Exploring Data

Data come in all types and amounts, from a handful of hard-earned analytical quantities obtained after days of meticulous work in a laboratory to terabytes of remotely sensed data simply gushing in from satellites and remotely operated vehicles. Having a common language to describe data and to make initial explorations of trends are thus desirable.

1.1 Classification of Data

All observational sciences require data that may be analyzed and explored, which give rise to new ideas for how the natural world works. Such ideas may be developed into simple hypotheses that can ultimately be tested against new data and either be rejected or live to fight another day. New data crush hypotheses every day, hence we have to be careful with and respectful of our data to a much greater extent than our hypotheses and models. We start our exploration of data by considering the various ways we can categorize data, discussing a few basic data properties, and introducing typical steps taken in exploratory data analysis.

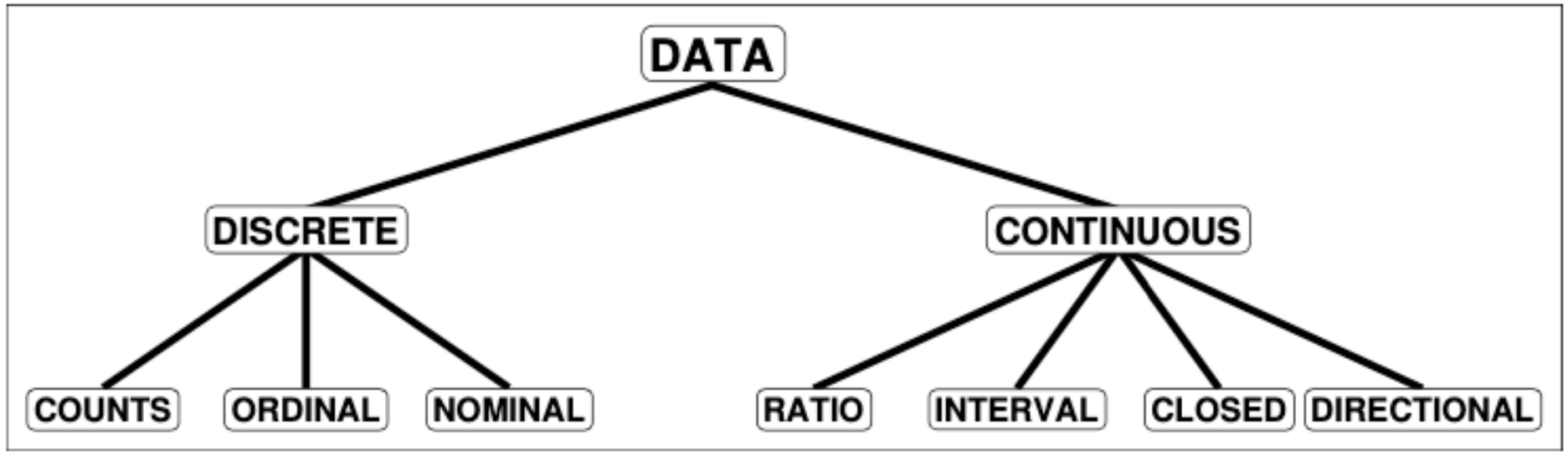

Figure 1.1: Classification of data types. All data we end up analyzing in a computer program are necessarily digitized and hence discrete, but they may represent a *phenomenon* that produced *continuous* output.

Exploring Data

1.1.1 Data types

Data represent measurements of either discrete or continuous quantities, often called *variables*. *Discrete* variables are those having discontinuous or individually distinct possible values (Figure 1.1). Examples of such data include:

- *Counts*: The flipping of a coin or rolling of dice, or the enumeration of individual items or groups of items. Such data can always be manipulated numerically.
- *Ordinal* data: These data can be ranked, but the intervals between consecutive items are not constant. For instance, consider Moh's hardness scale for minerals (Table 1.1): While *topaz* (hardness 8) is harder than *fluorite* (hardness 4), the values do not imply a doubling in hardness.
- *Nominal* data: These data cannot even be ranked. Examples of nominal data include categorizations or classifications, e.g., color of items (red *vs* blue marbles), lithology of rocks (sandstone, limestone, granite, etc.), and similar categorical data.

We will see in the following chapters that discrete data will often require specialized handling and testing.

Continuous variables are those that have an uninterrupted range of possible values (i.e., with no breaks). Consequently, they have an infinite number of possible values over a given range.

Our instruments, however, necessarily have finite precision and thus yield a finite number of recorded values. Examples of continuous data include:

- Seafloor depths, wingspans of birds, weights of specimens, and thickness of sedimentary layers
- Fault strikes, directions of wind and ocean currents.
- The Earth's geopotential fields, temperature anomalies, and dimensions of objects.
- Percentages of components (such data, which are closed and forced to a constant sum, sometimes require special care and attention).

