Chapter 2
Averages and Error Bars

2.1 Basic Analysis

A reference for the material in this subsection is the book by Taylor [1].

Suppose we have a set of data from a simulation, x_i, $(i = 1, \ldots, N)$, which we shall refer to as a *sample* of data. This data will have some random noise so the x_i are not all equal. Rather they are governed by a distribution $P(x)$, *which we don't know*.

The distribution is normalized,

$$\int_{-\infty}^{\infty} P(x)\, dx = 1, \tag{2.1}$$

and is usefully characterized by its moments, where the nth moment is defined by

$$\langle x^n \rangle = \int_{-\infty}^{\infty} x^n\, P(x)\, dx. \tag{2.2}$$

We will denote the average *over the exact distribution* by angular brackets. Of particular interest are the first and second moments from which one forms the mean μ and variance σ^2, by

$$\mu \equiv \langle x \rangle \tag{2.3a}$$

$$\sigma^2 \equiv \langle\, (x - \langle x \rangle)^2 \,\rangle = \langle x^2 \rangle - \langle x \rangle^2. \tag{2.3b}$$

The term "standard deviation" is used for σ, the square root of the variance.

In this section we will estimate the mean $\langle x \rangle$, and the uncertainty in our estimate, from the N data points x_i. The determination of more complicated averages and resulting error bars will be discussed in Sect. 2.2

© The Author(s) 2015

P. Young, *Everything You Wanted to Know About Data Analysis and Fitting but Were Afraid to Ask*, SpringerBriefs in Physics,
DOI 10.1007/978-3-319-19051-8_2

In order to obtain error bars we need to 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. In Monte Carlo simulations, measurements which differ by a sufficiently large number of Monte Carlo sweeps will be uncorrelated. More precisely the difference in sweep numbers should be greater than a "relaxation time". This is exploited in the "binning" method in which the data used in the analysis is not the individual measurements, but rather an average over measurements during a range of Monte Carlo sweeps, called a "bin". If the bin size is greater than the relaxation time, results from adjacent bins will be (almost) uncorrelated. A pedagogical treatment of binning has been given by Ambegaokar and Troyer [2]. Alternatively, one can do independent Monte Carlo runs, requilibrating each time, and use, as individual data in the analysis, the average from each run.

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]

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i\,, \tag{2.4a}$$

$$s^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \overline{x})^2\,. \tag{2.4b}$$

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. (2.3a) and (2.3b). 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.

[1] The factor of N is often replaced by $N-1$ in the expression for the sample variance in Eq. (2.4b). We note, though, that the final answer for the error bar on the mean, Eq. (2.16), will be independent of how the intermediate quantity s^2 is defined. The rationale for $N-1$ is that the N terms in Eq. (2.4b) are not all independent since $\overline{x}$, which depends on all the x_i, is subtracted. Rather, as will be discussed more in the section on fitting, Chap. 3, there are really only $N-1$ independent variables (called the "number of degrees of freedom" in the fitting context) and so dividing by $N-1$ rather than N also has a rational basis. Here we prefer to use N.

Next we derive two simple results which will be useful later:

1. The mean of the sum of N independent variables *with the same distribution* is N times the mean of a single variable, and
2. The variance of the sum of N independent variables *with the same distribution* is N times the variance of a single variable.

The result for the mean is obvious since, defining $X = \sum_{i=1}^{N} x_i$,

$$\mu_X \equiv \langle X \rangle = \sum_{i=1}^{N} \langle x_i \rangle = N \langle x_i \rangle = N\mu. \tag{2.5}$$

The result for the standard deviation needs a little more work:

$$\begin{aligned}
\sigma_X^2 &\equiv \langle X^2 \rangle - \langle X \rangle^2 && \text{(2.6a)}\\
&= \sum_{i,j=1}^{N} \left(\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle \right) && \text{(2.6b)}\\
&= \sum_{i=1}^{N} \left(\langle x_i^2 \rangle - \langle x_i \rangle^2 \right) && \text{(2.6c)}\\
&= N \left(\langle x^2 \rangle - \langle x \rangle^2 \right) && \text{(2.6d)}\\
&= N\sigma^2. && \text{(2.6e)}
\end{aligned}$$

To get from Eqs. (2.6b) to (2.6c) 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.) If the means and standard deviations are not all the same, then the above results generalize to

$$\mu_X = \sum_{i=1}^{N} \mu_i, \tag{2.7a}$$

$$\sigma_X^2 = \sum_{i=1}^{N} \sigma_i^2. \tag{2.7b}$$

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. 2.1.

