2 Estimation and uncertainty

In this chapter we look at estimators and how to use them to characterize a distribution. Relevant to this are the concepts of estimator bias, consistency, and efficiency. I shall discuss measurement models and measurement uncertainty and note the difference between distribution properties and estimates thereof. We will learn about the central limit theorem, how and when we can reduce errors through repeated measurements, and we will see how we can propagate uncertainties.

2.1 Estimators

An estimator is something that characterizes a set of data. This is often done in order to characterize the *parent distribution*, the distribution the data were drawn from. We often distinguish between a point estimator, which is a single value such as the mean or mode, and an interval estimator, which characterizes a range, such as the standard deviation or interquartile range. What we use as an estimator depends on what we want to characterise and what we know about the parent distribution.

Suppose we want to learn about the distribution of the heights of a particular species of tree, given a sample $\{x\}$ of the heights of N such trees in a forest. Specifically, we would like to estimate the mean of the parent distribution, which here means the set of all trees in the forest at this time (as this is finite it is sometimes called the *parent population*). How useful any estimate is depends on how the sample of trees was selected: were they taken from across the entire forest rather than in a particularly shady region? Were the tallest trees missed because they were harder to measure (or have already been cut down)? But let us assume that we have a sample that is representative of the forest in some useful sense. There are potentially many ways we could use this sample to estimate the mean of the parent distribution. We could

(1) sum all x and divide by N (arithmetic mean)
(2) sum the first K measurements ($K < N$) and divide by K
(3) sum all x and divide by $N - 1$
(4) average the smallest and largest values of x (midrange)
(5) sort all x and take the middle value; average the two middle ones if N is even (median)
(6) bin the data and take the most frequent (mode)
(7) disregard the data and report the value as 8 m (justified by a model for tree growth, perhaps).

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

All of these are estimators. But not all of them are good estimators. Ideally we want estimators that are *unbiased*, *consistent*, and *efficient*.

An estimator is said to be *unbiased* if the expectation value of the estimator is equal to the true value. As the name suggests, the expectation value is what we expect the result to be if we had the entire parent population and no uncertainty in the measurements (but see the formal definition in section 1.3). We don't demand that the estimator give the true value; that is generally unachievable. But an unbiased estimate is equal to the true value in some average sense: it is neither systematically too high nor too low. Sometimes our samples are not drawn from a discrete population, as they are in the forest example, but are measurements of a continuous distribution, such as when we measure the temperature in different parts of a lake. Here the parent population is effectively infinite in size, and the expectation value is given by an integral (equation 1.21).

An estimator is *consistent* if its value is expected to converge on the true value in the limit of a large amount of data. This is desirable because it means that as we gather more data we get an estimate closer to the truth.

Efficiency measures the variance of an estimator and is a relative term. The smaller the variance the more efficient it is.

Whether a particular estimator is unbiased and consistent, and how efficient it is, depends on the parent distribution. If the parent distribution is Gaussian, then of the estimators listed above, numbers 1, 4, and 5 are unbiased. Number 2 will be unbiased if the ordering is random. Number 3 is biased because dividing by $N-1$ overestimates the mean. Number 7 is biased, unless it happens to be the true value. Whether number 6 is biased or not depends on how we binned the data. Estimator 1 is consistent. So is 3, because there is no difference between N and $N-1$ in the limit $N \to \infty$. As the mean, median, and mode are all equal for a Gaussian, estimators 5 and 6 are consistent (provided we used ever narrower bins for the mode calculation), as is number 4. Estimators 2 and 7, on the other hand, are inconsistent because they do not change as we gather more data. Note that an estimator could be unbiased but inconsistent (number 2) or it could be biased but consistent (number 3). Estimator 1 has a smaller variance than estimator 2 because it is using more of the data. Number 4 likewise has a large variance for a Gaussian distribution. Note that despite its otherwise dubious properties, estimator 7 has the smallest variance: zero.

Sometimes the most efficient estimator is biased, or an unbiased estimator may not be very efficient. Which is the "best" estimator depends on your goals and on the parent distribution. As we shall see in section 4.8 (where we shall also define bias precisely) the error of an estimator can be written in terms of its bias and its variance.

2.2 Noise, measurement errors, and measurement models

Uncertainty arises in a number of ways when dealing with data. First, measurements cannot be made with arbitrary precision. When measuring the length of something with a ruler, the precision is limited to some fraction of the difference between two neighbouring

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

tick marks. It is further limited by how well these tick marks were machined in the first place. Second, there are invariably sources of interference – noise – when measuring. If we are trying to weigh a very light object by hanging it from a spring, air currents and mechanical vibrations will perturb the spring, preventing us from measuring a constant value. Third, many real-world processes are stochastic. For example, the number of photons received from a source in a given time interval is not deterministic, but is described by a Poisson distribution (section 1.4.2). In two different equal-length time intervals we will not necessarily receive the same number of photons. This is an intrinsic uncertainty which is unrelated to the measurement process.

All of this means that what we measure is not the true value of something, but rather an estimate of it. A useful way to think of this is to consider that the measurement process has a probability distribution, which we call the *measurement model*. This gives the probability of measuring some particular value of the data, given some parameters of the process and the measurement procedure. With a Poisson process, for example, if the expected number of photons received in some time interval is λ, then the measurement model is the Poisson distribution $P(r\,|\,\lambda)$. This gives the probability of measuring r when the expected (but unknown) value is λ. The measurement model is also called a *noise model*.

Measurement models are fundamental to inference and are usually described with the likelihood, a concept I will introduce in the next chapter.

A common measurement model is the Gaussian distribution (section 1.4.5). It is described by two parameters, the mean and the standard deviation. If we are given a measurement a and told it has a Gaussian uncertainty $\pm b$, we interpret this to mean that the measurement has a Gaussian distribution with mean a and standard deviation b. How should we understand this, given that we only have one measurement? From a frequency perspective it means that if we could somehow make an infinite number of (noisy) measurements, then we expect these measurements to show this distribution, which would have (for example) 68% of the measurements lying in the range $(a-b, a+b)$. But as we only made the measurement once, what we really mean is that if the true mean is a then there is a probability of 0.68 that a measurement will lie in the range $(a-b, a+b)$.

