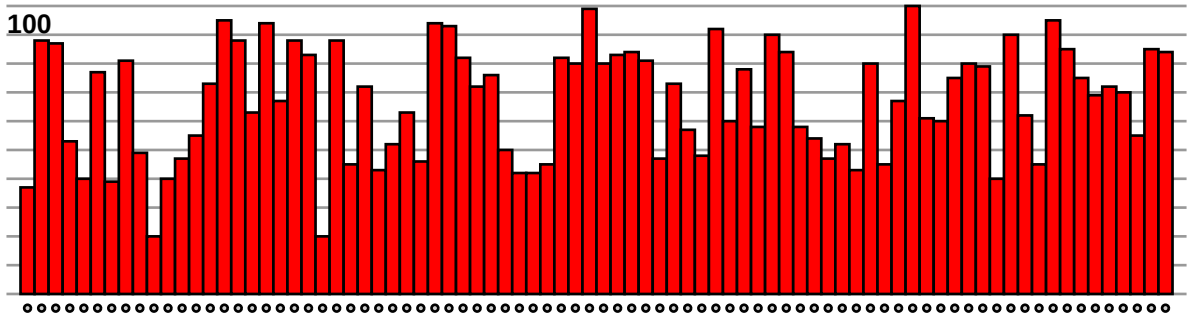
Mathematics 103 — Spring 2001

Averages and areas

In this note we shall look at ways of describing data, and in particular at various notions of **average**.

Here is a bar graph of the mid-term examination scores of a calculus section:



It happens that the scores are sorted alphabetically by the names of students. In other words, there isn't much intrinsic significance to the x-coordinate in this graph.

The basic problem of statistics is this:

- *How can you summarize and even comprehend such an apparent messy collection of data?*

This is not a simple problem. The simplest technique often applied is to calculate the **average** score, but there are in fact several different notions of average. We shall look at two of them here, the **arithmetic mean** (usually called just the mean) and the **median**. Both involve areas and, ultimately, integrals.

The mean

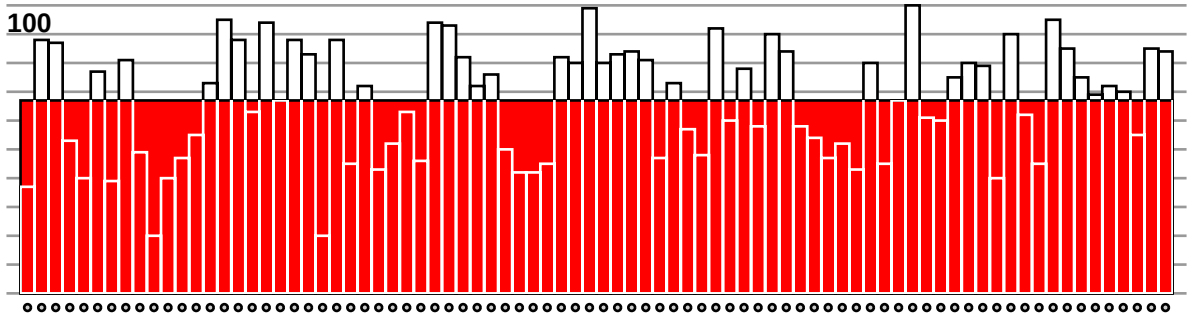
The arithmetic mean of N numbers s_1, s_2, ..., s_N is their sum divided by their number:

$$\overline{s} = \frac{s_1 + s_2 + \cdots + s_N}{N} \ .$$

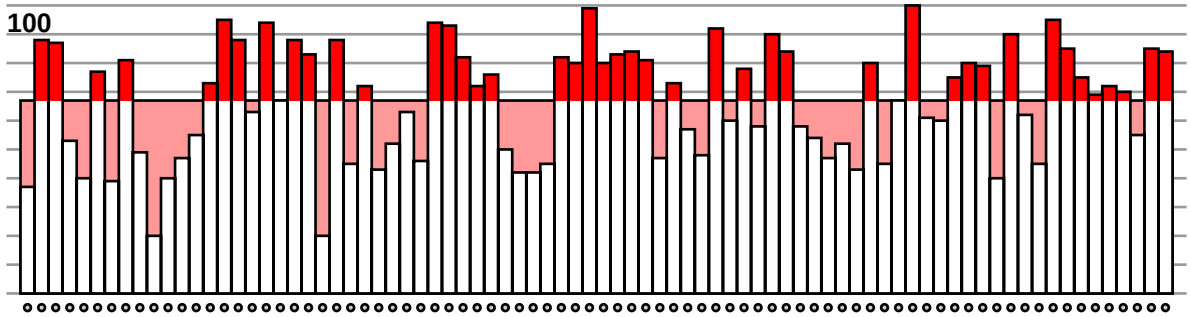
In terms of the original bar graph, there is a simple interpretation. The defining equation for $\overline{s}$ is equivalent to the equation

$$N\overline{s} = s_1 + s_2 + \cdots + s_N \ .$$

which means that if we replace each score in the bar graph by $\overline{s}$, the area of the new graph (which is a rectangle) will have the same area as the old one:



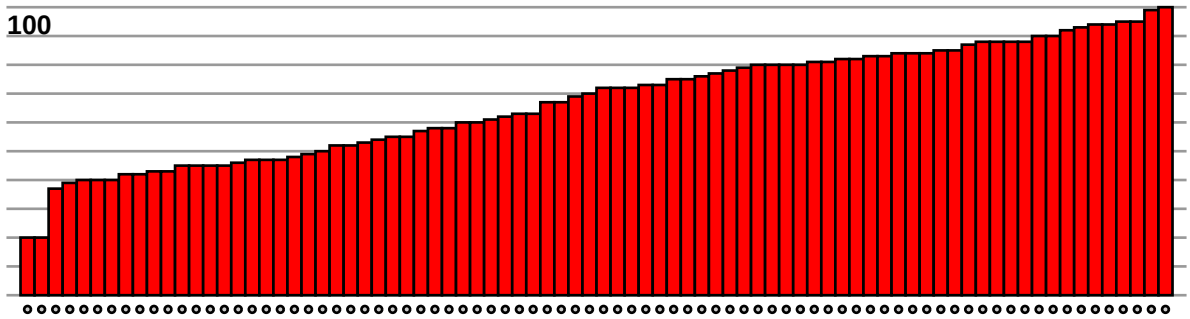
Another way to put this is that in the following graph the area of medium shade is equal to that of the darker one. Intuitively, we obtain the mean line by leveling off the original graph without removing any of it, so to speak. A process not unlike leveling a pile of dirt with a bulldozer.



The intuitive significance of the arithmetic mean is not so clear. About all we can say for sure is that it lies in the middle of the data, in that it lies between the highest value and the lowest one, and will be strictly between them if the two are different. How much information it contains seems to depend on the kind of data we are dealing with. For example, the mean cost of items in a bag of groceries does usually some kind of feeling for the relative cost of the stuff. The average price of houses sold in a year in Vancouver will give you some idea of what the market is like. But for the scores of a class on an examination, what someone usually wants to know is his relative placement in the class. The mean has nothing to say about this. It could happen, for example, that one student does far better than all the others. This might raise the mean considerably without affecting the ranking of scores. Similarly, if you bought 10 packages of potato chips and one bottle of fancy perfume, the average price of a bag of groceries won't really give you much of an idea of average cost.

Sorting the data

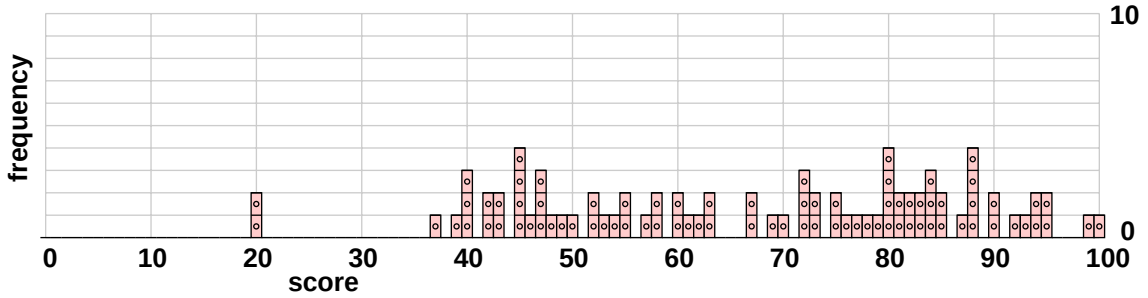
We can get a much better idea of how people did on the examination by sorting it according to score. If we do this and then make up a bar graph, we get this:



I think it is fair to say that this tell us much more. For example, a student with a grade of, say, 40 can see immediately how he did compared to the rest of the class—better than about 75% and worse than about 25%. The **median** grade is the one that lies half-way in this list—that is to say, half the students' scores are higher than the median and half are less. Of course if there is an odd number of students then we may have to modify this slightly, and if there were a gap in the middle, the median could be any of several possible scores, so that in general it is a **range of values**.

The frequency distribution

Even the bar graph of sorted data is not quite transparent to read. We can improve it by counting for each possible score n the number of people $f(n)$ with score n. We can make these data into a bar graph also, where now the x-axis is the range of possible scores, and y records the count for each value of x.



In this graph, the x-axis represents what the y-axis did in the previous pictures—namely, examination scores.

In many ways, this is easier to read than any of the previous ones. We can see, for example, that examination scores are rather evenly distributed in the upper range, with a slight clustering around 80.

How do we calculate the mean from this graph? The total number of students taking the exam is

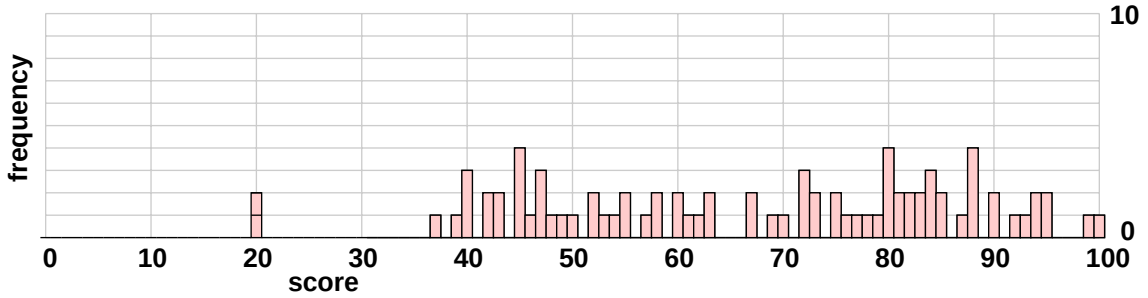
$$\sum_{x=0}^{x=100} f(x)$$

and the sum of the scores of all students can be rearranged so equal scores are grouped together. A given score x occurs in this sum $f(x)$ times. So the mean score is

$$\overline{x} = \frac{\sum_{x=0}^{x=100} xf(x)}{\sum_{x=0}^{x=50} f(x)} .$$

Each individual in this graph is now represented by a square in a bar. The number of individuals are represented by areas. For example, the total number of students taking the exam is the total area of the bar graph.

The number of students with, say, score at most 30 is equal to the area on the left:



How do we calculate the median from this graph? The median score of x is the value in the middle, in the sense that the area under the graph to the right of it is equal to the area to the left of it.