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

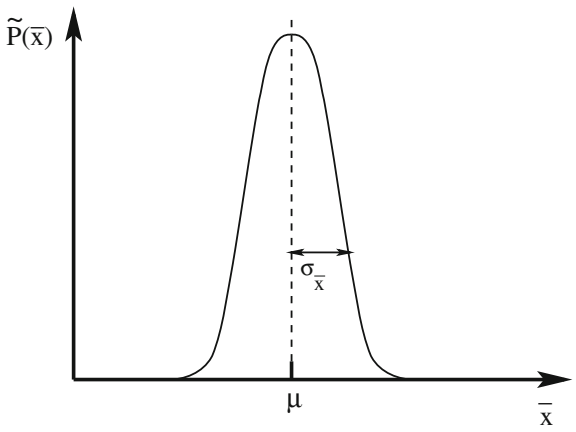

Fig. 2.1 The distribution of results for the sample mean $\overline{x}$ obtained by repeating the measurements of the N data points x_i many times. The average of this distribution is μ, the exact average value of x. The mean, $\overline{x}$, obtained from one sample of data typically differs from μ by an amount of order $\sigma_{\overline{x}}$, the standard deviation of the distribution $\widetilde{P}(\overline{x})$

$$\langle \overline{x} \rangle = \frac{1}{N} \sum_{i=1}^{N} \langle x_i \rangle = \langle x \rangle \equiv \mu, \tag{2.8}$$

i.e. it is the exact average over the distribution of x, as one would intuitively expect, see Fig. 2.1. Eq. (2.8) also follows from Eq. (2.5) by noting that $\overline{x} = X/N$.

In fact, though, we have only the *one* set of data, so we can not determine μ exactly. However, Eq. (2.8) shows that

$$\text{the best estimate of } \mu \text{ is } \overline{x}, \tag{2.9}$$

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 unbiased.

We would also like an estimate of the uncertainty, or "error bar", in our estimate of $\overline{x}$ for the exact average μ. 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 $\overline{x}$. The reason is that $\sigma_{\overline{x}}$ is the width of the distribution $\widetilde{P}(\overline{x})$, shown in Fig. 2.1, so a *single* estimate $\overline{x}$ typically differs from the exact result μ by an amount of this order. The variance $\sigma_{\overline{x}}^2$ is given by

$$\sigma_{\overline{x}}^2 \equiv \langle \overline{x}^2 \rangle - \langle \overline{x} \rangle^2 = \frac{\sigma^2}{N}, \tag{2.10}$$

which follows from Eq. (2.6e) with $\overline{x} = X/N$.

The problem with Eq. (2.10) 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. (2.4b), and the average of this over many repetitions of the N data points, is closely related to σ^2 since

$$\langle s^2 \rangle = \frac{1}{N} \sum_{i=1}^{N} \langle x_i^2 \rangle - \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \langle x_i x_j \rangle \tag{2.11a}$$

$$= \langle x^2 \rangle - \frac{1}{N^2} \left[N(N-1) \langle x \rangle^2 + N \langle x^2 \rangle \right] \tag{2.11b}$$

$$= \frac{N-1}{N} \left[\langle x^2 \rangle - \langle x \rangle^2 \right] \tag{2.11c}$$

$$= \frac{N-1}{N} \sigma^2. \tag{2.11d}$$

To get from Eqs. (2.11a) to (2.11b), we have separated the terms with $i = j$ in the last term of Eq. (2.11a) from those with $i \neq j$, and used the fact that each of the x_i is chosen from the same distribution and is statistically independent of the others. It follows from Eq. (2.11c) that

$$\text{the best estimate of } \sigma^2 \text{ is } \frac{N}{N-1} s^2, \tag{2.12}$$

since averaging s^2 over many repetitions of N data points gives σ^2. The estimate for σ^2 in Eq. (2.12) is therefore unbiased. Note that the expression for s^2 in Eq. (2.4a) is a sum of positive terms, so it is "self-averaging", which means that the deviation of the result for one sample of N data points from the average over many data sets (σ^2 in this case) tends to zero for $N \to \infty$.

Combining Eqs. (2.10) and (2.12) gives

$$\text{the best estimate of } \sigma_{\overline{x}}^2 \text{ is } \frac{s^2}{N-1}, \tag{2.13}$$

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

$$\mu = \overline{x} \pm \sigma_{\overline{x}}, \tag{2.14}$$

where

$$\sigma_{\overline{x}} = \frac{s}{\sqrt{N-1}}, \tag{2.15}$$

which we can write explicitly in terms of the data points as

$$\sigma_{\overline{x}} = \left[\frac{1}{N(N-1)} \sum_{i=1}^{N} (x_i - \overline{x})^2 \right]^{1/2}. \tag{2.16}$$

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. (2.4a) and (2.4b).

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

$$x_i = 10, 11, 12, 13, 14, \tag{2.17}$$

(not very random looking data it must be admitted!). Then, from Eq. (2.4a) we have $\overline{x} = 12$, and from Eq. (2.4b)

$$s^2 = \frac{1}{5}\left[(-2)^2 + (-1)^2 + 0^2 + 1^2 + 2^2\right] = 2. \tag{2.18}$$

Hence, from Eq. (2.15),

$$\sigma_{\overline{x}} = \frac{1}{\sqrt{4}}\sqrt{2} = \frac{1}{\sqrt{2}}. \tag{2.19}$$

so

$$\mu = \overline{x} \pm \sigma_{\overline{x}} = 12 \pm \frac{1}{\sqrt{2}}. \tag{2.20}$$

How does the error bar decrease with the number of statistically independent data points N? Equation (2.11d) shows that s^2 does not vary systematically with N, at large N (where we neglect the factor of -1 compared with N) and so from Eq. (2.15) we see that

the error bar in the mean goes down like $1/\sqrt{N}$ for large 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. (2.15) 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, derived in Appendix A, 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 $\sigma_{\overline{x}}$ 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 often a reasonable guide.