Note that we have done something subtle but important in this interpretation. We have taken the measurement a as an estimator of the mean of the signal and then used this to make a statement about the measurement process. That is, we have assumed that the measurements are described by a Gaussian, the standard deviation of which is known, and have then estimated its mean from this single measurement. You would not do this if you had good reason to believe that a is not a good estimator, e.g. if you think it is an outlier. In general, finding the mean – or indeed any other parameter of a distribution (such as the standard deviation) – is a process of inference: we use a set of data to make estimates of the values of the parameters. We will look at how to do this in some detail in later chapters, for example in section 3.5, as well as in section 4.4.5 for a Poisson distribution and in section 6.2 for a Gaussian distribution. In fact, the idea that we cannot measure parameters directly but can only infer them from data will appear frequently in later chapters.

The important point for now is that a measurement plus error bar can be thought of as summarizing a probability distribution that describes the measurement process. This works because we often adopt standard probability distributions as measurement models,

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

and many of these standard distributions, like the Gaussian, can be summarized by one point estimate and one width parameter.

2.3 The central limit theorem and $\sqrt{N}$ reduction from repeated measurements

Suppose we have N independent random variables $\{x_i\}$ for $i = 1, 2, \ldots N$, each drawn from a distribution with finite mean μ_i and finite variance V_i. Let

$$y = \sum_i x_i. \tag{2.1}$$

It follows from the definition of expectation (section 1.3) that

$$E[y] = \sum_i \mu_i \tag{2.2}$$

$$\mathrm{Var}(y) = \sum_i V_i \tag{2.3}$$

i.e. the expected mean is the sum of the means and the (expected) variance is the sum of the variances. It can further be proven that in the limit of very large N the distribution of y becomes a Gaussian. This is the *central limit theorem* (CLT). It essentially says that if we add together lots of independent variables, their distribution becomes a Gaussian with mean and variance as specified above.[1] If the variables were not independent then the expression for $\mathrm{Var}(y)$ would include additional terms to account for the covariance (section 1.6.1).

The central limit theorem is of great practical value because it means that if a measured quantity is affected by many independent noise sources, i.e. it is the sum of lots of independent random variables, then the measurement can be described with a Gaussian distribution. This simplifies many analyses and explains why the Gaussian distribution is so widespread. Note that there is no requirement that the original distributions be Gaussian or even identical. They just have to be independent and to have finite means and variances.

If we now define z as the mean of the N independent variables ($z = y/N$), it follows directly from the above that

$$E[z] = \frac{1}{N}\sum_{i=1}^{N} \mu_i \tag{2.4}$$

$$\mathrm{Var}(z) = \frac{1}{N^2}\sum_{i=1}^{N} V_i \quad \text{or} \quad \sigma_z = \frac{1}{N}\sqrt{\sum_{i=1}^{N} \sigma_i^2} \tag{2.5}$$

where $\sigma_z = \sqrt{\mathrm{Var}(z)}$ is the standard deviation in z and $\sigma_i = \sqrt{V_i}$. Why do we have

[1] One could argue that in the limit $N \to \infty$ the mean and variance of y would become infinite and so it makes little sense to talk of a distribution at all. But we need not be so pedantic when we realise that in practice the central limit theorem is a very good approximation even for relatively small N, as we will see in section 2.3.1.

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

a $1/N^2$ factor rather than a $1/N$ factor in the variance of z? It follows directly from the definition of variance:

$$\begin{aligned}\mathrm{Var}(z) &= E[(z - E[z])^2] \\ &= E\left[\left(\frac{y}{N} - E\left[\frac{y}{N}\right]\right)^2\right] \\ &= \frac{1}{N^2}E[(y - E[y])^2] \\ &= \frac{1}{N^2}\mathrm{Var}(y) \\ &= \frac{1}{N^2}\sum_{i=1}^{N} V_i. \end{aligned} \tag{2.6}$$

If all of the $\{x_i\}$ come from a common distribution with mean μ and variance σ^2, then setting $\mu_i = \mu$ and $\sigma_i = \sigma$ for all i, it follows that

$$E[z] = \frac{1}{N}N\mu = \mu \tag{2.7}$$

$$\mathrm{Var}(z) = \frac{1}{N^2}N\sigma^2 = \frac{\sigma^2}{N} \quad \text{or} \quad \sigma_z = \frac{\sigma}{\sqrt{N}}. \tag{2.8}$$

This result is of great practical relevance. It says that if we make N independent measurements of a quantity, each of which has the same uncertainty (standard deviation) of σ, then if we estimate the quantity by the mean of our measurements $\overline{x}$, this has an uncertainty (standard deviation) of $\sigma/\sqrt{N}$. In other words, repeated measurements reduce uncertainty. This fact underpins much of the scientific methodology for improving the precision of measurements.

2.3.1 Demonstration of the central limit theorem

We can demonstrate the central limit theorem using real numbers drawn from a uniform distribution $\mathcal{U}(0, 10)$ (this is very non-Gaussian!). We draw 10^4 numbers and plot their distribution using a histogram with the following.

```
hist(runif(1e4)*10)
```

This distribution has an expected mean of 5 and expected standard deviation of $(10 - 0)/\sqrt{12} = 2.89$ (see section 1.4.4). If the central limit theorem applies, then the distribution of the average of N numbers drawn from this should tend towards a Gaussian as N gets large, with mean 5 and standard deviation $2.89/\sqrt{N}$.

