AN ELEMENTARY INTRODUCTION TO DATA AND STATISTICS

PIETRO POGGI-CORRADINI

ABSTRACT. The first part of this material should be accessible to a fourth grader, the latter part to a middle-schooler.

For us a distribution, or simply the data, is a finite string of numbers, such as

$$\{49, 50, 52, 53, 54\}$$

These numbers could represent the weights in kilograms of five kids, or the monetary amounts in the pockets of five friends. The specific context only comes into play when interpreting the statistics. Statistics are computations that reveal some information about the data, but also "forget" about most of the distinguishing features of a given distribution.

For instance, the mean, or average of the distribution is computed by adding up all the data into a sum

$$S = \text{ the sum of the data}$$

and then dividing it up equally among the number of data.

Some algebra helps clarify the phrasing here. The first piece of data is usually represented by the letter x_1, the second by x_2, etc...and the last one by x_n. This makes it clear that we are dealing with n numbers. For instance, if we are dealing with the numbers $\{49, 50, 52, 53, 54\}$ then we would think of $x_1 = 49, x_2 = 50, x_3 = 52, x_4 = 53, x_5 = 54$. The number of data n here is equal to 5 and the sum S is....something. The average is usually written as $\bar{x}$ and in this case it would be computed by dividing S by 5.

The task of adding up all the data seems daunting at first, but here is a trick that allows you to do the calculation "mentally".

- First **make a guess**. If you haven't learned about negative numbers, then always make your guess to be the smallest data. In our example, your guess would then be 49. The notation for the guess is y.
- Now **subtract off** your initial guess from each of the data and get a new set of data that will be easier to manage. In our example with guess $y = 49$, the new distribution is $\{0, 1, 3, 4, 5\}$. We give a name to this new data, we will call them the errors and write $e_1, \ldots, e_5$.
- Now we **average the errors**. In our example it is much easier now to add up the errors and get $0 + 1 + 3 + 4 + 5 = 13$. Even though e_1 is zero it stills count as a piece

of data, so we still must divide by 5. Hence we get that the errors average to 13/5, or $2\frac{3}{5}$.

- Finally **add the averaged errors to your initial guess**, and voila'! For us:

$$49 + 2\frac{3}{5} = 51\frac{3}{5}.$$

You can practice this strategy on a few more examples and then you will be able to impress friends and family with amazing mental skills. Try and ask your mom to give you four numbers between 1350 and 1400. She'll say something like: $1357, 1366, 1381, 1372$. Give her a calculator and ask her to average these numbers (she should be able to do that) but tell her not to tell you the answer. Now guess a number roughly in between, say 1370. Subtract it off and get $-13, -4, 11, 2$. These add up to -4. What luck! $-4 \div 4 = -1$. So the average is $1370 - 1 = 1369$!

Why does this trick work? Why does it work no matter what your initial guess is? The best way to explain this is using some algebra. Luckily we've already set up all the necessary notation. Suppose you want to average data $x_1, \ldots, x_n$. Make a guess y. Subtract it off and get a new data set of errors $e_1 = x_1 - y, \ldots, e_n = x_n - y$. Now average the errors as such

$$\bar{e} = \frac{(x_1 - y) + \cdots + (x_n - y)}{n}$$

Getting rid of the parenthesis and rearranging we find that

$$\bar{e} = \frac{(x_1 + \cdots + x_n) - ny}{n} = \frac{x_1 + \cdots + x_n}{n} - \frac{ny}{n} = \bar{x} - y.$$

So when we add $\bar{e}$ to our initial guess y we see that y cancels out:

$$\bar{e} + y = (\bar{x} - y) + y = \bar{x}.$$

Is it possible to guess the average right the first time? Yes of course. In that case $y = \bar{x}$ and when you go and average the errors you find that $\bar{e} = \bar{x} - y = 0$. In fact this poperty characterizes the mean: *$\bar{x}$ is the only number that makes all the (signed) errors add up to* 0.

This property of $\bar{x}$ explains the physical intuition that is often given for the mean. Think of $49, 50, 52, 53, 54$ as the places where unit weights are lying on a thin tray. Then try to place a wedge ('fulcrum') under the tray pointing at some point with coordinate y. The tray will balance only when all the signed distances to y add up to zero, namely when $y = \bar{x}$. Otherwise, the tray will crash to the floor.