Hardness	Mineral	Chemical Formula
1	Talc	$Mg_3Si_4O_{10}(OH)_2$
2	Gypsum	$CaSO_4 \cdot 2H_2O$
3	Calcite	$CaCO_3$
4	Fluorite	CaF_2
5	Apatite	$Ca_5(PO_4)_3(OH^-,Cl^-,F^-)$
6	Feldspar	$KAlSi_3O_8$
7	Quartz	SiO_2
8	Topaz	$Al_2SiO_4(OH^-,F^-)_2$
9	Corundum	Al_2O_3
10	Diamond	C

Table 1.1: Traditional Moh's hardness scale.

Exploring Data

In addition, much data of interest to scientists and engineers vary as a function of *time* and/or *space* (the independent variables). Since time and space vary continuously themselves, our discrete or continuous variables will most often vary continuously as a function of one or more of these independent variables. Such data represent continuous *time series* data. They may also be referred to as *signals*, *traces*, *records*, and other names. We may further subdivide continuous data into sub-categories:

- Ratio scale are data that have a fixed zero point (e.g., weights, temperatures in degrees Kelvin).
- Interval scale are data that have an arbitrary zero point (e.g., temperatures in degrees Celsius or Fahrenheit).
- Closed data are forced to attain a constant sum (e.g., percentages).
- Directional data have components (e.g., vectors or orientations in two or higher dimensions).

Data can also be classified according to how they are recorded for use: Analog signals are those signals which have been recorded continuously (even though one might argue that this is impossible due to instrument limitations). Discrete data are those which have been recorded at discrete intervals of the independent variable. In either case, data must be discretized before they can be analyzed by a computer. Consequently, all data which are represented in computers are necessarily discrete.

1.1.2 Data limits

For a variety of reasons, such as lack of time or funds, our data tend to be limited in one or more ways. In particular, limits typically will apply to three important aspects of any data set:

Domain : No phenomena can be observed over all time or over all space, hence data have a limited *domain*. The domain may be one-dimensional for scalar quantities and *N*-dimensional for data in *N* dimensions (e.g., a spatial vector data set such as the Earth's magnetic field is three-dimensional).

Range : It is equally true that no measuring technique can record (or transmit) values that are arbitrarily large or small. The lower limit on very small quantities is often set by the noise level of the measuring instrument (because matter is quantized, all instruments will have internal noise). The *Dynamic Range* (*DR*) is the range over which the data can be measured (or exists). This range is usually given on a logarithmic scale measured in *decibels* (dB), i.e.,

Exploring Data

$$DR = 10\log_{10}\left(\frac{\text{maximum power}}{\text{minimum power}}\right)\ \text{dB}. \tag{1.1}$$

One can see that every time *DR* increases by 10, the ratio of the maximum to minimum values has increased by an order of magnitude. Strictly speaking, (1.1) is to be applied to data represented as a *power* measurement (a squared quantity, such as variance or square of the signal amplitude). If the data instead represent *amplitudes*, then the formula should be

$$DR = 20\log_{10}\left(\frac{\text{maximum amplitude}}{\text{minimum amplitude}}\right)\ \text{dB}. \tag{1.2}$$

For example, if the ratio between highest and lowest voltage (or current) is 10, then $DR = 20\log_{10}(10) = 20$ dB. If these same data were represented in watts (power), and since power is proportional to the voltage squared, the ratio would be 100, and $DR = 10\log_{10}(100) = 20$ dB. Thus, regardless of the manner in which we express our data we get the same result (provided we are careful). In most cases though (except for electrical data) the first formula given is the one to be used (so the data extrema should be expressed as powers). Few instruments have a dynamic range greater than 100 dB. In any case, because of the limited range and domain of data, any data set, say $f(t)$, can be enclosed as

$$t_0 < t < t_1 \text{ and } |f(t)| < M, \tag{1.3}$$

for some constant *M*. Such functions are always integrable and manageable and can be subjected to further analysis without any special treatment.

Frequency : Finally, most measuring devices cannot respond instantly to sudden change. The resulting data are thus said to be *band limited*. This means the data will not contain frequency information higher (or lower) than the signal representing the fastest (or slowest) response of the recording device.

Exploring Data

1.1.3 Noise

In almost all cases, real data contain information other than what is strictly desired ("desired" is a key word here since we all know the saying that "one scientist's signal is another scientist's noise"). We say such data consist of *samples* of *random variables*. This statement does not mean that the data are totally random, but instead imply that the value of any future observation can only be predicted in a *probabilistic* sense — it cannot be exactly predicted as is the case for a *deterministic* variable, which is completely predictable by a known law. In other words, because of inherent variability in natural systems and the imprecision of experiments or measuring devices, if we were to give an instrument the same input at two different times we would likely get two different outputs due to the difference in *noise* at the two different times. In analyzing data, one must never overlook or ignore the role of noise, as understanding the noise level provides the key to how subtle signals we can reliably resolve in our data. One of the main goals in data analysis is to detect the signal in the presence of noise or to reduce the degree of noise contamination. We therefore try to enhance the *signal-to-noise ratio* (S/N), defined (in decibels) as

$$\mathrm{S/N} = 10 \log_{10} \left(\frac{\text{power of signal}}{\text{power of noise}} \right) \text{ dB}. \tag{1.4}$$

Minimizing the influence of noise during data acquisition and stacking co-registered data are some of the approaches used to enhance the S/N ratio.

