

## Physics 250

**A summary of some important, and often poorly understood, results concerning the standard deviation of the distribution  $\sigma$ , the standard deviation of the sample  $\sigma_{\text{samp}}$ , and the error bar on the mean,  $\sigma_{\text{mean}}$ , of a sample of data.**

Peter Young

(Dated: December 3, 2007)

Suppose we have a set of experimental data,  $x_i$ , ( $i = 1, \dots, N$ ), which has some random noise. We shall often refer to this as a *sample* of data. The values of the  $x_i$  are governed by a distribution  $P(x)$ , *which we don't know*. This distribution has a mean  $\langle x \rangle$ , and a variance  $\sigma^2$ . (The term “standard deviation” is used for  $\sigma$ , the square root of the variance.) We denote an average *over the exact distribution* by angular brackets, e.g.

$$\langle x \rangle = \int x P(x) dx. \quad (1a)$$

$$\sigma^2 \equiv \langle (x - \langle x \rangle)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2 = \int x^2 P(x) dx - \left( \int x P(x) dx \right)^2. \quad (1b)$$

Our goal is to determine  $\langle x \rangle$ , and the uncertainty in our estimate of it, from the  $N$  data points  $x_i$ . In order to do this we will assume that the data are uncorrelated with each other. This is a crucial assumption, without which it is very difficult to proceed. However, it is not always clear if the data points are truly independent of each other; some correlations may be present but not immediately obvious. Here, we take the usual approach of assuming that even if there are some correlations, they are sufficiently weak to not significantly perturb the results of the analysis.

The information *from the data* is usefully encoded in two parameters, the sample mean  $\bar{x}$  and the sample standard deviation  $\sigma_{\text{samp}}$  which are defined by

$$\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i, \quad (2a)$$

$$\sigma_{\text{samp}}^2 \equiv \overline{(x - \bar{x})^2} = \overline{x^2} - (\bar{x})^2 = \frac{1}{N} \sum_{i=1}^N x_i^2 - \left( \frac{1}{N} \sum_{i=1}^N x_i \right)^2. \quad (2b)$$

In statistics, notation is often confusing but crucial to understand. Here, an average indicated by an over-bar,  $\overline{\dots}$ , is an average over the *sample of  $N$  data points*. This is to be distinguished from an exact average over the distribution  $\langle \dots \rangle$ , as in Eqs. (1a) and (1b). The latter is, however, just a theoretical construct since we *don't know* the distribution  $P(x)$ , only the set of  $N$  data points  $x_i$  which have been sampled from it.

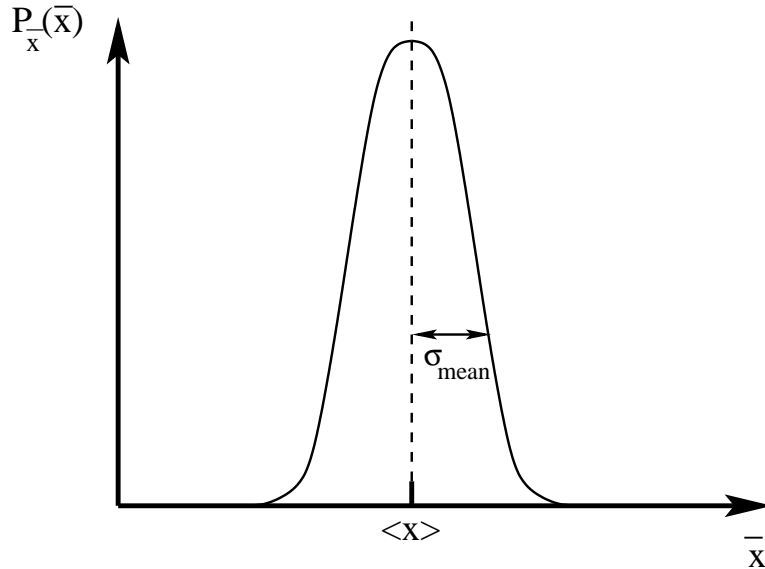


FIG. 1: The distribution of results for the sample mean  $\bar{x}$  obtained by repeating the measurements of the  $N$  data points  $x_i$  many times. The average of this distribution is the exact average value of  $x$ ,

Now we describe an important thought experiment. Let's *suppose* that we could repeat the set of  $N$  measurements *very many* many times, each time obtaining a value of the sample average  $\bar{x}$ . From these results we could construct a distribution,  $\tilde{P}(\bar{x})$ , for the sample average as shown in Fig. 1.