However, as emphasized in Appendix A, distributions which occur in nature typically have much more weight in the tails than a Gaussian. As a result, the weight in the tails of the distribution *of the sum* can also be much larger than for a Gaussian even for quite large values of N, see Fig. A.1. It follows that the probability of an "outlier" can be much higher than that predicted for a Gaussian distribution, as anyone who has invested in the stock market knows well!

We conclude this subsection by discussing the situation when there are several random variables, x, y, z, ..., for which we generate a sample of data: $(x_i, y_i, z_i, \ldots)$ with $i = 1, 2, \ldots, N$. We indicate the means and standard deviations of the different variables by suffices, i.e.

$$\mu_x \equiv \langle x \rangle \tag{2.21a}$$

$$\sigma_x^2 \equiv \langle x^2 \rangle - \langle x \rangle^2, \tag{2.21b}$$

for averages over the exact distribution, and

$$s_x^2 \equiv \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})^2 \,. \tag{2.22}$$

for the sample variance. The main new feature is the appearance of cross-correlations between different variables. One defines the "covariance" of x and y by

$$\mathrm{Cov}(x, y) \equiv \langle xy \rangle - \langle x \rangle \langle y \rangle = \langle (x - \langle x \rangle)\ (y - \langle y \rangle) \rangle. \tag{2.23}$$

It is convenient to have a more compact notation for the covariance, analogous to that in Eq. (2.21b) for the variance. I use the notation σ_{xy}^2 for the covariance of x and y, i.e.

$$\sigma_{xy}^2 \equiv \langle (x - \langle x \rangle)\ (y - \langle y \rangle) \rangle. \tag{2.24}$$

This notation is not ideal since there is no guarantee that the covariance σ_{xy}^2 is positive.[2] The standard notation is to write the covariance of x and y as σ_{xy} (no square), but I find this even more confusing.

By analogy to Eq. (2.24) I write the sample covariance of x and y as

$$s_{xy}^2 \equiv \frac{1}{N} \sum_{i=1}^{N} (x_i - \overline{x})\ (y_i - \overline{y}) \,. \tag{2.25}$$

2.2 Advanced Analysis

In Sect. 2.1 we learned how to estimate a simple average, such as $\mu_x \equiv \langle x \rangle$, plus the error bar in that quantity, from a set of data x_i. Trivially this method also applies to a *linear* combination of different averages, μ_x, μ_y, ... etc. However, we often need

[2] One should therefore think of σ_{xy}^2 as a single quantity, rather than the square of something, just as χ^2, discussed extensively in the section on fitting below, is never regarded as the square of an object called χ. Admittedly, though, χ^2 can not be negative.

more complicated, *non-linear* functions of averages. One example is the fluctuations in a quantity, i.e. $\langle x^2 \rangle - \langle x \rangle^2$. Another example is a dimensionless combination of moments, which gives information about the *shape* of a distribution independent of its overall scale. Such quantities are very popular in finite-size scaling (FSS) analyses since the FSS form is simpler than for quantities with dimension. An popular example, first proposed by Binder, is $\langle x^4 \rangle / \langle x^2 \rangle^2$, which is known as the "kurtosis" (frequently a factor of 3 is subtracted to make it zero for a Gaussian).

Hence, in this section we consider how to determine *non-linear functions* of averages of one or more variables, $f(\mu_y, \mu_z, \ldots)$, where

$$\mu_y \equiv \langle y \rangle, \tag{2.26}$$

etc. For example, the two quantities mentioned in the previous paragraph correspond to

$$f(\mu_y, \mu_z) = \mu_y - \mu_z^2, \tag{2.27}$$

with $y = x^2$ and $z = x$ and

$$f(\mu_y, \mu_z) = \frac{\mu_y}{\mu_z^2}, \tag{2.28}$$

with $y = x^4$ and $z = x^2$.

The natural estimate of $f(\mu_y, \mu_z)$ from the sample data is clearly $f(\overline{y}, \overline{z})$. However, it will take some more thought to estimate the error bar in this quantity. The traditional way of doing this is called "error propagation", described in Sect. 2.2.1 and Chap. 3 of Ref. [1]. However, it is now more common to use either "jackknife" or "bootstrap" procedures, described in Sects. 2.2.2 and 2.2.3. At the price of some additional computation, which is no difficulty when done on a modern computer (though it would have been tedious in the old days when statistics calculations were done by hand), these methods automate the calculation of the error bar.

In addition, the estimate of $f(\mu_y, \mu_z)$ turns out to have some *bias* if f is a non-linear function. Usually this is small effect because it is order $1/N$, see for example Eq. (2.34), whereas the statistical error is of order $1/\sqrt{N}$. Since N is usually large, the bias is generally much less than the statistical error and so can generally be neglected. In any case, the jackknife and bootstrap methods also enable one to eliminate the leading ($\sim 1/N$) contribution to the bias in a automatic fashion.

2.2.1 Traditional Method

First we will discuss the traditional method, known as error propagation [1], to compute the error bar and bias. We expand $f(\overline{y}, \overline{z})$ about $f(\mu_y, \mu_z)$ up to second order in the deviations:

$$f(\overline{y},\overline{z}) = f(\mu_y,\mu_z) + (\partial_{\mu_y} f)\,\delta_{\overline{y}} + (\partial_{\mu_z} f)\,\delta_{\overline{z}} + \frac{1}{2}\,(\partial^2_{\mu_y\mu_y} f)\,\delta^2_{\overline{y}} + (\partial^2_{\mu_y\mu_z} f)\,\delta_{\overline{y}}\delta_{\overline{z}} + \frac{1}{2}\,(\partial^2_{\mu_z\mu_z} f)\,\delta^2_{\overline{z}} + \cdots, \tag{2.29}$$

where

$$\delta_{\overline{y}} = \overline{y} - \mu_y, \tag{2.30}$$

etc.