Exploring Data

1.1.4 Accuracy versus precision

An *accurate* measurement is one that is very close to the true value of the phenomenon we are observing. A *precise* measurement is one that has very little scatter: Repeat measurements will give more or less the same values (Fig- ure 1.2). If the measured data have high precision but poor accuracy, one may suspect that a systematic *bias* has been introduced, e.g., we are using an instrument whose zero position has not been calibrated properly. If we do not know the expected value of a phenomenon but are trying to determine just that, it is obviously better to have accurate observations with poor precision than very precise, but inaccurate values, since the former will give a correct, but imprecise estimate while the latter will give a wrong, but very precise result! Fortunately, we will often have a good idea of what a result should be and can use that prior knowledge to detect any bias in the measurements.

Figure 1.2: Precision is a measure of repeatability, while accuracy refers to how close the average observed value is to the “true” value.

1.1.5 Randomness

Phenomena, and hence the data that represent them, may range from those that are completely determined by a natural law (*deterministic* data) to those that seem to have no inherent structure or patterns (*chaotic* or *random*). Most phenomena, and hence the data we collect, lie in-between these two extremes. Here, there is a more-or-less clear pattern or structure to the data, but each individual measurement also exhibits a random component that makes it impossible to predict the outcome of a future observation with 100% certainty. Such data are called *probabilistic* or *stochastic*. We will often fit simple, deterministic models to such data and hope that the residuals reflect insignificant, uncorrelated noise. We will also want to know if that hope is warranted using specific statistical tests.

Exploring Data

1.1.6 Analysis

Analysis means to separate something into its principal components in order to identify, interpret and/or study the underlying structure. In order to do this properly we should have some idea of what the components are likely to be. Therefore, we should have some concept of a model of the data in mind (whether this is a conceptual, physical, intuitive, or some other type of model is not important). We essentially need guidelines to aid our analysis. For example, it is not a good idea to take a data set and simply compute its Fourier series because you happen to know something about Fourier analysis. One needs to have an idea as to what to look for in the data. Often, this knowledge will grow with a set of well planned ongoing analyses, whose techniques and uses are the essence of this book.
The following steps are parts of most data analysis schemes:

1. *Collect* or obtain the data.
2. Perform *exploratory data analysis.*
3. *Reduce* the data to a few quantities (*statistics*) that *summarize* their bulk properties.
4. Compare data to various *hypotheses* using statistical *tests*.

We will briefly discuss step (1) while reviewing error analysis, which is the study and evaluation of uncertainty in measurements and how these propagate into our final statistical estimates. The main point of step (2) is to familiarize ourselves with the data set and its major structure. This acquaintance is almost always best done by graphing the data. Only an inexperienced analyst will use a sophisticated “black-box” technique to compile statistics from data and accept the validity of these statistics without actually looking at the data. Step (3) will usually include a model (simple or complicated) where the purpose is to extract a few representative parameters out of possibly millions of data points. These statistics can then be used in various tests (4) to help us decide which hypothesis the data favor, or rather *not* favor. That is the curse of statistics: You can never prove anything, just disprove! By disproving all possible hypotheses other than your pet theory, other scientists will eventually either grudgingly accept your views or they die of old age and then your theory will be accepted! Hence, persistence and longevity are important characteristics of a successful researcher. Joking aside, it is important to listen to your data as it is disturbingly easy to be convinced that your pet model or theory is right, data be damned. This phenomenon is called *conviction bias*.

Exploring Data

1.2 Exploratory Data Analysis

As mentioned, the main objective of exploratory data analysis (EDA) is to familiarize yourself with your observations and determine their main structure. Since simply staring at a table or computer printout of numbers will eventually lead to premature blindness or debilitating insanity, there are several standard techniques that we will classify under the broad EDA heading:

1. Scatter plots — show it all.
2. Schematic plots — simplify the sample.
3. Histograms — explore the distribution.
4. “Smoothing” of data — reduce the noise.
5. Residual plots — determine the trends.

We will briefly discuss each of these five categories of exploratory techniques

1.2.1 Scatter plots

If practical, consider plotting every individual data value on a single graph. Such “scatter” plots show graphically the correlation between points, the orientation of the data, bad outliers, and the spread of clusters (e.g., Figure 1.3). We will later (in Chapter 3) provide a more rigorous definition for what correlation is; at this stage it is just a visual appearance of a trend. Scatter plots are particularly useful in two dimensions, but even three-dimensional data are fairly easy to visualize. For higher dimensions we may choose to view *projections* of the data onto lower-dimensional spaces and thus examine only 2–3 components at the time.

Figure 1.3: Scatter plots showing all individual data points — the “raw” data — are invaluable in identifying outlying data points and other potential problems.