To discover the distribution empirically we draw N numbers from this distribution at random and calculate their mean. We repeat this many times and plot their distribution as a histogram. This is done by the following R code for $N = 1, 2, 5$ with the results shown in figure 2.1. As N gets larger, the distribution looks more like a Gaussian, the mean gets nearer to 5, and the standard deviation gets smaller. We see from the plot that the approximation is quite good even for small N. The computed standard deviations of the three distributions are 2.859, 2.050, 1.299 for $N = 1, 2, 5$ respectively. This compares

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

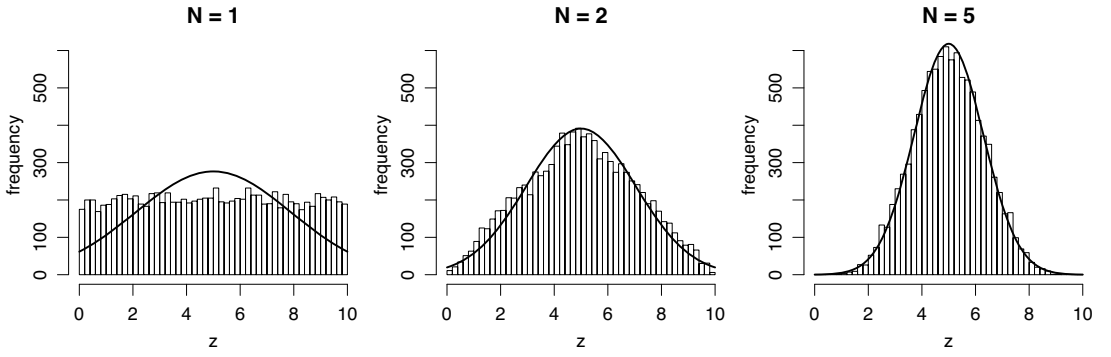

Fig. 2.1 Each histogram is the distribution of the mean of N numbers drawn from the uniform distribution $\mathcal{U}(0, 10)$. The smooth line is the Gaussian limit for the distribution expected from the central limit theorem.

to the standard deviation from the central limit theorem of 2.887, 2.041, 1.291 respectively. Note that because `dnorm` returns a normalized Gaussian (the area under the curve is unity), but I have plotted the data as a frequency, I have to scale the former to have the same total area as the histogram. This area is equal to the number of data points (`Nsamp`) multiplied by the histogram bin width (`binwidth`), which is 1000 in the example above. Alternatively I could have used used the option `freq=FALSE` in `hist` to plot the histogram as a probability density.

R file: `CLT_uniform.R`

```
##### Apply central limit theorem to draws from uniform distribution

pdf("CLT_uniform.pdf", 12, 4)
par(mfrow=c(1,3), mgp=c(2,0.8,0), mar=c(3.5,3.5,1,0), oma=0.1*c(1,1,5,5),
    cex=1.2)
set.seed(200)
for(N in c(1,2,5)) {
  Nsamp <- 10000
  z <- numeric(Nsamp)
  for (i in 1:Nsamp){
    z[i] <- mean(runif(N)*10)
  }
  binwidth <- 0.2
  hist(z, xlim=c(0,10), ylim=c(0,640), breaks=seq(0,10,binwidth),
       ylab="frequency", main=paste("N =",N))
  cat("mean, sd = ", mean(z), sd(z), "\n")
  # overplot Gaussian with mean and sd from CLT
  gsd <- (10/sqrt(12))*1/sqrt(N)
  x <- seq(0,10,0.01)
  y <- dnorm(x, mean=5, sd=gsd)*Nsamp*binwidth
  lines(x, y, lwd=2)
}
dev.off()
```

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

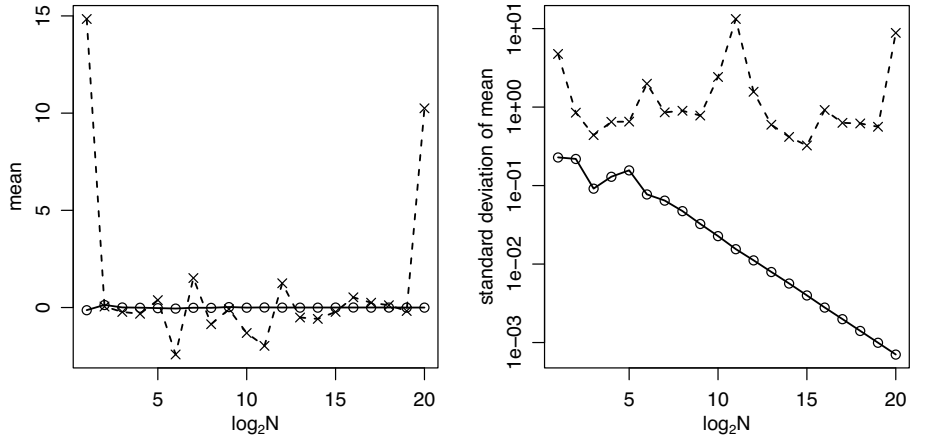

Fig. 2.2 The variation of the mean (left) and standard deviation of the mean (right) for N variables drawn from a Cauchy distribution (crosses and dashed line) and a Gaussian distribution (circles and solid line). Note the log scales.

2.3.2 A counter-example: estimating the mean of the Cauchy distribution

The central limit theorem is valid only if its assumptions are met. These are that the variables are independent and drawn from a distribution (or distributions) with finite mean and variance. Here is an example where not all of these assumptions are met.

We would like to estimate the location parameter a of the Cauchy distribution (section 1.4.7). The PDF is

$$P(x) = \frac{b}{\pi[b^2 + (x-a)^2]}. \tag{2.9}$$

The following R script draws N numbers at random, and independently, from a Cauchy distribution with $a = 0$ and unit half-width at half-maximum (HWHM), $b = 1$. As the distribution is symmetric we might expect the mean of these samples to be a good estimator of a. We investigate this by plotting this mean, as well as the standard deviation of this mean, as a function of N. These are shown in the left and right panels, respectively, of figure 2.2 using dashed lines and crosses. I calculate the standard deviation of the mean according to the central limit theorem, which tells us it is $\sigma/\sqrt{N}$ (equation 2.8), where σ is the sample standard deviation (computed using `sd` in R). For comparison I compute and plot the same quantities for samples drawn from a Gaussian distribution with the same mode and HWHM (plotted as solid lines and circles).

We see that the Gaussian distribution does what we expect from the central limit theorem. The mean converges (very quickly) towards its true value, and the standard deviation of the mean (also called the standard error in the mean) decreases[2] as $1/\sqrt{N}$. For the

[2] The right panel plots the log base 10 of the standard deviation of the mean σ_{mean} against $\log_2 N$. According

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

Cauchy distribution the mean is also close to zero, but it oscillates a lot. This is because the sample standard deviation does not decrease with increasing N, as can be seen in the right panel.

