$ means the sum of the x_k 's from $k=1$ to $k=41$.

In general, the mean \bar{x} of the collection of numbers $x_1, x_2, x_3, \dots, x_n$ is given by

$$\bar{x} = \frac{\sum_{k=1}^n x_k}{n} = \frac{1}{n} \sum_{k=1}^n x_k$$

A surprising fact is that this one number, the average or mean, is often taken as the complete description of the data. Thus, for example, we hear people saying that the average achievement test score of one school is higher than the average achievement test

score of another school, implying that the students of the first school are better (better educated or smarter) than the students of the second school. Surely the average is an important characteristic of a collection of data, but it just does not tell the whole story. We need at least one more characteristic to get a better idea about the data. This characteristic is the standard deviation associated with the collection defined as follows:

If $x_1, x_2, x_3, \dots, x_n$ are numbers with the mean \bar{x} , the **standard deviation s** is defined by

$$s = \sqrt{\frac{\sum_{k=1}^n (x_k - \bar{x})^2}{n}}$$

To see the significance of the standard deviation, compute the standard deviations of the data in the following exercises. Note that the collections of the data have the same mean 45.

Exercises:

Compute the standard deviations of the collections:

(a) 20, 30, 40, 50, 60, 70

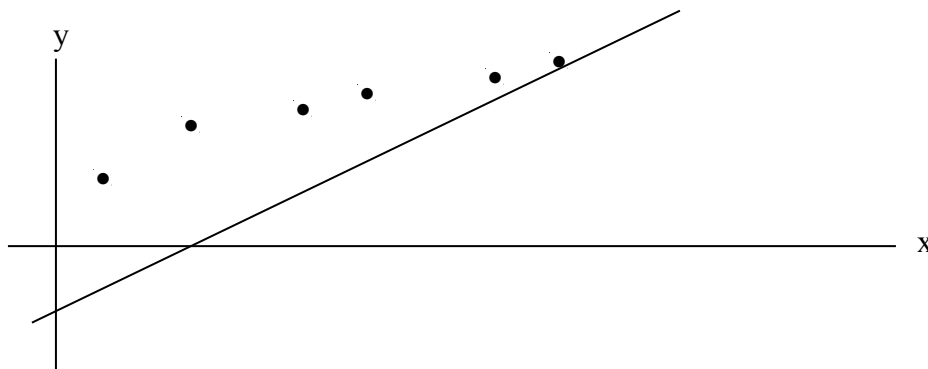
(b) 45, 45, 45, 45, 45, 45

(c) 20, 20, 20, 70, 70, 70

6.2 Linear regression

Another aspect with which the study of statistics is concerned is prediction. This takes several forms. For example, we can take a sample of a population and by analyzing the composition of the sample, we predict what the composition of the population is. The Gallup Poll is an example of this type. We can also study the present and the past trends and predict what will happen in the future. This is the kind of thing that is done in sciences, business, and industry. We will look at an example of this type.

The general problem is this: We have some data, $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots, (x_n, y_n)$, relating two quantities x and y , that seem to show some linear relationship when we plot the points on a graph paper or on a computer screen with some software. We want to fit a straight line so that it is “closest” to all the points.



More precisely, we want to determine m and b in the equation $y = mx + b$ so that the sum of the squares of the deviations

$$\sum_{k=1}^n (y_k - mx_k - b)^2$$

is smallest. The m and b that minimize the deviation $\sum_{k=1}^n (y_k - mx_k - b)^2$ are given by the following formulas:

$$m = \frac{\sum_{k=1}^n x_k y_k - n(\bar{x})(\bar{y})}{\sum_{k=1}^n (x_k)^2 - n(\bar{x})^2}$$

$$b = \bar{y} - m(\bar{x})$$

Here \bar{x} is the mean of $x_1, x_2, x_3, \dots, x_n$, and \bar{y} is the mean of $y_1, y_2, y_3, \dots, y_n$, and n is the number of the data.

The formula for m can be obtained by first bringing the deviation $\sum_{k=1}^n (y_k - mx_k - b)^2$ to the form $Am^2 + Bm + C$ and completing the square with respect to m , just as we have done in the last chapter. But since the computation is too messy, we will not give it here.

In any case, once we compute m and b by the above formulas, we substitute them into the equation $y = mx + b$, and get the equation of the **best fitting line** for the data. The line is also called the **regression line**.

Exercise 6.2

When sodium carbonate (a relative of the baking powder) is dissolved in water, the gas carbon dioxide is generated. The amount of gas generated is measured by measuring the pressure exerted by the gas. The more the sodium carbonate there is, the more the gas is generated and greater the pressure exerted by the gas. We want to determine the relationship between the amount of the sodium carbonate and the pressure exerted by the resulting gas. The following is an experimental result of a Chem 151 student:

W (in grams)	P (in cm of mercury)
0.000	0.40
0.100	0.90
0.250	2.15
0.400	2.90
0.550	4.00
0.700	4.90

- Plot the points on a graph paper or get a scatter diagram on a computer.
- Find the equation of the best fitting line for the data.
- An unknown sample is known to contain sodium carbonate and no other gas producing substance. When the sample is dissolved in water and the pressure exerted by the resulting gas was measured, the pressure turned out to be 4.25 cm of

mercury. Find the amount of sodium carbonate in the sample.