Exploring Data

1.2.2 Schematic plots

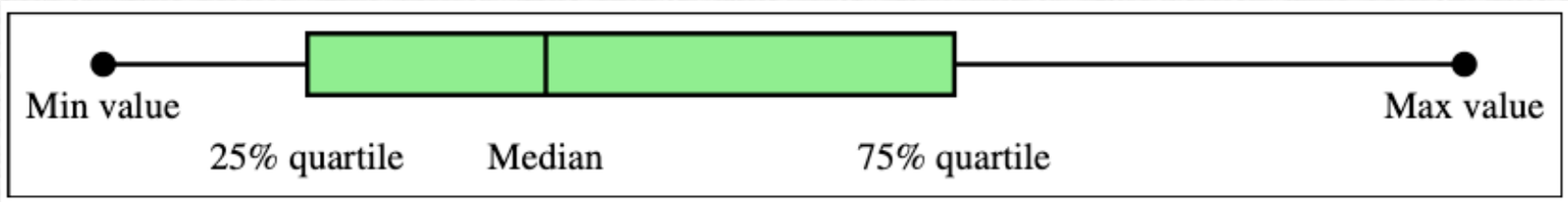

Figure 1.4: An example of a "box-and-whisker" diagram. The five quartiles give a visual representation of how one-dimensional data (or a single component of higher-dimensional data) are distributed.

The main objective here is to summarize a one-dimensional data distribution using a simple graph. One very common method is the *box-and-whisker* diagram, which graphically presents five informative measures of the sample. These five quantities are the *range* of the data (represented by the minimum and maximum values), the *median* (a line at the half-way point), and the *concentration* (represented by the 25% and 75% *quartiles*) of a data distribution. Schematically, these statistics can be conveniently illustrated as shown in Figure 1.4.

As an example of the successful use of box-and-whisker diagrams, we shall return to the winter of 1893–94 when Lord Rayleigh was investigating the density of nitrogen from various sources. Some of his previous experiments had indicated that there seemed to be a discrepancy between the densities of nitrogen produced by removing the oxygen from a sample of air and the nitrogen produced by decomposition of different chemical compounds. The 1893–94 results clearly established this difference and prompted further investigations into the composition of air, which eventually led him to the discovery of the inert gas *argon*. Part of his success in convincing his peers has been attributed to his use of box-and-whisker diagrams to emphasize the difference between the two data sets he was investigating. We will use Lord Rayleigh's data (reproduced in Table 1.2) to make a scatter plot and two schematic plots: The already mentioned box-and-whisker diagram and the *bar* graph.

Exploring Data

Date	Origin	Purifying Agent	Weight
29 Nov. 1893	NO	Hot iron	2.30143
5 Dec. 1893	”	”	2.29816
6 Dec. 1893	”	”	2.30182
8 Dec. 1893	”	”	2.29890
12 Dec. 1893	Air	”	2.31017
14 Dec. 1893	”	”	2.30986
19 Dec. 1893	”	”	2.31010
22 Dec. 1893	”	”	2.31001
26 Dec. 1893	N_2O	”	2.29889
28 Dec. 1893	”	”	2.29940
9 Jan. 1894	NH_4NO_2	”	2.29849
13 Jan. 1894	”	”	2.29889
27 Jan. 1894	Air	Ferrous hydrate	2.31024
30 Jan. 1894	”	”	2.31030
1 Feb. 1894	”	”	2.31028

Table 1.2: Lord Rayleigh’s density measurements of nitrogen (Lord Rayleigh, On an anomaly encountered in determinations of the density of nitrogen gas, *Proc. Roy. Soc. Lond., 55*, 340–344, 1894).

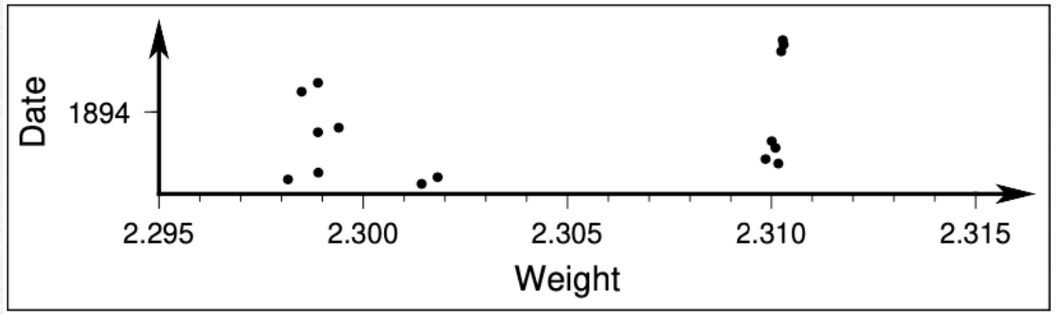

Figure 1.5: Scatter plot of Lord Rayleigh’s data on nitrogen. The plot reveals distinct groups.