The reason for this behaviour of the Cauchy-drawn samples is that the central limit theorem does not apply to the Cauchy distribution. This is because the central limit theorem assumes finite values of the mean and variance of the distribution from which the data are drawn. Recall that the definition of the variance of a distribution $P(x)$ is (equation 1.30)

$$\begin{aligned} \mathrm{Var}(x) &= E[x^2] - E[x]^2 \\ &= \int x^2 P(x)\,dx - \left(\int x\,P(x)\,dx\right)^2 . \end{aligned} \tag{2.10}$$

Thus if $P(x)$ does not drop off faster than $1/x^2$ with increasing $|x|$, the first interval will not converge. The Cauchy distribution drops off exactly as $1/x^2$ for large $|x|$, so its variance is infinite. In fact the mean – the second integral above – is not defined (and neither are the higher central moments) so the variance is not even defined. This is why we are careful to refer to a as the location parameter of the Cauchy distribution, and not as the mean.

Given a set of data drawn from a Cauchy distribution, we can better estimate the location parameter a and the HWHM b using the median and inter-quartile range, respectively.

R file: `cauchy_mean_estimation.R`

```
##### Is the mean a convergent estimator for the location of a Cauchy?

# Calculate mean and CLT prediction of its standard deviation, of a sample
# of numbers drawn from (i) Cauchy, (ii) Gaussian, both with mode=0, HWHM=1,
# with 2^lognsamp[i] no. of samples. Plot these against lognsamp[i]

set.seed(100)
lognsamp <- 1:20
GaussMean  <- vector(length=length(lognsamp))
CauchyMean <- vector(length=length(lognsamp))
GaussSD    <- vector(length=length(lognsamp))
CauchySD   <- vector(length=length(lognsamp))
for(i in 1:length(lognsamp)) {
  s <- rnorm(2^lognsamp[i], mean=0, sd=1/(2*log(2))) # sd=HWHM/2ln2=0.721
  GaussMean[i]  <- mean(s)
  GaussSD[i]    <- sd(s)/sqrt(length(s)) # standard deviation in mean
  s <- rcauchy(2^lognsamp[i], location=0, scale=1) # scale=HWHM
  CauchyMean[i] <- mean(s)
  CauchySD[i]   <- sd(s)/sqrt(length(s)) # standard deviation in mean
}

pdf("cauchy_mean_estimation.pdf", 8, 4)
par(mfrow=c(1,2), mgp=c(2.0,0.8,0), mar=c(3.5,3.5,1,1), oma=0.1*c(1,1,1,1))
plot(  lognsamp, GaussMean, ylim=range(c(GaussMean, CauchyMean)), type="n",
  xlab=expression(paste(log[2], N)), ylab="mean")
lines( lognsamp, GaussMean,  lwd=1.5)
```

to the central limit theorem $\sigma_{\mathrm{mean}} \propto 1/\sqrt{N}$. Therefore $\log \sigma_{\mathrm{mean}} \propto -(\log_2 N)/(2\log_2 10)$, i.e. the variation should be linear with a gradient of -0.151. Ignoring the few wiggles at low N due to small number statistics, you can confirm that this is indeed the gradient of the solid line.

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

```
points(lognsamp, GaussMean)
lines( lognsamp, CauchyMean, lwd=1.5, lty=2)
points(lognsamp, CauchyMean, pch=4)
plot(  lognsamp, GaussMean, ylim=range(c(GaussSD, CauchySD)), log="y",
       type="n", xlab=expression(paste(log[2], N)),
       ylab="standard deviation of mean")
lines( lognsamp, GaussSD,  lwd=1.5)
points(lognsamp, GaussSD)
lines( lognsamp, CauchySD, lwd=1.5, lty=2)
points(lognsamp, CauchySD, pch=4)
dev.off()
cbind(lognsamp, GaussSD, CauchySD)
```

2.4 Population properties vs sample estimates

It is important to understand the distinction between the variance of a distribution and the variance of an estimate of a property of that distribution, such as the mean. Assume that we have a set of N quantities $\{x_i\}$ drawn from a distribution with unknown mean and standard deviation. As already introduced in section 1.3, we can *estimate* these from the data by computing

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{2.11}$$

$$\hat{\sigma} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2}. \tag{2.12}$$

These are the *sample mean* and *sample standard deviation* respectively.[3] Note that the order of magnitude of the sample mean and standard deviation do not change as N increases: in equation 2.12 $(x_i - \overline{x})$ will have some typical size, call it δx, which is independent of N, so $\hat{\sigma} \sim \delta x\sqrt{N/(N-1)} \to \delta x$ in the limit $N \to \infty$. This is of course what we want, namely a *consistent* estimator.

According to the central limit theorem, if we draw N numbers from a distribution which has a population (true) standard deviation σ, and average these to get the sample mean $\overline{x}$, then this estimate of the mean has a standard deviation $\sigma/\sqrt{N}$ (equation 2.8).[4] As we don't usually know σ in advance, we estimate it from the data with $\hat{\sigma}$ (equation 2.12). In that case our estimate of the uncertainty in $\overline{x}$ is $\hat{\sigma}/\sqrt{N}$, which is the *standard deviation in the mean*

$$\mathrm{SEM} = \frac{\hat{\sigma}}{\sqrt{N}} = \sqrt{\frac{1}{N(N-1)}\sum_{i=1}^{N}(x_i - \overline{x})^2}. \tag{2.13}$$

[3] As noted previously, the presence of $N-1$ rather than N in the second equation can be understood as correcting for the sample bias.

[4] You can equally well replace the statement "draw N numbers from a distribution" with "make N measurements".

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

This is sometimes called the *standard error in the mean* instead. It is the precision of our estimate $\overline{x}$, and it clearly decreases as N gets larger. That is, our *inference* of the value of $\overline{x}$ gets more precise as we take more data. We will see in section 6.2 that this result is a direct consequence of the inference process when the data have been drawn from a Gaussian.

Naturally our estimate of the standard deviation (equation 2.12) also gets more precise as we take more data, and it can be shown that the variance in $\hat{\sigma}$ is approximately $\hat{\sigma}^2/[2(N-1)]$, i.e. its standard deviation is approximately $\hat{\sigma}/\sqrt{2(N-1)}$. The idea of a standard deviation in a standard deviation may seem strange at first, but it's perfectly okay to ask what the standard deviation ("uncertainty") is in any quantity estimated from data.