The terms of first order in the $\delta's$ in Eq. (2.29) give the leading contribution to the error, but would average to zero if the procedure were to be repeated many times. However, the terms of second order do not average to zero and so give the leading contribution to the bias. We now estimate that bias.

Averaging Eq. (2.29) over many repetitions, and noting that

$$\langle\delta^2_{\overline{y}}\rangle = \langle\overline{y}^2\rangle - \langle\overline{y}\rangle^2 \equiv \sigma^2_{\overline{y}}, \quad \langle\delta^2_{\overline{z}}\rangle = \langle\overline{z}^2\rangle - \langle\overline{z}\rangle^2 \equiv \sigma^2_{\overline{z}}, \quad \langle\delta_{\overline{y}}\delta_{\overline{z}}\rangle = \langle\overline{y}\,\overline{z}\rangle - \langle\overline{y}\rangle\langle\overline{z}\rangle \equiv \sigma^2_{\overline{y}\,\overline{z}}, \tag{2.31}$$

we get

$$\langle f(\overline{y},\overline{z})\rangle - f(\mu_y,\mu_z) = \frac{1}{2}\,(\partial^2_{\mu_y\mu_y} f)\,\sigma^2_{\overline{y}} + (\partial^2_{\mu_y\mu_z} f)\,\sigma^2_{\overline{y}\,\overline{z}} + \frac{1}{2}\,(\partial^2_{\mu_z\mu_z} f)\,\sigma^2_{\overline{z}} + \cdots. \tag{2.32}$$

As shown in Eq. (2.13) $\sigma^2_{\overline{y}}$ is $(N-1)^{-1}$ times the average sample variance $\langle s^2_y\rangle$. Furthermore, as noted below Eq. (2.12), s^2_y is self averaging, which means that the difference between the value of s^2_y from one data set and the average over all data sets, σ^2_y, tends to zero for $N \to \infty$. Hence we can replace $\sigma^2_{\overline{y}}$ by $(N-1)^{-1}s^2_y$, and similarly replace $\sigma^2_{\overline{z}}$ by $(N-1)^{-1}s^2_z$. In the same way, we can replace $\sigma^2_{\overline{y}\,\overline{z}}$ by $(N-1)^{-1}$ times s^2_{yz}, the sample *covariance* of y and z, defined in Eq. (2.25). Hence, from Eq. (2.32), we have

$$f(\mu_y,\mu_z) = \langle f(\overline{y},\overline{z})\rangle - \frac{1}{(N-1)}\left[\frac{1}{2}\,(\partial^2_{\mu_y\mu_y} f)\,s^2_y + (\partial^2_{\mu_y\mu_z} f)\,s^2_{yz} + \frac{1}{2}\,(\partial^2_{\mu_z\mu_z} f)\,s^2_z\right] + \cdots. \tag{2.33}$$

The leading contribution to the bias is the $1/(N-1)$ term. It follows from Eq. (2.33) that if one wants to eliminate the leading contribution to the bias one should

$$\text{estimate } f(\mu_y,\mu_z) \text{ from } f(\overline{y},\overline{z}) - \frac{1}{(N-1)}\left[\frac{1}{2}\,(\partial^2_{\mu_y\mu_y} f)\,s^2_y + (\partial^2_{\mu_y\mu_z} f)\,s^2_{yz} + \frac{1}{2}\,(\partial^2_{\mu_z\mu_z} f)\,s^2_z\right]. \tag{2.34}$$

As claimed earlier, the bias correction is of order $1/N$. Note that it vanishes if f is a linear function, as shown in Sect. 2.1. The generalization to functions of more than two averages, $f(\mu_y, \mu_z, \mu_w, \ldots)$, is obvious.

Next we discuss the leading *error* in using $f(\overline{y}, \overline{z})$ as an estimate for $f(\mu_y, \mu_z)$. This comes from the terms linear in the δ's in Eq. (2.29). Just including these terms we have

$$\langle f(\overline{y}, \overline{z}) \rangle = f(\mu_y, \mu_z), \tag{2.35a}$$

$$\langle f^2(\overline{y}, \overline{z}) \rangle = f^2(\mu_y, \mu_z) + (\partial_{\mu_y} f)^2 \langle \delta_{\overline{y}}^2 \rangle + 2(\partial_{\mu_y} f)(\partial_{\mu_z} f) \langle \delta_{\overline{y}} \delta_{\overline{z}} \rangle + (\partial_{\mu_z} f)^2 \langle \delta_{\overline{z}}^2 \rangle. \tag{2.35b}$$

Hence

$$\begin{aligned} \sigma_f^2 &\equiv \langle f^2(\overline{y}, \overline{z}) \rangle - \langle f(\overline{y}, \overline{z}) \rangle^2 \\ &= (\partial_{\mu_y} f)^2 \langle \delta_{\overline{y}}^2 \rangle + 2(\partial_{\mu_y} f)(\partial_{\mu_z} f) \langle \delta_{\overline{y}} \delta_{\overline{z}} \rangle + (\partial_{\mu_z} f)^2 \langle \delta_{\overline{z}}^2 \rangle. \end{aligned} \tag{2.36}$$

