Chapter 6
Mean, Median, and Average Values of Variables

Abstract The data analyst often faces the question of what is the "best" value to report from N measurements of a random variable. In this chapter we investigate the use of the linear average, the weighted average, the median and a logarithmic average that may be applicable when the variable has a log-normal distribution. The latter may be useful when a variable has errors that are proportional to their measurements, avoiding the inherent bias arising in the weighted average from measurements with small values and small errors. We also introduce a relative-error weighted average that can be used as an approximation for the logarithmic mean for log-normal distributions.

6.1 Linear and Weighted Average

In the previous chapter (see Sect. 5.1.3) we have shown that the weighted mean is the most likely value of the mean of the random variable. Therefore, the weighted mean is a commonly accepted quantity to report as the best estimate for the value of a measured quantity. If the measurements have the same standard deviation, then the weighted mean becomes the linear average; in general, the linear and weighted means differ unless all measurement errors are identical.

The difference between linear average and weighted mean can be illustrated with an example. Consider the $N = 25$ measurements shown in Table 6.1, which reports the measurement of the energy of certain astronomical sources made at a given radius [5]. This dataset is illustrative of the general situation of the measurement of a quantity (in this example, the ratio between the two measurements) in the presence of different measurement error. The weighted mean is 0.90 ± 0.02, while the linear average is 1.01 (see Problem 6.1). The difference is clearly due to the presence of a few measurements with a low value of the ratio that carry higher weight because of the small measurement error (for example, source 15).

Which of the two values is more representative? This question can be addressed by making the following observations. The measurement error reported in the table reflects the presence of such sources of uncertainty as Poisson fluctuations in the detection of photons from the celestial sources. The same type of uncertainty would also apply to other experiments, in particular those based on the counting of events.

M. Bonamente, *Statistics and Analysis of Scientific Data*, Graduate Texts in Physics, DOI 10.1007/978-1-4939-6572-4_6

Table 6.1 Dataset with measurement of energy for $N = 25$ different sources and their ratio