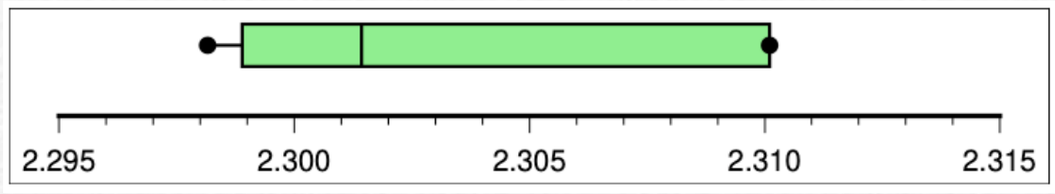

Figure 1.6: Schematic box-and-whisker plot of Lord Rayleigh’s nitrogen data. While representing the entire sample the graph obscures the pattern so clearly revealed by the scatter plot.

We will first look at all the data using a scatter plot. It may look something like Figure 1.5. It is immediately clear that we probably have two different data groupings here. Note that this is only apparent if you plot the “raw” data points. Plotting all the values as one data set using the box-and-whisker approach would result in a confusing graph (Figure 1.6), which tells us very little that is meaningful. Even the median, traditionally a stable indicator of “average” value, is questionable since it lies between the two data clusters. Clearly, it is important to find out if our data consist of a single population or if it contains a mix of two (or even more) data components.

Exploring Data

Fortunately, in this example we know how to separate the two data sets based on their origins. It appears that we are better off plotting the data sets separately instead of treating them as a single population. However, the choice of diagram is also important. Consider a simple bar graph (here indicating the average value) summarizing the data given in Table 1.2. It would simply look as shown in Figure 1.7. In this presentation, it is just barely visible that the weight of "nitrogen" extracted from the air is slightly heavier than nitrogen extracted from the chemical compounds. Given the way they are shown, the data present no clear indication that the two data sets are *significantly* different. Part of the problem here is the fact that we are drawing the bars from an origin at zero, whereas all the variation actually takes place in the 2.29–2.32 interval (again evident from the scatter plot in Figure 1.5). By expanding the scale and choosing a box-and-whisker plot we concentrate on the differences and produce an illustration as the one shown in Figure 1.8.

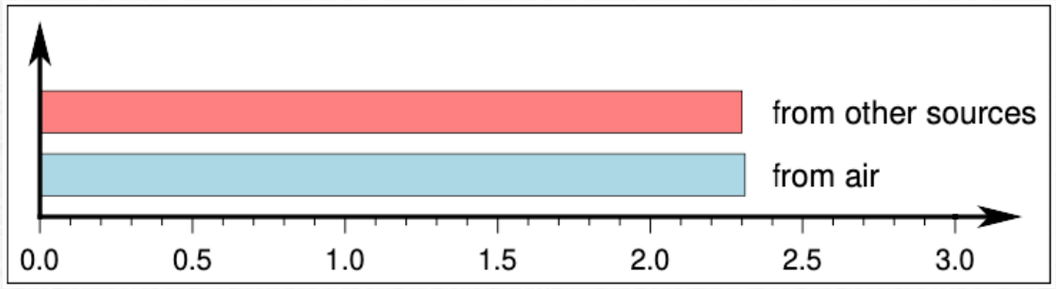

Figure 1.7: Bar graph of the average values from Lord Rayleigh's nitrogen data. Because the average values are very similar the two bars look very similar and do not tell us much.

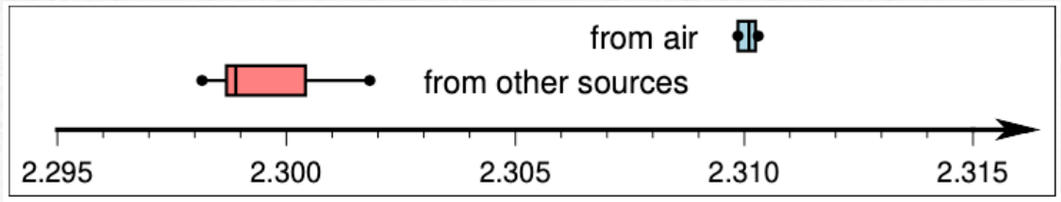

Figure 1.8: Separate box-and-whisker diagrams of nitrogen weight given in Table 1.2 based on origin. Their separation and extent clearly convey the finding of two separate sources.

It is obvious that the second box-and-whisker diagram allows a clearer interpretation than the bar graph. The diagram also benefits from the stretched scale, which highlights the different ranges of the data groupings. In Lord Rayleigh's situation the convincing diagram was accepted as strong evidence for the existence of a new element (later determined to be *argon*), and a Nobel prize in physics followed in 1904.