As above, we use $s_{yy}^2/(N-1)$ as an estimate of $\langle \delta_{\overline{y}}^2 \rangle$ and similarly for the other terms. Hence

$$\text{the best estimate of } \sigma_f^2 \text{ is } \frac{1}{(N-1)} (\partial_{\mu_y} f)^2 s_y^2 + 2(\partial_{\mu_y} f)(\partial_{\mu_z} f) s_{yz}^2 + (\partial_{\mu_z} f)^2 s_z^2. \tag{2.37}$$

This estimate is unbiased to leading order in N. Note that we need to keep track not only of fluctuations in y and z, characterized by their variances s_y^2 and s_z^2, but also cross correlations between y and z, characterized by their covariance s_{yz}^2.

Hence, still to leading order in N, we get

$$f(\mu_y, \mu_z) = f(\overline{y}, \overline{z}) \pm \sigma_f, \tag{2.38}$$

where we estimate the error bar σ_f from Eq. (2.37) which shows that it is of order $1/\sqrt{N}$. The generalization to functions of more than two averages is obvious.

Note that in the simple case studied in Sect. 2.1 where $f(\mu_x)$ is a linear function, $f = \mu_x$, Eq. (2.33) tells us that there is no bias, which is correct, and Eq. (2.37) gives an expression for the error bar which agrees with Eq. (2.15).

In Eqs. (2.34) and (2.37) we need to keep track how errors in the individual quantities like $\overline{y}$ propagate to the estimate of the function f. This requires inputting by hand the various partial derivatives into the analysis program, and keeping track of all the variances and covariances. In the next two sections we see how *resampling* the data automatically takes account of error propagation without needing to input the partial derivatives and keep track of variances and covariances. There are two resampling approaches, called jackknife and bootstrap, and each provide a *fully automatic* method of determining error bars and bias.

2.2.2 Jackknife

We define the ith jackknife estimate, y_i^J $(i = 1, 2, \ldots, N)$ to be the average over all data in the sample *except the point* i, i.e.

$$y_i^J \equiv \frac{1}{N-1} \sum_{j \neq i} y_j \quad \left(= \overline{y} + \frac{1}{N-1} (\overline{y} - y_i) \right). \tag{2.39}$$

We also define corresponding jackknife estimates of the function f (again for concreteness we will assume that f is a function of just 2 averages but the generalization will be obvious):

$$f_i^J \equiv f(y_i^J, z_i^J). \tag{2.40}$$

In other words, we use the jackknife values, y_i^J, z_i^J, rather than the sample means, $\overline{y}, \overline{z}$, as the arguments of f. For example a jackknife estimate of the Binder ratio $\langle x^4 \rangle / \langle x^2 \rangle^2$ is

$$f_i^J = \frac{(N-1)^{-1} \sum_{j,(j \neq i)} x_j^4}{\left[(N-1)^{-1} \sum_{j,(j \neq i)} x_j^2 \right]^2} \tag{2.41}$$

The overall jackknife estimate of $f(\mu_y, \mu_z)$ is then the average over the N jackknife estimates f_i^J:

$$\overline{f^J} \equiv \frac{1}{N} \sum_{i=1}^{N} f_i^J. \tag{2.42}$$

It is straightforward to show that if f is a linear function of μ_y and μ_z then $\overline{f^J} = f(\overline{y}, \overline{z})$, i.e. the jackknife and standard averages are identical, see e.g. Eq. (2.39). However, when f is not a linear function, so there is bias, there *is* a difference, and we will now show the resampling carried out in the jackknife method can be used to determine bias and error bars in an automated way.

We proceed as for the derivation of Eq. (2.33), which we now write as

$$f(\mu_y, \mu_z) = \langle f(\overline{y}, \overline{z}) \rangle - \frac{A}{N} - \frac{B}{N^2} + \cdots, \tag{2.43}$$

where A is the term in rectangular brackets in Eq. (2.33), and we have added the next order correction. The jackknife data sets have $N-1$ points with the same distribution as the N points in the actual distribution, and so the bias in the jackknife average will be of the same form, with the same values of A and B, but with N replaced by $N-1$, i.e.

$$f(\mu_y, \mu_z) = \langle \overline{f^J} \rangle - \frac{A}{N-1} - \frac{B}{(N-1)^2} \cdots. \tag{2.44}$$

We can therefore eliminate the leading contribution to the bias by forming an appropriate linear combination of $f(\overline{y},\overline{z})$ and $\overline{f^J}$, namely

$$f(\mu_y,\mu_z) = N\langle f(\overline{y},\overline{z})\rangle - (N-1)\langle\overline{f^J}\rangle + O\left(\frac{1}{N^2}\right). \tag{2.45}$$

It follows that, to eliminate the leading bias without computing partial derivatives, one should

$$\text{estimate } f(\mu_y,\mu_z) \text{ from } Nf(\overline{y},\overline{z}) - (N-1)\overline{f^J}. \tag{2.46}$$

The bias is then of order $1/N^2$. However, as mentioned earlier, bias is usually not a big problem because, even without eliminating the leading contribution, the bias is of order $1/N$ whereas the statistical error is of order $1/\sqrt{N}$ which is much bigger if N is large. In most cases, therefore, N is sufficiently large that one can use *either* the usual average $f(\overline{y},\overline{z})$, or the jackknife average $\overline{f^J}$ in Eq. (2.42), to estimate $f(\mu_y,\mu_z)$, since the difference between them will be much smaller than the statistical error. In other words, elimination of the leading bias using Eq. (2.46) is usually not necessary.