Source	Radius	Energy Method #1	Method #2	Ratio
1	$221.1\pm^{11.0}_{12.3}$	$8.30\pm^{0.76}_{0.88}$	$9.67\pm^{1.14}_{1.12}$	$0.86\pm^{0.08}_{0.07}$
2	$268.5\pm^{22.1}_{20.7}$	$4.92\pm^{0.77}_{0.70}$	$4.19\pm^{0.82}_{0.70}$	$1.17\pm^{0.16}_{0.15}$
3	$138.4\pm^{12.7}_{11.9}$	$3.03\pm^{0.53}_{0.49}$	$2.61\pm^{0.59}_{0.49}$	$1.16\pm^{0.20}_{0.18}$
4	$714.3\pm^{23.5}_{34.5}$	$49.61\pm^{3.15}_{3.19}$	$60.62\pm^{4.84}_{6.13}$	$0.82\pm^{0.06}_{0.05}$
5	$182.3\pm^{18.5}_{15.1}$	$2.75\pm^{0.49}_{0.43}$	$3.30\pm^{0.81}_{0.61}$	$0.83\pm^{0.14}_{0.14}$
6	$72.1\pm^{5.5}_{5.7}$	$1.01\pm^{0.23}_{0.20}$	$0.86\pm^{0.14}_{0.13}$	$1.17\pm^{0.24}_{0.21}$
7	$120.3\pm^{8.6}_{7.5}$	$5.04\pm^{0.66}_{0.57}$	$3.80\pm^{0.72}_{0.57}$	$1.33\pm^{0.16}_{0.15}$
8	$196.2\pm^{15.1}_{15.5}$	$5.18\pm^{0.73}_{0.70}$	$6.00\pm^{1.17}_{1.11}$	$0.86\pm^{0.14}_{0.11}$
9	$265.7\pm^{8.7}_{8.6}$	$12.17\pm^{1.22}_{1.17}$	$10.56\pm^{0.93}_{0.95}$	$1.14\pm^{0.13}_{0.10}$
10	$200.0\pm^{9.6}_{10.7}$	$7.74\pm^{0.57}_{0.58}$	$6.26\pm^{0.78}_{0.83}$	$1.24\pm^{0.14}_{0.11}$
11	$78.8\pm^{5.6}_{5.1}$	$1.08\pm^{0.16}_{0.15}$	$0.73\pm^{0.11}_{0.10}$	$1.49\pm^{0.26}_{0.24}$
12	$454.4\pm^{20.3}_{20.3}$	$17.10\pm^{2.64}_{2.03}$	$23.12\pm^{2.36}_{2.32}$	$0.75\pm^{0.07}_{0.06}$
13	$109.4\pm^{8.3}_{8.3}$	$3.31\pm^{0.34}_{0.34}$	$3.06\pm^{0.54}_{0.52}$	$1.09\pm^{0.18}_{0.15}$
14	$156.5\pm^{11.5}_{10.2}$	$2.36\pm^{0.61}_{0.58}$	$2.31\pm^{0.36}_{0.31}$	$1.02\pm^{0.26}_{0.23}$
15	$218.0\pm^{6.6}_{5.9}$	$14.02\pm^{0.75}_{0.75}$	$21.59\pm^{1.82}_{1.82}$	$0.65\pm^{0.04}_{0.04}$
16	$370.7\pm^{7.6}_{8.0}$	$31.41\pm^{1.56}_{1.56}$	$29.67\pm^{1.56}_{1.57}$	$1.06\pm^{0.06}_{0.06}$
17	$189.1\pm^{16.4}_{15.4}$	$2.15\pm^{0.45}_{0.39}$	$2.52\pm^{0.57}_{0.51}$	$0.86\pm^{0.22}_{0.18}$
18	$150.5\pm^{4.2}_{4.6}$	$3.39\pm^{0.57}_{0.50}$	$4.75\pm^{0.44}_{0.46}$	$0.72\pm^{0.11}_{0.11}$
19	$326.7\pm^{12.1}_{9.9}$	$15.73\pm^{1.43}_{1.30}$	$18.03\pm^{1.54}_{1.26}$	$0.87\pm^{0.06}_{0.06}$
20	$189.1\pm^{9.9}_{9.1}$	$5.04\pm^{0.65}_{0.55}$	$4.61\pm^{0.61}_{0.50}$	$1.09\pm^{0.12}_{0.12}$
21	$147.7\pm^{8.0}_{11.1}$	$2.53\pm^{0.29}_{0.30}$	$2.76\pm^{0.37}_{0.48}$	$0.93\pm^{0.12}_{0.10}$
22	$504.6\pm^{12.5}_{11.2}$	$44.97\pm^{2.99}_{2.74}$	$43.93\pm^{3.08}_{2.59}$	$1.02\pm^{0.05}_{0.05}$
23	$170.5\pm^{8.6}_{8.1}$	$3.89\pm^{0.30}_{0.29}$	$3.93\pm^{0.49}_{0.42}$	$0.98\pm^{0.10}_{0.09}$
24	$297.6\pm^{13.1}_{13.6}$	$10.78\pm^{1.04}_{1.02}$	$10.48\pm^{1.34}_{1.22}$	$1.04\pm^{0.10}_{0.11}$
25	$256.2\pm^{13.4}_{14.4}$	$7.27\pm^{0.81}_{0.77}$	$7.37\pm^{0.97}_{0.95}$	$0.99\pm^{0.09}_{0.09}$