Exploring Data

1.2.3 Histograms

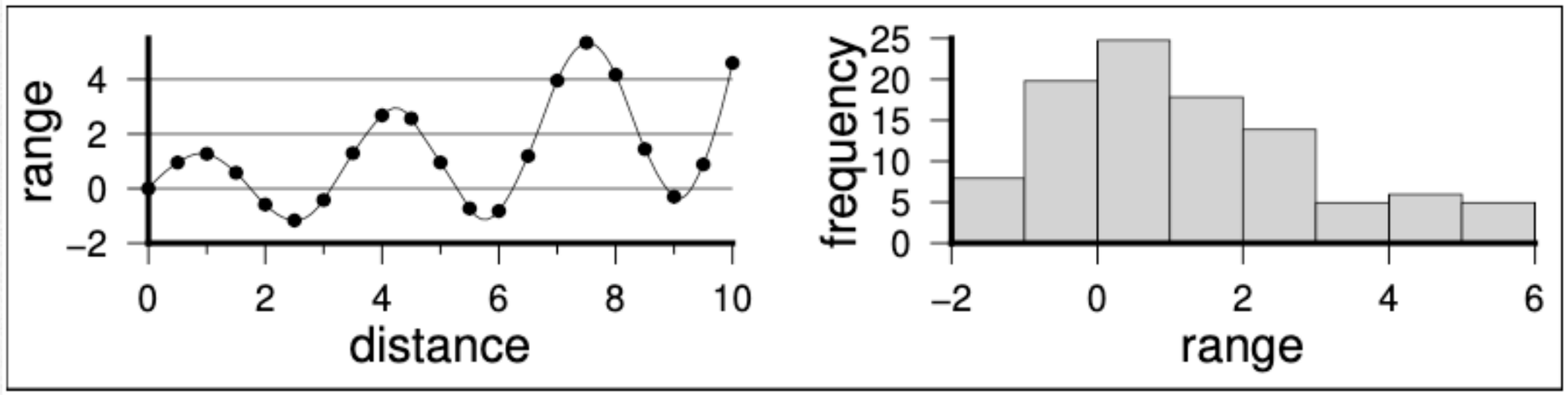

Figure 1.9: A data set, here a function of distance, can be converted to a histogram by counting the frequency of occurrence within each sub-range. The histogram only uses the *y*-values of the (*x*, *y*) points shown in the left diagram.

Histograms convey an accurate impression of the data *distribution* even if it is multimodal. One simply breaks the data range into equidistant sub-ranges and plots the frequency or occurrences for each range (e.g., Figure 1.9). Obviously, the width of the sub-range determines the level of detail you will see in the final histogram. Because of this, it is usually a good idea to plot the discrete values as individual points since the “binning” throws away some information about the distribution. If the amount of data is moderate, then one can plot the individual values inside the histogram bars. Furthermore, you should explore how the *shape* of the distribution changes as you vary the *bin width*. Clearly, as the histogram bin width approaches zero you will end up with one point (or none) per bin, while at the other extreme (a very wide bin width) you will simply have a single bin with all your data. Try to select a width that yields a representative distribution, but at the same time try to understand what is going on when your widths give different shapes.

Exploring Data

1.2.4 Smoothing

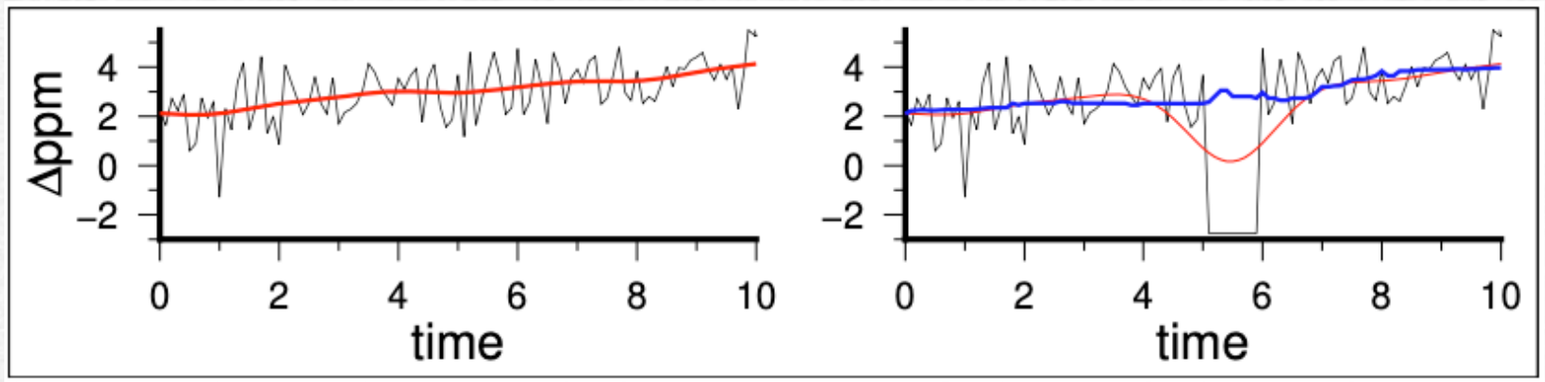

Figure 1.10: Smoothing of data is usually done by filtering. The left panel shows a noisy data set and a smooth Gaussian filtered curve (red), while the right panel shows the same data with a glitch between $x = 5$ and $x = 6$. For such data the Gaussian filter is unduly influenced by the outlying data whereas a median filter (blue) is much more tolerant.