If we do enough repetitions we are effectively averaging over the exact distribution. Hence the average of the sample mean,  $\bar{x}$ , over very many repetitions of the data is given by

$$\langle \bar{x} \rangle = \frac{1}{N} \sum_{i=1}^N \langle x_i \rangle = \langle x \rangle, \quad (3)$$

i.e. it is the exact average over the distribution of  $x$ , as one would intuitively expect, see Fig. 1.

In fact, though, we have only the *one* set of data, so we can not determine  $\langle x \rangle$  exactly. However, Eq. (3) shows that

$$\boxed{\text{the best estimate of } \langle x \rangle \text{ is } \bar{x},} \quad (4)$$

i.e. the sample mean, since averaging the sample mean over many repetitions of the  $N$  data points gives the true mean of the distribution,  $\langle x \rangle$ . An estimate like this, which gives the exact result if averaged over many repetitions of the experiment, is said to be *unbiased*.

We would also like an estimate of the uncertainty or “error bar” in our estimate of  $\bar{x}$  for the exact average  $\langle x \rangle$ . This would be useful, for example, if we have a theoretical prediction for its

value and would like to know if the experiment agrees with it. We can't tell unless we know the uncertainty in our estimate.

We take  $\sigma_{\text{mean}}$ , the standard deviation in  $\bar{x}$  (obtained if one did many repetitions of the  $N$  measurements), to be the uncertainty, or error bar, in  $\bar{x}$ . This is the width of the distribution  $\tilde{P}(\bar{x})$  shown in Fig. 1. A *single* estimate  $\bar{x}$  typically differs from the exact result  $\langle x \rangle$  by an amount of order  $\sigma_{\text{mean}}$ .

We shall now show that the variance of the *mean* of a set of  $N$  random variables is the variance of the distribution of *one* variable divided by  $N$ . To see this, we have

$$\sigma_{\text{mean}}^2 \equiv \langle \bar{x}^2 \rangle - \langle \bar{x} \rangle^2 = \left\langle \left( \frac{1}{N} \sum_{i=1}^N x_i \right)^2 \right\rangle - \left\langle \left( \frac{1}{N} \sum_{i=1}^N x_i \right) \right\rangle^2 \quad (5a)$$

$$= \frac{1}{N^2} \sum_{i=1}^N (\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle) \quad (5b)$$

$$= \frac{1}{N^2} \sum_{i=1}^N (\langle x_i^2 \rangle - \langle x_i \rangle^2) \quad (5c)$$

$$= \frac{1}{N} (\langle x^2 \rangle - \langle x \rangle^2) \quad (5d)$$

$$\boxed{= \frac{\sigma^2}{N}}. \quad (5e)$$

To get from Eq. (5b) to Eq. (5c) note that for  $i \neq j$ , we have  $\langle x_i x_j \rangle = \langle x_i \rangle \langle x_j \rangle$  since  $x_i$  and  $x_j$  are assumed to be statistically independent. (This is where the statistical independence of the data is needed.)

The problem with Eq. (5e) is that **we don't know**  $\sigma^2$  since it is a function of the exact distribution  $P(x)$ . We do, however, know the *sample* variance  $\sigma_{\text{samp}}^2$ , see Eq. (2b), and the average of this over many repetitions of the  $N$  data points, is related to  $\sigma^2$  since

$$\langle \sigma_{\text{samp}}^2 \rangle = \frac{1}{N} \sum_{i=1}^N \langle x_i^2 \rangle - \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \langle x_i x_j \rangle \quad (6a)$$

$$= \langle x^2 \rangle - \frac{1}{N^2} [N(N-1) \langle x \rangle^2 + N \langle x^2 \rangle] \quad (6b)$$

