Physics 116C

A summary of some important, and often poorly understood, results concerning the mean of the distribution, μ, the mean of of a sample of N data points, $\overline{x}$, the standard deviation of the distribution, σ, the standard deviation of the data points, s, and the error bar on the mean, $\sigma_{\overline{x}}$.

Peter Young

(Dated: November 23, 2009)

Suppose we have a set of experimental data, $x_i, (1 = 1, \cdots, N)$, 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 $\mu \equiv \langle x \rangle$, and a variance σ^2. (The term "standard deviation" is used for σ, the square root of the variance.) We denote an average *over the exact distribution* by angular brackets, e.g.

$$\mu \equiv \langle x \rangle = \int x\, P(x)\, dx\,. \tag{1a}$$

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

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 so as not to significantly perturb the results of the analysis.

The information *from the data* is usefully encoded in two parameters, the sample mean $\overline{x}$ and

the sample standard deviation s which are defined by[1]

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \tag{2a}$$

$$s^2 \equiv \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^2 \tag{2b}$$

$$= \overline{(x-\overline{x})^2} = \overline{x^2} - (\overline{x})^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \left(\frac{1}{N}\sum_{i=1}^{N} x_i\right)^2 . \tag{2c}$$

In statistics, notation is often confusing but crucial to understand. Here, an average indicated by an over-bar, $\overline{\cdots}$, is an average over the *sample of N data points*. This is to be distinguished from an exact average over the distribution $\langle\cdots\rangle$, 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.

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 $\overline{x}$. From these results we could construct a distribution, $\widetilde{P}(\overline{x})$, 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, $\overline{x}$, over very many repetitions of the data, is given by

$$\langle\overline{x}\rangle = \frac{1}{N}\sum_{i=1}^{N}\langle x_i\rangle = \langle x\rangle \equiv \mu\,, \tag{3}$$

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 μ exactly. However, Eq. (3) shows that

$$\boxed{\text{the best estimate of } \mu \text{ is } \overline{x},} \tag{4}$$

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, μ. An estimate like this, which gives the exact result if averaged over many repetitions of the experiment, is said to be $\boxed{\text{unbiased}}$.

[1] The sample variance is often defined with a factor of $N-1$ rather than N in Eq. (2b). However, to me this seems unnatural, and I prefer to define the sample variance in the naive way as the variance over the data. The reason that $N-1$ is often put here, is so the factor of $(N-1)/N$ in Eqs. (6d) and (7) becomse unity, and the factor of $N-1$ in Eqs. (8) and (10) becomes N. Of course, these differences are negligible for large N.

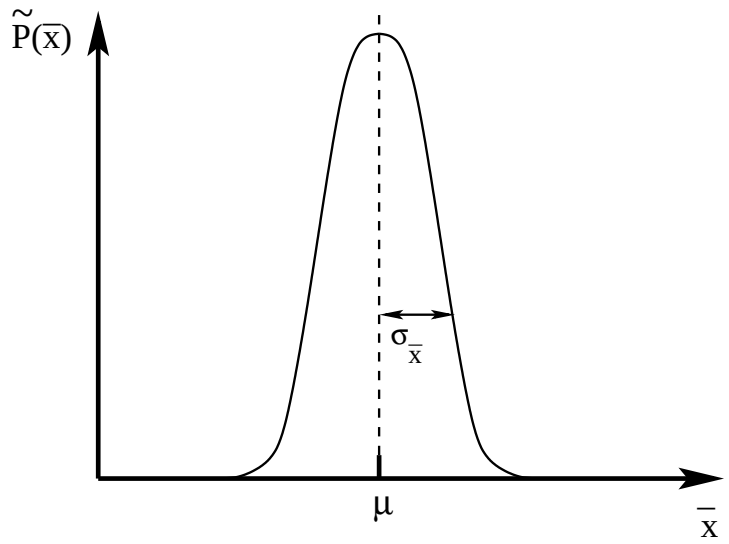

FIG. 1: The distribution of results for the sample mean $\overline{x}$ obtained by repeating the measurements of the N data points x_i many times. The average of this distribution is μ, the exact average value of x,

We would also like an estimate of the uncertainty or "error bar" in our estimate of $\overline{x}$ for the exact average μ. 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 the experimental estimate.

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

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

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

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

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

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

$$\boxed{= \frac{\sigma^2}{N} .} \tag{5e}$$

To get from Eq. (5b) to Eq. (5c) we note that, for $i \neq j$, $\langle x_i x_j \rangle = \langle x_i \rangle \langle x_j \rangle$ since x_i and x_j are

assumed to be statistically independent. (This is where the statistical independence of the data is needed.) We have already met Eq. (5e) in the handout on "The Distribution of the Sum of Random Variables".

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

$$\langle s^2 \rangle = \frac{1}{N} \sum_{i=1}^{N} \langle x_i^2 \rangle - \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \langle x_i x_j \rangle \tag{6a}$$

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

$$= \frac{N-1}{N} \left[\langle x^2 \rangle - \langle x \rangle^2 \right] \tag{6c}$$

$$= \frac{N-1}{N} \, \sigma^2 \, . \tag{6d}$$

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 facts that each of the x_i is chosen from the same distribution and is statistically independent of the others. It follows from Eq. (6d) that

$$\boxed{\text{the best estimate of } \sigma^2 \text{ is } \frac{N}{N-1} \, s^2 \, ,} \tag{7}$$

since averaging $[N/(N-1)]s^2$ over many repetitions of N data points gives σ^2. The estimate for σ^2 in Eq. (7) is therefore unbiased.

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

$$\boxed{\text{the best estimate of } \sigma_{\overline{x}}^2 \text{ is } \frac{s^2}{N-1} \, .} \tag{8}$$

This estimate is also unbiased. We have now obtained, using only information from from the data, that the mean is given by

$$\boxed{\mu = \overline{x} \, \pm \, \sigma_{\overline{x}} \, .} \tag{9}$$

where we estimate

$$\boxed{\sigma_{\overline{x}} = \frac{s}{\sqrt{N-1}} \, .} \tag{10}$$

Remember that $\overline{x}$ and s are the mean and standard deviation of the (one set) of data that is available to us, see Eqs. (2a) and (2b).

As an example, suppose $N = 5$ and the data points are

$$x_i = 10, 11, 12, 13, 14, \tag{11}$$

(not very random looking data it must be admitted). Then, from Eq. (2a) we have $\overline{x} = 12$, and from Eq. (2b)

$$s^2 = \frac{1}{5}\left[(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2\right] = \frac{10}{5} = 2. \tag{12}$$

Hence, from Eq. (10),

$$\sigma_{\overline{x}} = \frac{1}{\sqrt{4}}\sqrt{2} = \frac{1}{\sqrt{2}}. \tag{13}$$

so

$$\mu = \overline{x} \pm \sigma_{\overline{x}} = 12 \pm \frac{1}{\sqrt{2}}. \tag{14}$$

Neglecting factors of -1 compared with N (which is fine if we are dealing with N quite large, the usual case) we see from Eq. (6d) that s is equal to σ and hence, from Eq. (10), that

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 μ lies more than $\sigma_{\overline{x}}$ away from our estimate $\overline{x}$. Fortunately, for large N the central limit theorem tells us (for distributions where the first two moments are finite) that the distribution of $\overline{x}$ 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 $\overline{x}$ will be within $\sigma_{\overline{x}u}$ of the correct result μ, and only occasionally will be more than two times $\sigma_{\overline{x}}$ 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 $\overline{x}$ by several times $\sigma_{\overline{x}}$, or more, there is likely to be either some systematic error in the experiment, or the theory does not apply.

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