Let's go back to the interpretation of the mean in specific examples. When $49, 50, 52, 53, 54$ represent amounts of money, then the mean $51\frac{3}{5}$ (51 dollars and 60 cents) represents what everyone would end up with if we tried to redistribute the money, "level the playing field", in such a way that everyone has the same amount. That amount is the mean. Clearly this interpretation fails if $49, 50, 52, 53, 54$ were representing heights instead. No way could we redistribute heights.

So in applications the interpretation of the mean must vary from context to context, and some times the information that is lost from the data when computing the mean might overshadow whatever "statistic" is obtained. Unfortunately, in politics and in the social sciences, too often, the error is made of speaking as if the mean is representing everything one would want to know about a specific data set.

Of course statisticians have a partial answer to this problem. If it's true that two quite different data sets may share the same average (thus losing all the information that distinguishes the two data sets), we can come up with a way of measuring how "dispersed" a data set may be around its average. This is a "second order" analysis. Let's consider again our friends, the errors $e_1 = x_1 - \bar{x}, \ldots, e_n = x_n - \bar{x}$. We know that they add up to zero, but if we remove the signs and consider instead $|e_1|, \ldots, |e_n|$, i.e. the distances of each piece of data from the average, how do they behave? What is their average? In words that would be "the average distance from the average". There is a term for this quantity, it's called the mean deviation:

$$\text{MD} = \frac{|x_1 - \bar{x}| + \cdots + |x_n - \bar{x}|}{n}.$$

Mathematicians, turns out, are not satified with this. Instead of just removing the sign of the errors $e_1, \ldots, e_n$, we'd rather do it simply by "squaring" the errors. So instead of the mean deviation, we prefer to compute the variance:

$$\text{Var} = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

and then, to make amends, we take the square root of the variance and call that the standard deviation.

In words, the standard deviation is "the square root of the average square-distance from the average". Why on earth would one want to square errors? There are many deep reasons for this and appealing to a vague resemblance to the Pythagorean Theorem would go a long way in explaining this. Instead let me give you an idea of why the variance is better, by doing a simple calculation.

What happens if we make an initial guess y which turns out not to be the right guess: $y \neq \bar{x}$, and then we happily go ahead and start averaging the square distances to y instead? What can we say about the number we would end up computing in relation to the variance? It turns out, that no matter what our initial guess is, we would always get something larger than the variance. In other words, we get another characterization of the mean: *$\bar{x}$ is the unique value of y that minimizes the sum of the square errors* $(x_1 - y)^2 + \cdots + (x_n - y)^2$.

To see this, let's focus on one term of the sum at a time, say the first one. We want to compare $(x_1 - y)^2$ to $(x_1 - \bar{x})^2$. Let's take the difference! Then we can maybe use the

remarkable identity

$$a^2 - b^2 = (a - b)(a + b).$$

(this can be checked simply by unfolding the left hand-side).

We get

$$\begin{aligned}(x_1 - y)^2 - (x_1 - \bar{x})^2 &= [(x_1 - y) - (x_1 - \bar{x})][(x_1 - y) + (x_1 - \bar{x})] \\ &= [\bar{x} - y][2x_1 - y - \bar{x}]\end{aligned}$$

The same exact computation holds with x_1 replaced by any other x_j. So adding these identities up and factoring out the common term $[\bar{x} - y]$, we get

$$\begin{aligned}&((x_1 - y)^2 + \cdots + (x_n - y)^2) - ((x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2) \\ &= [(x_1 - y)^2 - (x_1 - \bar{x})^2] + \cdots + [(x_n - y)^2 - (x_n - \bar{x})^2)] \\ &= [\bar{x} - y][2x_1 - y - \bar{x}] + \cdots + [\bar{x} - y][2x_n - y - \bar{x}] \\ &= [\bar{x} - y][2(x_1 + \cdots + x_n) - ny - n\bar{x}] \\ &= [\bar{x} - y][2n\bar{x} - ny - n\bar{x}] \\ &= n(\bar{x} - y)^2 > 0.\end{aligned}$$

where I used the fact that $x_1 + \cdots + x_n = n\bar{x}$.

What this shows is that if we go ahead and compute the average square error having made a guess y, we always get a larger quantity than the variance and in fact we overshoot exactly by $(\bar{x} - y)^2$. The magic of squares!

Department of Mathematics, Cardwell Hall, Kansas State University, Manhattan, KS 66506, USA

E-mail address: pietro@math.ksu.edu