The purpose of *smoothing* is to highlight the general trend of the data and suppress high-frequency oscillations. We will briefly look at two types of smoothing: The *median* filter and the *Hanning* filter. A detailed introduction to filtering will be presented in the second volume of this series. The median filter is typically a three-point filter and simply replaces a point with the median value of the point and its two immediate neighbors. The filter then shifts one step further to the right and the process repeats. This technique is very efficient at removing isolated spikes or *outliers* in the data since the bad points will be completely ignored as they will never occupy the median position, unless they appear in groups of two or more (in that case a wider median filter, say 5-point, would be required). In contrast, the Hanning filter is simply a moving average of three points where the center point is given twice the weight of the neighbor points, i.e.,

$$y'_i = \frac{y_{i-1} + 2y_i + y_{i+1}}{4} \qquad (1.5)$$

Note that while such a filter works well for data that have random high-frequency noise, it gives disastrous results for spiky data since the outliers are averaged into the filtered value and never simply ignored. For noisy data with occasional outliers one might consider running the data first through a median filter, followed by a treatment of the Hanning filter. In building automated data procedures one should always have the worst-case scenario in mind and seek to protect the analysis by preprocessing with a median filter. Figure 1.10 illustrates the use of simple smoothing with Gaussian filters, with or without protection from outliers by a median filter.

Exploring Data

1.2.5 Residual plots

We can always make the assumption that our data can be decomposed into two parts: A smooth trend plus noisier residuals. This separation forms the basis for *regional-residual* analysis. The simplest trend (or regional) is just a straight line. One can easily define such a line by picking two representative points (x_1,y_1) and (x_2,y_2) and then compute the trend as

$$T(x) = y_1 + \left(\frac{y_2 - y_1}{x_2 - x_1}\right)(x - x_1). \qquad (1.6)$$

We then can form residuals $r_i = y_i - T(x_i)$ (in Chapter 6 we will learn more rigorous ways to find linear trends in $x - y$ data). If a significant trend still exists, one can try several standard transformations to determine the nature of the "smooth" trend, such as the family of trends represented by

$$y^n, \ldots, y^1, y^{1/2}, \log(y), y^{-1/2}, y^{-1}, \ldots, y^{-n}. \qquad (1.7)$$

The residual plot procedure we will follow is:

1. Take log(y) of the data and plot the values.
2. If the data is convex then choose a transformation closer to y^{-n} .
3. If the line is concave then choose a transformation closer to y^{n}

While approximate, this method gives you a feel for how the data vary. Note that if your y-values contain or straddle zero then you cannot explore a logarithmic relationship for that axis.

Exploring Data

We will conclude this section with a quote from J. Tukey's book "Exploratory Data Analysis", which is worth contemplating:
"Many people would think that plotting y against x for simple data is something requiring little thought. If one wishes to learn but little, it is true that a very little thought is enough, but if we want to learn more, we must think more."

The moral of it all is: *Always* plot your data. *Always*! *Never* trust output from software performing statistical analyses without comparing the results to your data. Very often, such software implements statistical methods that are based on certain assumptions about the data distribution, which may or may not be appropriate in your case. Plotting your data may be the easiest thing to do (for dimensionality less than or equal to three) or it may be quite challenging. Exploring sub-spaces of just a few dimensions is a simple starting point, while more sophisticated methods will examine the dimensionality of the data to see if some dimensions simply provide redundant information; we will explore such methods in the second volume of this book series.

REVIEW OF ERROR ANALYSIS

"An error does not become truth by reason of multiplied propagation, nor does the truth become error because nobody will see it."
Mahatma Gandhi, India Independence Leader

Error analysis is the study and evaluation of uncertainty in measurements of continuous data (discrete data may have no errors). We know that no measurement, however carefully made, can be completely free of uncertainties. Since science depends on measurements, it is crucially important to be able to evaluate these uncertainties and to keep them to a minimum. Thus errors are not mistakes and you cannot avoid them by being very careful, but you should strive to make them as small as possible. Understanding uncertainties is critical for determining the significance of models that you may construct to explain your data. In this chapter we will review the basic rules that govern the reporting of measurements with uncertainties and how these uncertainties propagate when used to obtain derived quantities (such as sums and products) as well as how they affect the uncertainties in more complicated situations (such as being arguments to nonlinear functions).

2.1 Reporting Uncertainties

When reporting the value of a measurement, care should be taken to give the best possible estimate of the uncertainty or error in the measurement. Values read off scales or measured with mechanical instruments can usually be bracketed between lower and upper limits. E.g., we may know that a temperature measurement T is not less than 23° and not larger than 24°. Hence, we shall report the value of T as

$$T = 23.5 \pm 0.5^\circ, \tag{2.1}$$

or, in general, $x \pm \delta x$. For many types of measurements we can state with absolute certainty that x must be within these bounds. Unfortunately, very often we cannot make such a categorical statement. To do so, we would have to specify unreasonably large values for δx to be absolutely confident that the actual quantity lies within the stated interval. In most cases, we will lower our confidence to, say, 90% and use a smaller δx. We need more detailed knowledge of the statistical laws that govern the process of measurement for finding a suitable δx, so we will return to this issue later in this book.