Next we show that the jackknife method gives error bars, which agree with Eq. (2.37) but without the need to explicitly keep track of the partial derivatives and the variances and covariances.

We define the variance of the jackknife averages by

$$s^2_{f^J} \equiv \overline{\left(f^J\right)^2} - \left(\overline{f^J}\right)^2, \tag{2.47}$$

where

$$\overline{\left(f^J\right)^2} = \frac{1}{N}\sum_{i=1}^{N}\left(f_i^J\right)^2. \tag{2.48}$$

Using Eqs. (2.40) and (2.42), we expand $\overline{f^J}$ away from the exact result $f(\mu_y,\mu_z)$. Just including the leading contribution gives

$$\begin{aligned}
\overline{f^J} - f(\mu_y,\mu_z) &= \frac{1}{N}\sum_{i=1}^{N}\left[(\partial_{\mu_y} f)\,(y_i^J-\mu_y) + (\partial_{\mu_z} f)\,(z_i^J-\mu_z)\right] \\
&= \frac{1}{N(N-1)}\sum_{i=1}^{N}\Big[(\partial_{\mu_y} f)\,\left\{N(\overline{y}-\mu_y) - (y_i-\mu_y)\right\} \\
&\qquad\qquad + (\partial_{\mu_z} f)\,\{N(\overline{z}-\mu_z) - (z_i-\mu_z)\}\Big] \\
&= (\partial_{\mu_y} f)\,(\overline{y}-\mu_y) + (\partial_{\mu_z} f)\,(\overline{z}-\mu_z),
\end{aligned} \tag{2.49}$$

where we used Eq. (2.39). Similarly we find

$$\begin{aligned}\overline{(f^J)^2} &= \frac{1}{N}\sum_{i=1}^{N}\left[f(\mu_y,\mu_z)+(\partial_{\mu_y}f)\,(y_i^J-\mu_y)+(\partial_{\mu_z}f)\,(z_i^J-\mu_z)\right]^2\\ &= f^2(\mu_y,\mu_z)+2f(\mu_y,\mu_z)\left[(\partial_{\mu_y}f)\,(\overline{y}-\mu_y)+(\partial_{\mu_z}f)\,(\overline{z}-\mu_z)\right]\\ &\quad+(\partial_{\mu_y}f)^2\left[(\overline{y}-\mu_y)^2+\frac{s_y^2}{(N-1)^2}\right]+(\partial_{\mu_z}f)^2\left[(\overline{z}-\mu_z)^2+\frac{s_z^2}{(N-1)^2}\right]\\ &\quad+2(\partial_{\mu_y}f)(\partial_{\mu_z}f)\left[(\overline{y}-\mu_y)(\overline{z}-\mu_z)+\frac{s_{yz}^2}{(N-1)^2}\right].\end{aligned} \tag{2.50}$$

Hence, from Eqs. (2.47) to (2.49), the variance in the jackknife estimates is given by

$$s_{f^J}^2 = \frac{1}{(N-1)^2}\left[(\partial_{\mu_y}f)^2\,s_y^2+(\partial_{\mu_z}f)^2\,s_z^2+2(\partial_{\mu_y}f)(\partial_{\mu_z}f)s_{yz}^2\right], \tag{2.51}$$

which is just $1/(N-1)$ times σ_f^2, the estimate of the square of the error bar in $f(\overline{y},\overline{z})$ given in Eq. (2.37). Hence

$$\text{the jackknife estimate for } \sigma_f \text{ is } \sqrt{N-1}\,s_{f^J}. \tag{2.52}$$

Note that this is directly obtained from the jackknife estimates without having to put in the partial derivatives by hand. Note too that the $\sqrt{N-1}$ factor is in the *numerator* whereas the factor of $\sqrt{N-1}$ in Eq. (2.15) is in the *denominator*. Intuitively the reason for this difference is that the jackknife estimates are very close, much closer than the error in the means, since they would all be equal except that each one omits just one data point.

If N is very large, roundoff errors could become significant from having to subtract large, almost equal, numbers to get the error bar from the jackknife method. It is then advisable to group the N data points into N_{group} groups (or "bins") of data and take, as individual data points in the jackknife analysis, the average of the data in each group. The above results clearly go through with N replaced by N_{group}.

To summarize this subsection, to estimate $f(\mu_y,\mu_z)$ one can use either $f(\overline{y},\overline{z})$ or the jackknife average $\overline{f^J}$ in Eq. (2.42). The error bar in this estimate, σ_f, is related to the standard deviation in the jackknife estimates s_{f^J} by Eq. (2.52).