Some books, publications, and computer programs do not distinguish carefully between "sample standard deviation" and "standard deviation". Some just define the latter using $1/N$ and then either apply $N/(N-1)$ corrections to reach the former or just say N is large so it doesn't matter. You have been warned! The R function `sd(x)` gives the *sample* standard deviation of the vector of values `x` (equation 2.12), which of course must estimate the mean from the data given.

2.5 The mean is not necessarily the most efficient estimator

I stated in section 2.1 that a good estimator is one that is unbiased, consistent, and efficient. Achieving all of these simultaneously can be impossible in practice, and there is no universal estimator – such as the mean, median, or mode for characterizing the central value of a distribution – which one can single out as always being "optimal". What makes a good estimator depends on the probability density function the data were drawn from, and often this is itself unknown and must be approximated from the data.

The central limit theorem tells us that by averaging N independent measurements of a quantity we can reduce its standard deviation by a factor of $\sqrt{N}$ compared to a single measurement (provided its assumptions are met). However, the central limit theorem does not claim that the average is the most efficient estimator, and sometimes it is not, as we will now see.

Consider the uniform distribution $\mathcal{U}(-1,+1)$. This has an expectation value $E[x]$ of zero, and a standard deviation σ of $1/\sqrt{3}$ (see section 1.4.4). We draw a sample $\{x\}$ of size N from this distribution and wish to use these to estimate $E[x]$. An obvious estimator is the mean of $\{x\}$. An alternative is the midrange, which is the average of the highest and lowest values in $\{x\}$. In the limit of infinite data we expect them both to be zero, because they are both consistent estimators. To compare their efficiencies we examine how their standard deviations vary with N. We proceed as follows.

Draw a sample of size N from the distribution and compute the two estimators, the mean and midrange. Repeat this K times (I use $K = 1000$) and compute the mean and standard

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

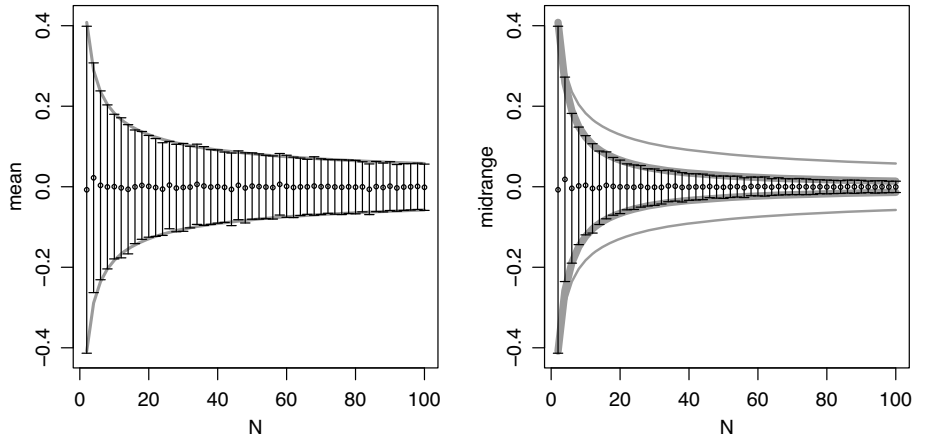

Fig. 2.3 The left panel shows the mean (circles) and standard deviation of the mean (error bars) for the mean estimator, $(1/N)\sum x_i$, of a sample of size N drawn from the uniform distribution $\mathcal{U}(-1,+1)$. The right panel shows the same but for the midrange estimator, $(x_{\max}+x_{\min})/2$. The thin grey lines (identical in both panels) show the expected standard deviation of the mean estimator, $\sigma/\sqrt{N}$ (where $\sigma=1/\sqrt{3}$ is the standard deviation of the uniform distribution). The thick grey lines in the right panel – which more or less overlap the ends of the error bars – show the expected standard deviation of the midrange estimator, $\sigma\sqrt{6/(N+2)(N+1)}$.

deviation of the two estimators over these K sets.[5] We then plot this as a point and an error bar for each estimator at this value of N, as is done in figure 2.3. We then repeat this for a range of different values of N. R code to do this is given below.

We see in figure 2.3 how the standard deviation of the midrange estimator (right panel) drops more rapidly with N than the standard deviation of the mean estimator (left panel). It can be shown that the standard deviation of the midrange estimator (for the uniform distribution) varies as $\sigma\sqrt{6/(N+2)(N+1)}$, which for large N is approximately $\sqrt{6}\,\sigma/N$. This drops with increasing N considerably faster than the standard deviation of the mean, which is $\sigma/\sqrt{N}$.

The point of this example is not to show that the central limit theorem is wrong. The central limit theorem applies to the uniform distribution, but it does not apply to the midrange estimator because it is not an average of N estimators. The point of this example is instead to show that there is nothing very special about the mean, and that estimators can have an efficiency that improves faster than $1/\sqrt{N}$. The properties that estimators have depend on the underlying distribution.

I use the function `Vectorize` in the following R code to generate an in-place version of `runif` which can accept a vector argument. I use this to generate `nsamp` ($=K$) vectors of

[5] If we used only one set for each N we would hardly see the effect we want to demonstrate, because the estimators are very noisy for small N.

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

size n ($= N$) in each cycle of the loop, which is stored in the matrix samp. The function apply then computes the mean and midrange (over n) for each of the nsamp vectors.