This type of uncertainty is usually referred to as *statistical error*. Many experiments and measurements are also subject to other sources of uncertainty that may not be explicitly reported in the dataset. For example, the measurement of events recorded by a detector is affected by the calibration of the detector, and a systematic offset in the calibration would affect the numbers recorded. In the case of the data of Table 6.1, the uncertainty due to the calibration of the detector is likely to affect by the same amount of all measurements, regardless of the precision indicated by the statistical error. This type of uncertainty is typically referred to as *systematic error*, and the inclusion of such additional source of uncertainty would modify the value of the weighted mean. As an example of this effect, if we add an error of ± 0.1 to all values of the ratio of Table 6.1, the weighted mean becomes 0.95 ± 0.04 (see Problem 6.2). It is clear that the addition of a constant error for each measurement causes a de-weighting of datapoints with small statistical errors, and in the limit of a large systematic error the weighted mean becomes the linear average. Therefore, the linear average can be used when the data analyst wants to weigh equally all datapoints, regardless of the precision indicated by the statistical errors. Systematic errors are discussed in more detail in Chap. 11.

6.2 The Median

Another quantity that can be calculated from the N measurements is the median, defined in Sect. 2.3.1 as the value of the variable that is greater than 50 % of the measurements, and also lower than 50 % of the measurements. In the case of the measurement of the ratios in Table 6.1, this is simply obtained by ordering the 25 measurements in ascending order, and using the 13th measurement as an approximation for the median. The value obtained in this case is 1.02, quite close to the value of the linear average, since both statistics do not take into account the measurement errors.

One useful feature of the median is that it is not very sensitive to "outliers" in the distribution. For example, if one of the measurements was erroneously reported as 0.07 ± 0.01 (instead of 0.72 ± 0.11, such as source 18 in the Table), both linear and weighted averages would be affected by the error, but the median would not. The median may therefore be an appropriate value to report in cases where the analyst suspects the presence of outliers in the dataset.

6.3 The Logarithmic Average and Fractional or Multiplicative Errors

The quantity "Ratio" in Table 6.1 can be used to illustrate a type of variables that may require a special attention when calculating their averages. Consider a variable whose errors are proportional to their measured values. In this case, a weighted average will be skewed towards *lower* values because of the smaller errors in those measurements. The question we want address is whether it is appropriate to use a weighted average of these measurements or whether one should use a different approach.

To illustrate this situation, let's use two measurements such as $x_1 = 1.2 \pm 0.24$ and $x_2 = 0.80 \pm 0.16$. Both measurements have a relative error of 20 %, the linear average is 1.00 and the weighted average is 0.923. The base-10 logarithm of these measurements are $\log x_1 = 0.0792$ and $\log x_2 = -0.0969$, with the same error. In fact, using the error propagation method (Sect. 4.7.6), the error in the logarithm is proportional to the fractional error according to

$$\sigma_{\log x} = \frac{\sigma_x}{x} \frac{1}{\ln 10}. \tag{6.1}$$

For our measurements, this equation gives a value of $\sigma_{\log x} = 0.087$ for both measurements. The weighted average of these logarithms is therefore the linear average $\overline{\log x} = -0.0088$, leading to an average of $\overline{x} = 0.980$. This value is much closer to the linear average of 1.00 than to the weighted average.

Errors that are exactly proportional to the measurement, or

$$\sigma_x = x\sigma_r \tag{6.2}$$

may be called *fractional* or *multiplicative errors*. The quantity σ_r is the relative error and it remains constant for purely multiplicative errors. In most cases, including that of Table 6.1, the relative error σ_x/x varies among the measurements, and therefore (6.2) applies only as an approximation. In the following we investigate when it is in fact advisable to use the logarithm of measurements, instead of the measurements themselves, to obtain a more accurate determination of the mean of a variable that has multiplicative errors.

6.3.1 The Weighted Logarithmic Average

The maximum likelihood method applied to the logarithm of measurements of a variable X can be used to estimate the mean and the error of $\log X$. The weighted logarithmic average of N measurements x_i is defined as

$$\overline{\log x} = \frac{\sum_{i=1}^{N} \dfrac{\log x_i}{\sigma^2_{\log x_i}}}{\sum_{i=1}^{N} \dfrac{1}{\sigma^2_{\log x_i}}} \tag{6.3}$$

where $\sigma^2_{\log x_i}$ is the variance of the logarithm of the measurements, which can be obtained from (6.1). The uncertainty in the weighted logarithmic average is given by

$$\sigma^2_{\log x} = \frac{1}{\sum_{i=1}^{N} \dfrac{1}{\sigma^2_{\log x_i}}}. \tag{6.4}$$