2.2.3 Bootstrap

The bootstrap, like the jackknife, is a resampling of the N data points. A brief discussion, in the context of data analysis is given in Ref. [3]. Whereas jackknife considers N new data sets, each of containing all the original data points minus

one, bootstrap uses N_{boot} data sets each containing N points obtained by random (Monte Carlo) sampling of the original set of N points. During the Monte Carlo sampling, the probability that a data point is picked is $1/N$ irrespective of whether it has been picked before. (In the statistics literature this is called picking from a set "with replacement".) Hence a given data point x_i will, *on average*, appear once in each Monte Carlo-generated data set, but may appear not at all, or twice, and so on. The probability that x_i appears n_i times is close to a Poisson distribution with mean unity. However, it is not exactly Poissonian because of the constraint in Eq. (2.53). It turns out that we shall need to include the deviation from the Poisson distribution even for large N. We shall use the term "bootstrap" to denote the Monte Carlo-generated data sets.

More precisely, let us suppose that the number of times x_i appears in a bootstrap data set is n_i. Since each bootstrap dataset contains exactly N data points, we have the constraint

$$\sum_{i=1}^{N} n_i = N. \tag{2.53}$$

Consider one of the N variables x_i. Each time we generate an element in a bootstrap dataset the probability that it is x_i is $1/N$, which we will denote by p. From standard probability theory, the probability that x_i occurs n_i times is given by a binomial distribution

$$P(n_i) = \frac{N!}{n_i!\,(N-n_i)!}\, p^{n_i}(1-p)^{N-n_i}. \tag{2.54}$$

The mean and standard deviation of a binomial distribution are given by

$$[n_i]_{\text{MC}} = Np = 1, \tag{2.55}$$

$$[n_i^2]_{\text{MC}} - [n_i]_{\text{MC}}^2 = Np(1-p) = 1 - \frac{1}{N}, \tag{2.56}$$

where $[\ldots]_{\text{MC}}$ denotes an exact average over bootstrap samples (for a fixed original data set x_i). For $N \to \infty$, the binomial distribution goes over to a Poisson distribution, for which the factor of $1/N$ in Eq. (2.56) does not appear. We assume that N_{boot} is sufficiently large that the bootstrap average we carry out reproduces this result with sufficient accuracy. Later, we will discuss what values for N_{boot} are sufficient in practice. Because of the constraint in Eq. (2.53), n_i and n_j (with $i \neq j$) are not independent and we find, by squaring Eq. (2.53) and using Eqs. (2.55) and (2.56), that

$$[n_i n_j]_{\text{MC}} - [n_i]_{\text{MC}}[n_j]_{\text{MC}} = -\frac{1}{N} \qquad (i \neq j). \tag{2.57}$$

First of all we just consider the simple average $\mu_x \equiv \langle x \rangle$, for which, of course, the standard methods in Sect. 2.1 suffice, so bootstrap is not necessary. However, this will show how to get averages and error bars in a simple case, which we will then generalize to non-linear functions of averages.

We denote the average of x for a given bootstrap data set by x_α^B, where α runs from 1 to N_{boot}, i.e

$$x_\alpha^B = \frac{1}{N}\sum_{i=1}^{N} n_{i,\alpha} x_i . \tag{2.58}$$

We then compute the bootstrap average of the mean of x and the bootstrap variance in the mean, by averaging over all the bootstrap data sets. We assume that N_{boot} is large enough for the bootstrap average to be exact, so we can use Eqs. (2.56) and (2.57). The result is

$$\overline{x^B} \equiv \frac{1}{N_{\text{boot}}}\sum_{\alpha=1}^{N_{\text{boot}}} x_\alpha^B = \frac{1}{N}\sum_{i=1}^{N}[n_i]_{\text{MC}} x_i = \frac{1}{N}\sum_{i=1}^{N} x_i = \overline{x} \tag{2.59}$$

$$s_{x^B}^2 \equiv \overline{\left(x^B\right)^2} - \left(\overline{x^B}\right)^2 = \frac{1}{N^2}\left(1-\frac{1}{N}\right)\sum_i x_i^2 - \frac{1}{N^3}\sum_{i\neq j} x_i x_j = \frac{1}{N}\left(\overline{x^2} - \overline{x}^2\right) = \frac{s^2}{N}, \tag{2.60}$$

where

$$\overline{\left(x^B\right)^2} \equiv \frac{1}{N_{\text{boot}}}\sum_{\alpha=1}^{N_{\text{boot}}}\left[\left(x_\alpha^B\right)^2\right]_{\text{MC}} . \tag{2.61}$$

We now average Eqs. (2.59) and (2.60) over many repetitions of the original data set x_i. Averaging Eq. (2.59) gives

$$\langle \overline{x^B} \rangle = \langle \overline{x} \rangle = \langle x \rangle \equiv \mu_x . \tag{2.62}$$

This shows that the bootstrap average $\overline{x^B}$ is an unbiased estimate of the exact average μ_x. Averaging Eq. (2.60) gives

$$\left\langle s_{x^B}^2 \right\rangle = \frac{1}{N}\left\langle s^2 \right\rangle = \frac{N-1}{N^2}\sigma^2 = \frac{N-1}{N}\sigma_{\overline{x}}^2, \tag{2.63}$$

where we used Eqs. (2.10) and (2.11c). Since $\sigma_{\overline{x}}$ is the uncertainty in the sample mean, we see that

$$\text{the bootstrap estimate of } \sigma_{\overline{x}} \text{ is } \sqrt{\frac{N}{N-1}}\, s_{x^B} . \tag{2.64}$$

Remember that s_{x^B} is the standard deviation of the bootstrap data sets. Usually N is sufficiently large that the square root in Eq. (2.64) can be replaced by unity.