R file: midrange_estimator.R

```
##### Show that midrange estimator of the uniform distribution has a
##### standard error which drops as ~1/N. Cf. mean which drops as 1/sqrt(N).

library(gplots) # for plotCI
midrange <- function(x) {0.5*sum(range(x))}
set.seed(150)

nmax  <- 1e2
nsamp <- 1e3
sigma <- 2/sqrt(12) # standard deviation of U(-1,+1)
nVec  <- seq(from=2, to=nmax, by=2)
est <- matrix(NA, nrow=nmax, ncol=4) # mean(mu), mean(mr), sd(mean), sd(mr)
for(n in nVec) { # samp is a n*nsamp matrix
  samp <- Vectorize(runif, "n")(n=rep(n, nsamp), min=-1, max=1)
  mu <- apply(samp, 2, mean)     # vector size nsamp, mean estimator
  mr <- apply(samp, 2, midrange) # vector size nsamp, midrange estimator
  est[n,1] <- mean(mu)
  est[n,2] <- sd(mu)
  est[n,3] <- mean(mr)
  est[n,4] <- sd(mr)
}

pdf("midrange_estimator.pdf", 8, 4)
par(mfrow=c(1,2), mgp=c(2.0,0.8,0), mar=c(3.5,3.5,1,1), oma=0.1*c(1,1,1,1))
plot(nVec, est[nVec,1], type="n", ylim=c(-0.45, 0.45), yaxs="i",
     xlab="N", ylab="mean")
lines(nVec,  sigma*1/sqrt(nVec), col="grey60", lw=2.5)
lines(nVec, -sigma*1/sqrt(nVec), col="grey60", lw=2.5)
plotCI(nVec, est[nVec,1], uiw=est[nVec,2], gap=0, cex=0.5, add=TRUE)
plot(nVec, est[nVec,3], type="n", ylim=c(-0.45, 0.45), yaxs="i",
     xlab="N", ylab="midrange")
lines(nVec,  sigma*1/sqrt(nVec), col="grey60", lw=2)
lines(nVec, -sigma*1/sqrt(nVec), col="grey60", lw=2)
lines(nVec,  sigma*sqrt(6/((nVec+2)*(nVec+1))), col="grey60", lw=6)
lines(nVec, -sigma*sqrt(6/((nVec+2)*(nVec+1))), col="grey60", lw=6)
plotCI(nVec, est[nVec,3], uiw=est[nVec,4], gap=0, cex=0.5, add=TRUE)
dev.off()
```

2.6 Outliers and robust statistics

Real data are often affected by outliers. These are measurements that do not follow an assumed noise model. Consider taking an image of an empty region of the night sky with a digital camera. Our simple model might be that the sky is uniform across the image, with small variations about a constant value due to photon statistics (Poisson noise) and various sources of detector and electronics noise (Gaussian). We could bundle all of these into a Gaussian noise model, perhaps justified by the central limit theorem on the grounds that

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

combining lots of independent random effects will be approximately Gaussian. One could calculate a mean and standard deviation, then plot a Gaussian over a suitable histogram of the pixel measurements. The agreement may be quite good overall, but chances are that there are some pixels in this image that have very deviant values: extremely high, zero, or even negative. These *outliers* can come about for a large number of reasons, and depend on the physical set up and the data processing that has been done. In this example, origins of outliers could be: stars in the field; scattered light in the camera; cosmic rays hitting the detector; broken pixels; processing errors. The values involved could be so large that they would distort the calculation of the mean, standard deviation, or other statistics. The reason that they are outliers – and so need special treatment – is that they do not originate from the specified noise model.

Broadly speaking there are two things we can do with outliers: remove them or model them. Physical knowledge can be used to remove some outliers, for example if we know that flux values cannot be negative. But one must be careful about this, as some processing of the data may mean that such values are in fact acceptable. For example, if a so-called dark image has been subtracted from the sky image (done because the detector gives a signal even when not exposed to light), then because the dark image itself will contain noise we will now have negative values in the dark-subtracted sky image.

The effect of outliers can can be ameliorated by adopting a more robust statistic. For example, rather than using the mean, which weights all of the data equally, the median can be used to estimate the mean for distributions where they are theoretically equal. One could consider using a weighted mean (section 2.8.7), but this requires reliable uncertainty estimates, whereas the very nature of outliers normally means that these are absent. Instead of using the standard deviation, which is a squared quantity and thus strongly dependent on deviations from the mean, one could calculate the mean absolute deviation, the full-width at half-maximum, or the interquartile range, and then scale this to what the standard deviation would be for a theoretical distribution. The literature is full of suggestions of such robust estimators.

Another approach is to adopt an iterative approach to clip outliers. For example, to estimate the mean and standard deviation of a set of data that we believe is Gaussian but has outliers we can

(1) calculate the median and scaled interquartile range as initial estimates of the mean $\overline{x}$ and standard deviation $\hat{\sigma}$, respectively;
(2) using all of the data, remove ("clip") those values that lie more than $n\hat{\sigma}$ from $\overline{x}$, where n depends on how much of the data you expect to be outliers, but is typically 3–5;
(3) calculate the mean $\overline{x}$ and standard deviation $\hat{\sigma}$ from the remaining data;
(4) iterate steps 2 and 3 until convergence is reached, i.e. no more points are clipped.

We use the median instead of the mean in the first iteration because this provides a more robust estimation when the distribution is skewed by the outliers. It corresponds to an extreme clipping, as the median clips all points except the central one.

There are many ad hoc approaches to robust statistics and dealing with outliers, many of which need to be tailored to specific applications and which take into account what we know about potential outliers. In section 9.3 we will see how we can use a mixture model

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

to simultaneously identify and model outliers when fitting a model to data without having to clip them.

2.7 Errors, accuracy, and precision

Even within statistics the term *error* does not have a rigorous definition, as its broad usage in different books and publications makes clear.

In some contexts "error" might mean the difference between the measured value (or the inferred value) and the true value. This is a useful concept, although as the true value is usually unknown (unless we are doing simulations) this error cannot usually be computed.

A closely related term is *residual*, which usually refers to the difference between a measured value and a model prediction of this value (neither of which is usually equal to the true value). However, the word "error" is often use synonymously with "residual".

The term "error" is often used in a more colloquial sense to mean the uncertainty in a measurement; it is frequently used to indicate the standard deviation of the measurement model, as discussed in section 2.2. One often distinguishes between random errors and systematic errors.

Random errors refer to differences between measurements or potential measurements of the same thing. Such variations arise from the fact that the measurement process does not always deliver exactly the same value (the probability distribution describing the measurements has a non-zero width), for reasons discussed in section 2.2. We normally think of random errors as having zero mean, in which case repeated, independent measurements often allow us to reduce the uncertainty in the thing we are measuring in accordance with the central limit theorem (section 2.3).