The use of this logarithmic average is justified when the variable X has a log-normal distribution, i.e., when $\log X$ has a Gaussian distribution, rather than the variable X itself. An example of a log-normal variable is illustrated in Fig. 6.1. In this case, the maximum likelihood method estimator of the mean of $\log X$ is the logarithmic mean of (6.3). Clearly, a variable can only be log-normal when the variable has positive values, such as the ratio of two positive quantities. The

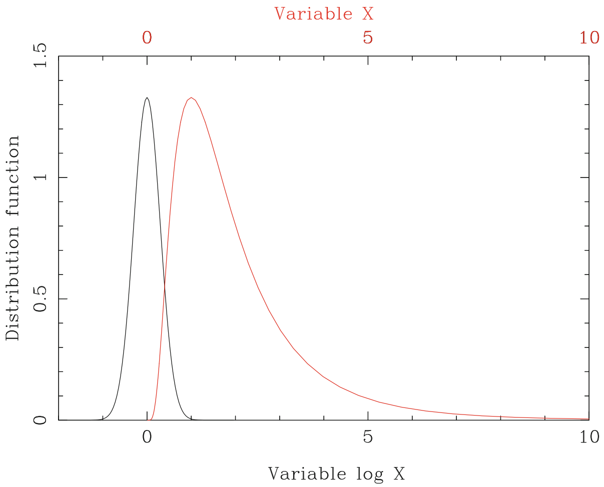

Fig. 6.1 Log-normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 0.3$ (*black line*) and linear plot of the same distribution (*red line*). A heavier right-hand tail in the linear plot may be indicative of a log-normal distribution

determination of the log-normal shape can be made if one has available random samples from its distribution.

In the limit of measurements with the same fractional error and small deviations from the mean μ, the weighted logarithmic average is equivalent to the linear average.

Proof This can be shown by proving that

$$\overline{\log x} = \log \overline{x}$$

where $\overline{x}$ is the ordinary linear average. Notice that $\log x$ in (6.3) is a base-10 logarithm. In this proof we make use the base-e logarithm ($\ln x$), the two are related by

$$\log x = \ln x / \ln 10.$$

Consider N measurements x_i in the neighborhood of the mean μ of the random variable, $x_i = \mu + \Delta x_i$. A Taylor series expansion yields

$$\ln x_i = \ln \mu (1 + \frac{\Delta x_i}{\mu}) = \ln \mu + \frac{\Delta x_i}{\mu} - \frac{(\Delta x_i/\mu)^2}{2} + \ldots$$

If the deviation is $\Delta x_i \ll \mu$, one can neglect terms of the second order and higher. The average of the logarithms of the N measurements can thus be approximated as

$$\overline{\log x} = \frac{1}{N}\sum_{i=1}^{N} \log x_i \simeq \frac{1}{\ln 10}\left(\ln \mu + \frac{1}{N}\sum_{i=1}^{N}\frac{\Delta x_i}{\mu}\right).$$

On the other hand, the logarithm of the mean $\overline{x}$ is

$$\log \overline{x} = \log \frac{1}{N}\sum_{i=1}^{N} x_i = \log\left(\frac{1}{N}\sum_{i=1}^{N}\mu(1+\frac{\Delta x_i}{\mu})\right).$$

This leads to

$$\log \overline{x} = \frac{1}{\ln 10}\left(\ln \mu + \ln\left(1 + \frac{1}{N}\sum_{i=1}^{N}\frac{\Delta x_i}{\mu}\right)\right) \simeq$$
$$\frac{1}{\ln 10}\left(\ln \mu + \frac{1}{N}\sum_{i=1}^{N}\frac{\Delta x_i}{\mu}\right) = \overline{\log x}$$

where we retained only the first-order term in the Taylor series expansion of the logarithm since $\sum \Delta x_i/\mu \ll N$. □

As discussed earlier in this section, the logarithmic average is an appropriate quantity for log-normal distributed variables. The results of this section show that this average is closer to the linear average of the measurements than the standard weighted average, when measurement errors are positively correlated to the measurements themselves.