As for the jackknife, these results can be generalized to finding the error bar in some possibly non-linear function, $f(\mu_y, \mu_z)$, rather than for μ_x. One computes the bootstrap estimates for $f(\mu_y, \mu_z)$, which are

$$f_\alpha^B = f(y_\alpha^B, z_\alpha^B). \tag{2.65}$$

In other words, we use the bootstrap values, y_α^B, f_α^B, rather than the sample means, $\overline{y}$, $\overline{z}$, as the arguments of f. The final bootstrap estimate for $f(\mu_y, \mu_z)$ is the average of these, i.e

$$\overline{f^B} = \frac{1}{N_{\text{boot}}} \sum_{\alpha=1}^{N_{\text{boot}}} f_\alpha^B. \tag{2.66}$$

Following the same methods in the jackknife section, one obtains the error bar, σ_f, in $f(\mu_y, \mu_z)$. The result is

$$\text{the bootstrap estimate for } \sigma_f \text{ is } \sqrt{\frac{N}{N-1}}\, s_{f^B}, \tag{2.67}$$

where

$$s_{f^B}^2 = \overline{\left(f^B\right)^2} - \left(\overline{f^B}\right)^2, \tag{2.68}$$

is the variance of the bootstrap estimates. Here

$$\overline{\left(f^B\right)^2} \equiv \frac{1}{N_{\text{boot}}} \sum_{\alpha=1}^{N_{\text{boot}}} \left(f_\alpha^B\right)^2. \tag{2.69}$$

Usually N is large enough that the factor of $\sqrt{N/(N-1)}$ is Eq. (2.67) can be replaced by unity. Equation (2.67) corresponds to the result Eq. (2.64) which we derived for the special case of $f = \mu_x$.

Again, following the same path as in the jackknife section, it is straightforward to show that the bias of the estimates in Eqs. (2.66) and (2.67) is of order $1/N$ and so vanishes for $N \to \infty$. However, if N is not too large it may be useful to eliminate the leading contribution to the bias in the mean, as we did for jackknife in Eq. (2.46). The result is that one should

$$\text{estimate } f(\mu_y, \mu_z) \text{ from } 2f(\overline{y}, \overline{z}) - \overline{f^B}. \tag{2.70}$$

The bias in Eq. (2.70) is of order $1/N^2$, whereas $f(\overline{y}, \overline{z})$ and $\overline{f^B}$ each have a bias of order $1/N$. However, it is not normally necessary to eliminate the bias since, if N is large, the bias is much smaller than the statistical error.

I have not systematically studied the values of N_{boot} that are needed in practice to get accurate estimates for the error. It seems that N_{boot} in the range 100–500 is typically chosen, and this seems to be adequate irrespective of how large N is.

It is sometimes stated, e.g. [3], that the bootstrap method can give error bars correctly even when there are correlations in the data. This is not so. If one applies bootstrap to the direct average of a set of data, it simply reproduces the results of

the standard analysis. Bootstrap is useful both to get error bars when one is looking at combination of averages of the data, and to get confidence limits when the noise on the data is not Gaussian, see Sect. 3.7. Unfortunately, bootstrap does not work miracles and cannot give correct error bars for correlated data.

To summarize this subsection, to estimate $f(\mu_y, \mu_z)$ one can either use $f(\overline{y}, \overline{z})$, or the bootstrap average in Eq. (2.66), and the error bar in this estimate, σ_f, is related to the standard deviation in the bootstrap estimates by Eq. (2.67).

2.2.4 Jackknife or Bootstrap?

The jackknife approach involves less calculation than bootstrap, and is fine for estimating combinations of moments of the measured quantities. Furthermore, identical results are obtained each time jackknife is run on the same set of data, which is not the case for bootstrap. However, the range of the jackknife estimates is very much smaller, by a factor of $\sqrt{N}$ for large N, than the scatter of averages which would be obtained from individual data sets, see Eq. (2.52). By contrast, for bootstrap, σ_{f^B}, which measures the deviation of the bootstrap estimates f_α^B from the result for the single actual data set $f(\overline{y}, \overline{z})$, *is equal to* σ_f, the deviation of the average of a single data set from the exact result $f(\mu_y, \mu_z)$ (if we replace the factor of $N/(N-1)$ by unity, see Eq. (2.67)). This is the main strength of the bootstrap approach; it samples the full range of the distribution of the sample distribution. Hence, if you want to generate data which covers the full range then should use bootstrap. This is useful in fitting, see for example, Sect. 3.7. However, if you just want to generate error bars on combinations of moments quickly and easily, then use jackknife.

References

1. J.R. Taylor, *The Study of Uncertainties in Physical Measurements* (University Science Books, Sausalito, California, 1997)
2. V. Ambegaokar, M. Troyer, Estimating errors reliably in Monte Carlo simulations of the Ehrenfest model. Am. J. Phys. **78**, 150 (2009)
3. M.E.J. Newman, G.T. Barkema, *Monte Carlo Methods in Statistical Physics* (Oxford University Press Inc., New York, USA, 1999)

http://www.springer.com/978-3-319-19050-1

Everything You Wanted to Know About Data Analysis and Fitting but Were Afraid to Ask

Young, P.

2015, X, 85 p. 15 illus., 11 illus. in color., Softcover
ISBN: 978-3-319-19050-1