Systematic errors, in contrast, mean that measurements are consistently offset (biased) with respect to the true value. The expectation value of the difference between the measured and true value is then not zero: no matter how many measurements we average we will not beat down systematic errors. If a weighing scale reads 5 mg when there is nothing on it, then all measurements will be systematically wrong by 5 mg. We can sometimes detect and remove systematic errors in the experimental set up by measuring objects for which we know the true values and calibrating accordingly (here by adjusting the zero point of the weighing scales).

If x' is the true, unknown value for something, then an example of a measurement model which includes both random and systematic errors is $\hat{x} = x' + \mathcal{N}(\Delta x, \sigma_x)$, where $\hat{x}$ indicates our measurement, or an estimate obtained through some analysis. We can think of σ_x as the standard deviation of the random errors and Δx as the systematic error. If we know this measurement model we can estimate σ_x from a set of measurements of $\hat{x}$ (e.g. by computing their sample standard deviation). But we cannot infer Δx, because it is degenerate with the unknown value x'. Of course, if we were told Δx then we could remove it by subtracting it from all the measurements. But we could never remove the actual random error from each measurement, because it is stochastic and remains unknown for each measurement.

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

Many scientists make the following distinction: a *precise* measurement is one which has a small random error and so has a small variance. An *accurate* measurement is one which has a small systematic error and so a small bias.

2.8 Propagation and combination of uncertainties

We are often given, or have determined, the uncertainty in one variable, but would like to know what the corresponding uncertainty is in a function of that variable. Here we look at a few simple yet common situations.

2.8.1 Linear function of one variable

Suppose we have a variable x with variance $\mathrm{Var}(x)$, or equivalently standard deviation $\sigma_x = \sqrt{\mathrm{Var}(x)}$. We take a linear function of x

$$f = a + bx \tag{2.14}$$

where a and b are fixed constants. What is the standard deviation of f?

We can derive this exactly from the definition of variance and the properties of the expectation operator (section 1.3)

$$\begin{aligned} \mathrm{Var}(f) &= E[(f - E[f])^2] \\ &= E[(a + bx - E[a + bx])^2] \\ &= E[(bx - bE[x])^2] \\ &= b^2 E[(x - E[x])^2] \\ &= b^2 \mathrm{Var}(x). \end{aligned} \tag{2.15}$$

Taking the square root to get an equation for the standard deviations gives

$$\sigma_f = |b|\sigma_x. \tag{2.16}$$

This is logical: from a dimensional point of view there must be a linear scale factor b in the conversion. a shouldn't play any role, as it is just a constant offset on all data points. So apart from a potential numerical factor, the above solution should have been obvious.

2.8.2 Arbitrary function of one variable

We can apply the same approach to propagate the uncertainties to a more general, nonlinear function $f(x)$. We do this by making a Taylor expansion of the function about its mean

$$f(x) \simeq f(x_0) + (x - x_0)\left.\frac{df}{dx}\right|_{x=x_0}. \tag{2.17}$$

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

As $E[f(x)] = f(x_0)$ and with $\delta x = x - x_0$, we have[6]

$$\begin{aligned} \mathrm{Var}(f) &= E[(f - E[f])^2] \\ &\simeq E\left[\left(\delta x \frac{df}{dx}\right)^2\right] = \left(\frac{df}{dx}\right)^2 \mathrm{Var}(x) \quad \text{so} \\ \sigma_f &\simeq \left|\frac{df}{dx}\right| \sigma_x. \end{aligned} \tag{2.19}$$

This approximation only holds for small uncertainties, such that the derivative does not change much with x. If we had a larger uncertainty, $x - x_0$ would be large and this first-order Taylor expansion would break down (we would need higher order terms). But we've not made any assumption about a Gaussian distribution of the uncertainties.

Example: angular errors

If the (small) uncertainty in an angle α is δ_α radians, what is the corresponding uncertainty in $\sin\alpha$?

From equation 2.19 it is $\delta_\alpha |\cos\alpha|$. Note that this is maximum when the angle is 0 or π radians, but zero when the angle is $\pi/2$. This is because when $\alpha = \pi/2$ the sine function has its maximum and is flat, so small changes in the angle have negligible effect on the sine.

2.8.3 Linear function of two variables

Now consider

$$f = a + bx + cy \tag{2.20}$$

for a, b and c fixed constants and x and y random variables. Proceeding as before

$$\begin{aligned} \mathrm{Var}(f) &= E[(f - E[f])^2] \\ &= E[(a + bx + cy - E[a + bx + cy])^2] \\ &= E[(bx + cy - bE[x] - cE[y])^2] \\ &= b^2(E[x^2] - E[x]^2) + c^2(E[y^2] - E[y]^2) + 2bc\,(E[xy] - E[x]E[y]) \\ &= b^2\mathrm{Var}(x) + c^2\mathrm{Var}(y) + 2bc\,\mathrm{Cov}(x, y) \end{aligned} \tag{2.21}$$

in which some tedious algebra has been omitted between lines 3 and 4. If x and y are uncorrelated then this reduces to

$$\sigma_f^2 = b^2\sigma_x^2 + c^2\sigma_y^2. \tag{2.22}$$

[6] Alternatively write

$$f(x) \simeq f(x_0) - x_0 \frac{df}{dx}\Big|_{x=x_0} + x \frac{df}{dx}\Big|_{x=x_0} \tag{2.18}$$

which is just $f \simeq a + bx$ where the first two terms are a (independent of x) and $b = df/dx$. So from equation 2.15 we have $\sigma_f \simeq |b|\sigma_x$

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

If $b = c = 1$, i.e. if f is an unweighted linear sum of x and y, then

$$\sigma_f^2 \;=\; \sigma_x^2 + \sigma_y^2. \tag{2.23}$$

This is often summarized by saying "errors sum in quadrature".