Example 6.1 The data of Table 6.1 can be used to calculate the logarithmic average of the column "Ratio" according to (6.3) and (6.4) as $\overline{\log x} = -0.023 \pm 0.018$. These quantities can be converted easily to linear quantities taking into account the error propagation formula $\sigma_{\log x} = \sigma/(x \ln 10)$, to obtain a value of 0.95 ± 0.04.

Notice how the logarithmic mean has a value that is somewhat between that of the linear average $\overline{x} = 1.01$ and the traditional weighted average of 0.90 ± 0.02. It should not be surprising that the logarithmic mean is not exactly equal to the linear average. In fact, the measurements of Table 6.1 have different relative errors. Only in the case of identical relative errors for all measurements we expect that the two averages have the same value. ◇

6.3.2 *The Relative-Error Weighted Average*

Although transforming measurements to their logarithms is a simple procedure, we also want to investigate another type of average that deals directly with the measurements without the need to calculate their logarithms.

We introduce the *relative-error weighted average* as

$$\overline{x_{RE}} = \frac{\sum_{i=1}^{N} x_i/(\sigma_i/x_i)^2}{\sum_{i=1}^{N} 1/(\sigma_i/x_i)^2}. \tag{6.5}$$

The only difference with the weighted mean defined in Sect. 5.1.3 is the use of the extra factor of x_i in the error term, so that σ_i/x_i is the relative error of each measurement.

The reason to introduce this new average is that, for log-normal variables, this relative-error weighted mean is equivalent to the logarithmic mean of (6.3). This can be proven by showing that $\overline{\ln x} = \ln \overline{x_{RE}}$.

Proof Start with the logarithm of the relative-error weighted average,

$$\ln \overline{x_{RE}} = \ln \left(\frac{\sum_{i=1}^{N} x_i/(\sigma_i/x_i)^2}{\sum_{i=1}^{N} 1/(\sigma_i/x_i)^2} \right) = \ln \left(\frac{\sum_{i=1}^{N} x_i/\sigma^2_{\log x_i}}{\sum_{i=1}^{N} 1/\sigma^2_{\log x_i}} \right).$$

From this, expand the measurement term $x_i = \mu + \Delta x_i$, where μ is the parent mean of the variable X,

$$\ln \left(\mu + \frac{\sum_{i=1}^{N} \Delta x_i/\sigma^2_{\log x_i}}{\sum_{i=1}^{N} 1/\sigma^2_{\log x_i}} \right) = \ln \mu + \ln \left(1 + \frac{\sum_{i=1}^{N} \Delta x_i/(\mu\sigma^2_{\log x_i})}{\sum_{i=1}^{N} 1/\sigma^2_{\log x_i}} \right).$$

If $\Delta x_i \ll \mu$, then

$$\ln \overline{x_{RE}} = \ln \mu + \frac{\sum_{i=1}^{N} \Delta x_i/(\mu\sigma^2_{\log x_i})}{\sum_{i=1}^{N} 1/\sigma^2_{\log x_i}}$$

leading to

$$\log \overline{x_{RE}} = \log \mu + \frac{1}{\ln 10} \frac{\sum_{i=1}^{N} \Delta x_i/(\mu\sigma^2_{\log x_i})}{\sum_{i=1}^{N} 1/\sigma^2_{\log x_i}}.$$

The logarithmic average can also be expanded making use of

$$\sum_{i=1}^{N} \frac{\log x_i}{\sigma^2_{\log x_i}} = \sum_{i=1}^{N} \frac{\log \mu + \log(1 + \Delta x_i)/\mu}{\sigma^2_{\log x_i}} \simeq \sum_{i=1}^{N} \left(\frac{\log \mu}{\sigma^2_{\log x_i}} + \frac{\Delta x_i/\mu}{\sigma^2_{\log x_i} \ln 10} \right).$$