$$= \frac{N-1}{N} [\langle x^2 \rangle - \langle x \rangle^2] \quad (6c)$$

$$= \frac{N-1}{N} \sigma^2. \quad (6d)$$

To get from Eq. (6a) to Eq. (6b), we have separated the terms with  $i = j$  in the last term of Eq. (6a) from those with  $i \neq j$ , and used the fact that each of the  $x_i$  is chosen from the same

distribution. It follows from Eq. (6d) that

$$\boxed{\text{the best estimate of } \sigma^2 \text{ is } \frac{N}{N-1} \sigma_{\text{samp}}^2,} \quad (7)$$

since averaging  $(N/(N-1))\sigma_{\text{samp}}^2$  over many repetitions of  $N$  data points gives  $\sigma^2$ . The estimate for  $\sigma^2$  in Eq. (7) is therefore unbiased.

Combining Eqs. (5e) and (7) gives

$$\boxed{\text{the best estimate of } \sigma_{\text{mean}}^2 \text{ is } \frac{\sigma_{\text{samp}}^2}{N-1}.} \quad (8)$$

Hence, we have now obtained, using only information from from the data, that

$$\boxed{\langle x \rangle = \bar{x} \pm \sigma_{\text{mean}}.} \quad (9)$$

where we estimate

$$\boxed{\sigma_{\text{mean}} = \frac{\sigma_{\text{samp}}}{\sqrt{N-1}}.} \quad (10)$$

Remember that  $\bar{x}$  and  $\sigma_{\text{samp}}$  are obtained from the (one set) of data that is available to us, see Eqs. (2a) and (2b). Neglecting factors of  $-1$  compared with  $N$  (which is usually fine since we are generally dealing with  $N$  quite large) we see from Eq. (6d) that  $\sigma_{\text{samp}}$  is equal to  $\sigma$  and hence, from Eq. (10), that

$$\boxed{\text{the error bar in the mean goes down like } 1/\sqrt{N}.$$

Hence, to reduce the error bar by a factor of 10 one needs 100 times as much data. This is discouraging, but is a fact of life when dealing with random noise.

For Eq. (10) to be really useful we need to know the probability that the true answer  $\langle x \rangle$  lies more than  $\sigma_{\text{mean}}$  away from our estimate  $\bar{x}$ . Fortunately, for large  $N$  the central limit theorem tells us (for distributions where the first two moments are finite) that the distribution of  $\bar{x}$  is a Gaussian. For this distribution we know that the probability of finding a result more than one standard deviation away from the mean is 32%, more than two standard deviations is 4.5% and more than three standard deviations is 0.3%. Hence we expect that most of the time  $\bar{x}$  will be within  $\sigma_{\text{mean}}$  of the correct result and only occasionally will be more than two times  $\sigma_{\text{mean}}$  from it. Even if  $N$  is not very large, so there are some deviations from the Gaussian form, the above numbers are usually a reasonable guide.

Hence, if the theoretical prediction differs from the experimental value of  $\bar{x}$  by several times  $\sigma_{\text{mean}}$ , or more, there is likely to be either some systematic error in the experiment, or else the theory does not apply.

As an aside, although scientists quote  $\sigma_{\text{mean}}$  as the statistical uncertainty in  $\bar{x}$ , by convention, surveys of voters in elections use  $2\sigma_{\text{mean}}$  as a measure of the statistical uncertainty.

Finally, we should mention that Eq. (7) is only an *estimate* of  $\sigma$ , and so Eq. (8) is only an estimate of  $\sigma_{\text{mean}}$ , the error bar on the sample mean. Hence there also is an error bar on  $\sigma_{\text{mean}}$  (i.e. an error bar on the error bar). We won't go into details but, not surprisingly, in most circumstances (enough moments of the distribution of  $x$  are finite), the error on  $\sigma_{\text{mean}}$  goes down like  $1/\sqrt{N}$  relative to its estimated value. Hence the above calculations are only reliable if  $N$  is large (in which case the factor  $N/(N-1)$  in Eqs. (6d) and (7) can be replaced by unity, and the factor of  $-1$  in Eqs. (8) and (10) can be neglected).