2.8.4 Arbitrary function of two variables

Expanding a general two-parameter function $f(x, y)$ as a Taylor series up to first order gives

$$f(x,y) \;\simeq\; f(x_0,y_0) \;+\; (x - x_0)\,\frac{\partial f}{\partial x} \;+\; (y - y_0)\,\frac{\partial f}{\partial y} \tag{2.24}$$

where the partial derivatives are computed at (x_0, y_0). This has the form of equation 2.20 (b and c are the derivatives). Equation 2.21 therefore tells us that

$$\mathrm{Var}(f) \;\simeq\; \left(\frac{\partial f}{\partial x}\right)^2 \mathrm{Var}(x) \;+\; \left(\frac{\partial f}{\partial y}\right)^2 \mathrm{Var}(y) \;+\; 2\frac{\partial f}{\partial x}\frac{\partial f}{\partial y}\,\mathrm{Cov}(x,y). \tag{2.25}$$

2.8.5 Arbitrary function of many variables

We can generalize the result in the previous section to a function of J variables $f(x_1, \ldots, x_j, \ldots, x_J)$ using the J-dimensional first-order Taylor expansion, in which case it's best to write down the result using matrix algebra in terms of the covariance matrix (see section 1.6.1). We get

$$\mathrm{Var}(f) \;\simeq\; (\nabla f)^{\mathsf{T}}\, C_x\, \nabla f \tag{2.26}$$

where C_x is the $J \times J$ covariance matrix of the data with terms $\sigma_{x_j}^2$ on the leading diagonal, and ∇f is a J-dimensional (column) vector of the gradients $\partial f / \partial x_j$. $\mathrm{Var}(f)$ is the resulting variance in f (a scalar). If the variables are uncorrelated this reduces to

$$\sigma_f^2 \;\simeq\; \sum_j \left(\frac{\partial f}{\partial x_j}\right)^2 \sigma_{x_j}^2. \tag{2.27}$$

In the case of having K functions $f_1, \ldots, f_k, \ldots, f_K$ of the J variables, then in addition to a variance in each f_k there will also, in general, be covariances between them. We can write these as the $K \times K$ covariance matrix C_f (the leading diagonal of which is the variances). Applying again the first-order Taylor expansion we get

$$C_f \;\simeq\; A\, C_x\, A^{\mathsf{T}} \tag{2.28}$$

where A is the $K \times J$ matrix of partial first derivatives – the *Jacobian matrix* – in which the element in the kth row and jth column is $\partial f_k / \partial x_j$. This approximation of the covariance becomes exact if the functions are linear in the $\{x_j\}$.

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

2.8.6 Fractional errors

If $f = xy$ and x and y are independent, then from equation 2.25 we have

$$\mathrm{Var}(f) = y^2\mathrm{Var}(x) + x^2\mathrm{Var}(y) \tag{2.29}$$

which is exact because the second and higher derivatives are zero. Dividing by f^2 gives

$$\left(\frac{\sigma_f}{f}\right)^2 = \left(\frac{\sigma_x}{x}\right)^2 + \left(\frac{\sigma_y}{y}\right)^2 \tag{2.30}$$

which we can remember by saying "fractional errors add in quadrature". You can check that we get the same result for the ratio of two variables x/y (although then the result is just approximate, because $\partial^2 f/\partial y^2 \neq 0$).

2.8.7 Weighted mean

If we have positive weights $\{w_i\}$ for independent measurements $\{x_i\}$ then their weighted mean is

$$\hat{\mu}_w = \frac{1}{\sum w_i}\sum_i w_i x_i. \tag{2.31}$$

We can use the result in section 2.8.5 to show that the variance in $\hat{\mu}_w$ is

$$\mathrm{Var}(\hat{\mu}_w) = \frac{1}{(\sum_i w_i)^2}\sum_i w_i^2\sigma_i^2 \tag{2.32}$$

where $\{\sigma_i^2\}$ are the variances of the individual measurements. The square root of $\mathrm{Var}(\hat{\mu}_w)$ is the standard deviation (or standard error) in the weighted mean.[7]

Often one takes N measurements of some quantity x, each of which has its own uncertainty estimate. How should we combine these into an estimate of x which takes the uncertainties into account, and what is the resulting uncertainty? Let's suppose we know the standard deviations $\{\sigma_i\}$ of the measurements. We can weight the measurements according to the inverse variance, $w_i = 1/\sigma_i^2$. From the above we then have that

$$\hat{\mu}_w = \frac{1}{\sum_i 1/\sigma_i^2}\sum_i \frac{x_i}{\sigma_i^2} \tag{2.33}$$

and the variance in this weighted mean is

$$\mathrm{Var}(\hat{\mu}_w) = \frac{1}{\sum_i 1/\sigma_i^2}. \tag{2.34}$$

As we will see later (section 4.4.1), this weighting is "best" in terms of being the maximum likelihood solution for Gaussian data. Be aware that the variances $\{\sigma_i^2\}$ are the true variances, not those estimated from the data. Variances estimated from the data themselves

[7] This is a biased estimator of the standard deviation, because we have estimated $\hat{\mu}_w$ from the data. One can derive an unbiased estimator, but in the interests of simplicity we will just live with this bias in this section. It will be of little practical significance once N is larger than about 10 or so. Let's not forget that we are anyway dealing with *estimates* from noisy data.

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003

have variances (uncertainties), and if we used these we would derive a different expression for the mean (and its variance). This is the case no matter what distribution the data are drawn from. Note the particular case of the Poisson distribution, for which the true variance equals the true mean. Having calculated an estimate of that mean, it is tempting to use this as the weight in the above.

It is also important to realise that $\mathrm{Var}(\hat{\mu}_w)$ is the variance of (i.e. a measure of our uncertainty in) the weighted mean. It is *not* the variance of the weighted sample. If we want a measure for the "spread" in the weighted sample (whatever that means exactly), we could generalize the definition of variance to include a weight, and quote

$$\frac{1}{\sum_i w_i} \sum_i w_i (x_i - \hat{\mu}_w)^2. \tag{2.35}$$

With $w_i = 1/\sigma_i^2$, this takes into account both the spread in the measurements and their individual uncertainties.

Note that $\mathrm{Var}(\hat{\mu}_w)$ is independent of the values $\{x_i\}$: moving the data points doesn't change how precisely we can determine the weighted mean from them. This is consistent with what we get when we don't use weights. Equation 2.34 then gives the variance in the mean as σ^2/N, where σ^2 is the assumed common variance of the data points. This is just what we got from the central limit theorem (section 2.3).

Downloaded from https://www.cambridge.org/core. Universitaet Heidelberg, on 06 Aug 2018 at 07:59:04, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/9781108123891.003