This leads to

$$\overline{\log x} = \log \mu + \frac{1}{\ln 10} \frac{\sum_{i=1}^{N} \Delta x_i / (\mu \sigma^2_{\log x_i})}{\sum_{i=1}^{N} 1/\sigma^2_{\log x_i}} = \log \overline{x_{RE}}.$$

□

The use of the relative-error weighted average should be viewed as an ad hoc method to obtain an average value that is consistent with the logarithmic average, especially in the limit measurements with equal relative errors. The statistical uncertainty in this error-weighted average can be simply assigned as the error in the traditional weighted average (5.8). In fact, the statistical error should be determined by the "physical" uncertainties in the measurements, as is the case for the variance in (5.8). It would be tempting to use the inverse of the denominator of (6.5) as the variance; however, the result would be biased by our somewhat arbitrary choice of weighing the measurements by the relative errors, instead of the error themselves.

Example 6.2 Continuing with the values of "Ratio" in Table 6.1, the error-weighted average is calculated as $\overline{x_{RE}} = 0.96$. The error in the traditional weighted average was 0.02, therefore we may report the result as 0.96 ± 0.02. Comparison with the values of 0.95 ± 0.04 for the logarithmic average shows the general agreement between these two values.

◇

Summary of Key Concepts for this Chapter

- □ *Linear average*: The mean $\overline{x}$ of N measurements.
- □ *Median*: The 50 % quantile, or the number below and above which there are 50 % of the variable's values.
- □ *Logarithmic average*: In some cases (e.g., when errors are proportional to the measured values) it is meaningful to calulate the weighted average of the logarithm of the variable,

$$\begin{cases} \overline{\log x} = \dfrac{\sum \log x_i/\sigma^2_{logx_i}}{\sum 1/\sigma^2_{logx_i}} \\ \sigma^2_{\log x} = \dfrac{1}{\sigma^2_{logx_i}} \end{cases}$$

  where $\sigma_{logx_i} = \sigma_i/(x_i \ln 2)$.
- □ *Relative-error weighted average*: An approximation of the logarithmic average that does not require logarithms,

$$\overline{x_{RE}} = \frac{\sum x_i/(\sigma_i/x_i)^2}{\sum 1/(\sigma_i/x_i)^2}.$$

Problems

6.1 Calculate the linear average and the weighted mean of the quantity "Ratio" in Table 6.1.

6.2 Consider the 25 measurements of "Ratio" in Table 6.1. Assume that an additional uncertainty of ± 0.1 is to be added linearly to the statistical error of each measurement reported in the table. Show that the addition of this source of uncertainty results in a weighted mean of 0.95 ± 0.04.

6.3 Given two measurements x_1 and x_2 with values in the neighborhood of 1.0, show that the logarithm of the average of the measurements is approximately equal to the average of the logarithms of the measurements.

6.4 Given two measurements x_1 and x_2 with values in the neighborhood of a positive number A, show that the logarithm of the average of the measurements is approximately equal to the average of the logarithms of the measurements.

6.5 For the data in Table 6.1, calculate the linear average, weighted average and median of each quantity (Radius, Energy Method 1, Energy Method 2 and Ratio). You may assume that the error of each measurements is the average of the asymmetric errors of each measurement reported in the table.

6.6 Table 6.1 contains the measurement of the thermal energy of certain sources using two independent methods labeled as method #1 and method #2. For each source, the measurement is made at a given radius, which varies from source to source. The error bars indicate the 68 %, or 1σ, confidence intervals; the fact that most are asymmetric indicate that the measurements do not follow exactly a Gaussian distribution. Calculate the weighted mean of the ratios between the two measurements and its standard deviation, assuming that the errors are Gaussian and equal to the average of the asymmetric errors, as it is often done in this type of situation.