Massachusetts Institute of Technology
Physics Department

8.01X **Fall 2000**

Averages and Standard Deviations: Sampling a Population

By a population we mean a group of objects or events which we wish to examine in terms of a quantifiable characteristic. For example, if our population is a group of people, we can consider height, age, or political affiliation: for a series of throws of two dice we can examine the product or sum of the two numbers: for baseballs we can measure coefficients of restitution (*i.e.*, the ratios of the speeds, with respect to the bat, before and after contact). We would not, however, wish to include the age of the dice in our sample of people, or the product of digits of the people's social security numbers in the samples of dice throws.

Of course, we can't always sample the entire population. Once every ten years we try to find out what we can about all the people in the country, but it makes no sense to try to measure "every throw of two dice" or speak of "home runs not hit in one season".

For a population of N elements, denote the random variables measured as x, so we have N values of x, denoted x_i, $i = 1$ to N, or just x_i. For our purposes we will consider

i) what the average of the population is (the *mean*), and

ii) how close a measurement of a random element of the sample is likely to be from the mean (the *standard deviation*).

The mean, denoted as $\langle x \rangle$, is simply the arithmetic average of the sample:

$$\langle x \rangle = \frac{1}{N} \sum_i x_i$$

This is simple enough and very likely quite familiar, but this definition also has the property that the *variance* with respect to $\langle x \rangle$ is a minimum. That is, let the variance with respect to a value c be

$$V_c = \frac{1}{N} \sum_i (x_i - c)^2.$$

Then, V_c clearly is positive ($V_c = 0$, which can only occur when all of the measurements are the same, will not be considered) and depends on c. To minimize

V_c, note that

$$\begin{aligned} V_c &= \frac{1}{N}\sum_i (x_i^2 - 2cx_i + c^2) \\ &= \frac{1}{N}\sum_i x_i^2 - 2c\langle x\rangle + c^2 \\ &= \frac{1}{N}\sum_i x_i^2 - \langle x\rangle^2 + \left(c^2 - 2c\langle x\rangle + \langle x\rangle^2\right) \\ &= \frac{1}{N}\sum_i x_i^2 - \langle x\rangle^2 + (c - \langle x\rangle)^2 \end{aligned}$$

The value of c that minimizes this epression is that which makes the term in parantheses zero. If one chooses to use calculus,

$$\begin{aligned} \frac{dV_c}{dc} &= -\frac{1}{N}\sum_i 2(x_i - c) = -\frac{2}{N}\left(\sum_i x_i - \sum_i c\right) \\ &= -\frac{2}{N}(N\langle x\rangle - Nc) = 0. \end{aligned}$$

Either way, V_c is minimized at $c = \langle x\rangle$. In the future, this least variance will be denoted as σ^2, and will be called "the variance" without qualification. Explicitly,

$$\begin{aligned} \sigma^2 &= \frac{1}{N}\sum_i (x_i - \langle x\rangle)^2 = \frac{1}{N}\sum_i \left(x_i^2 - 2\langle x\rangle x_i + \langle x\rangle^2\right) \\ &= \frac{1}{N}\left(\sum_i x_i^2 - 2\langle x\rangle\sum_i x_i + N\langle x\rangle^2\right) \\ &= \frac{1}{N}\sum_i x_i^2 - 2\langle x\rangle\langle x\rangle + \langle x\rangle^2 = \langle x^2\rangle - \langle x\rangle^2. \end{aligned}$$

This convenient algebraic expression allows us to calculate only two averages, the average square $\langle x^2\rangle$ and the mean $\langle x\rangle$, instead of having to calculate $\langle x\rangle$, then recalculate $(x_i - \langle x\rangle)$ N times, then squaring and averaging. So, as a result,

$$\sigma = \sqrt{\langle x^2\rangle - \langle x\rangle^2};$$

σ is known as the "root mean square of the deviation" of the population, that is, the square root of the average square of the distance from the mean. (In many situations, σ is known as the "population standard deviation".)

There's one slight catch. How do we know the $\langle x\rangle$ measured from our N trials is the true mean? Actually, sometimes we know it can't be. For example, if we toss a fair coin N times, assigning $x_i = 0$ if toss i is heads and $x_i = 1$ if toss i is tails,

we know that the true mean is 1/2. If N is odd, however, we can't possibly get $\langle x \rangle = 1/2$.

To account for this, we use as the standard deviation a revised version of the square root of the variance;

$$s^2 = \frac{1}{N-1} \sum_i (x_i - \langle x \rangle)^2 = \frac{N}{N-1} \sigma^2$$

$$s = \sqrt{\frac{N}{N-1}}\, \sigma.$$

The factor of $N-1$ instead of N is taken to mean that "one measurement is needed to find the mean, so $N-1$ are left to find the standard deviation". Another way to see the necessity of this factor is to realize that one measurement tells us nothing about the standard deviation. Often, s is known as the "sample standard deviation", both to indicate that s will depend on which elements are sampled, and to distinguish s from σ. For our purposes, the term "standard deviation", without qualification, will refer to s.

If $N \gg 1$, s is essentially the same as σ. If, however, we know the mean in advance, all we are measuring is the standard deviation, so we use σ. For example, with two honest dice, we know that the average of the sum of the numbers is seven (2 is as probable as 12, 3 as likely as 11, etc).

Most handheld scientific calculators with statistics functions have keys for both the sqaure root of the variance and for the standard deviation, and the symbols used to denote these quantities is not universal; if you use a handheld, you should check to see which keys give which values.

We've alluded to the fact that we can't expect to measure the mean precisely. How close can we reasonably expect to be? This sounds like a standard deviation and it is; it's known as the 'standard deviation of the mean', or $(SDM)_x$. To find the (SDM), we assume that the deviation of each x_i is precisely σ, but each x_i contributes one part in N to the variance, so

$$(SDM)_x^2 = \frac{1}{N-1} \sum_i \frac{\sigma^2}{N} = \frac{\sigma^2}{(N-1)} = \frac{s^2}{N}$$

$$(SDM)_x = \frac{s}{\sqrt{N}}.$$

To summarize:

- Mean = $\langle x \rangle$ = estimate of mean of population based on a sample size N.
- Standard deviation = s = estimate of spread of the population about the mean.
- Standard deviation of the mean = (SDM) = estimate of how close the mean of the sample is to the population mean.