REVIEW OF ERROR ANALYSIS

2.1.1 Significant figures

In general, the last significant figure of any stated answer should be of the same order of magnitude (i.e., same decimal position) as the uncertainty. For intermediate calculations you should use one extra decimal (if calculating by hand) or all available decimals (if using calculators or computers).

> *Example 2–1. If a measured distance d is 6051.78 m with an uncertainty of ±30 m, you should report the distance as d = 6050 ± 30 m. Similarly, if a time t is measured to be 3 seconds and the uncertainty is given as 1 part in 100, you should report the time as t = 3.00 ± 0.03 s.*

2.2 Fractional Uncertainty

While the statement $x = x_0 \pm \delta x$ indicates the *precision* of the measurement, it is clear that such a statement will have different meanings depending on the value of x_0 relative to δx. Clearly, with δx = 1 m, we imply a different precision for x_0 = 3 m than for x_0 = 1000 km. Thus, we should consider the *fractional uncertainty*, written as

$$\frac{\delta x}{|x_0|}. \tag{2.2}$$

Note that the fractional uncertainty is a *nondimensional* quantity.

> *Example 2–2. We wish to report the uncertainty in the length of a 3m crocodile, which (due to a shortage of available arms) we only were able to measure with a precision of 6 cm. Using fractional uncertainty, we report the uncertainty*
>
> $$\delta l = \pm 100 \frac{0.0\bar{6}}{3}\% = \pm 2\%$$

REVIEW OF ERROR ANALYSIS

2.3 Uncertainty in Derived Quantities

It is common to obtain measurements (and estimate their uncertainties) and then use these values in expressions that lead to derived quantities. In such situations we must be careful to *propagate* the uncertainties of the initial measurements so that the derived quantities can be reported with their combined uncertainties. The answers will differ depending on whether or not the individual measurements are *independent* or *dependent* on each other.

2.3.1 Uncertainty in sums and differences

Consider the two values $x \pm \delta x$ and $y \pm \delta y$. We would like to determine the correct expressions for the uncertainties in the four derived quantities

$$s = x + y, \tag{2.3}$$

$$d = x - y, \tag{2.4}$$

$$p = x \cdot y, \tag{2.5}$$

$$q = x/y. \tag{2.6}$$

For the sum, common sense would suggest that the maximum value of s must be $s = x + y + \delta x + \delta y$, while the minimum value is $s = x+y-\delta x-\delta y$. Thus

$$s \approx (x+y) \pm (\delta x + \delta y) = s_0 \pm \delta s \tag{2.7}$$

and similarly for differences we reason that

$$d \approx (x-y) \pm (\delta x + \delta y) = d_0 \pm \delta d. \tag{2.8}$$

Here, the subscript 0 indicates the nominal or “best” value of the results. We use the approximate sign ≈ since we anticipate that the stated uncertainties δs and δd probably overestimate the true uncertainties in the sum and difference, respectively. Note that the uncertainties in both the sum and the difference of x and y are identical.

REVIEW OF ERROR ANALYSIS

2.3.2 Uncertainty in products and quotients

For the product, we first use fractional uncertainties and rewrite *x* and *y* as

$$x = x_0\left(1 \pm \frac{\delta x}{|x_0|}\right),$$

$$y = y_0\left(1 \pm \frac{\delta y}{|y_0|}\right).$$

Then, the maximum value of *p* is

$$p_{high} = x_0\left(1 + \frac{\delta x}{|x_0|}\right) \cdot y_0\left(1 + \frac{\delta y}{|y_0|}\right), \tag{2.9}$$

which becomes

$$p_{high} = p_0\left(1 + \frac{\delta x}{|x_0|} + \frac{\delta y}{|y_0|} + \ldots\right), \tag{2.10}$$

where the higher order term proportional to $\delta x \cdot \delta y$ has been ignored. We note that the best value is

$$p_0 = x_0 \cdot y_0. \tag{2.11}$$

The minimum value is found by reversing the signs of δx and δy. Both expressions yield the uncertainty in *p* as

$$\frac{\delta p}{|p_0|} \approx \frac{\delta x}{|x_0|} + \frac{\delta y}{|y_0|}. \tag{2.12}$$

For quotients, the maximum value will be

$$q_{high} = \frac{x_0\left(1 + \frac{\delta x}{|x_0|}\right)}{y_0\left(1 - \frac{\delta y}{|y_0|}\right)} = q_0\frac{1+a}{1-b}. \tag{2.13}$$

REVIEW OF ERROR ANALYSIS

Using the binomial theorem, we expand $(1-b)^{-1}$ as $1+b+b^2+b^3...$, hence

$$\frac{1+a}{1-b} \approx (1+a)(1+b) = 1+a+b+ab \approx 1+a+b, \tag{2.14}$$

where we again ignore higher-order terms in b by assuming that $(\delta x/|x_0|) \ll 1$ and $(\delta y/|y_0|) \ll 1$. Similarly, for the minimum value, find

$$\frac{1-a}{1+b} \approx (1-a)(1-b) = 1-a-b+ab \approx 1-(a+b). \tag{2.15}$$

It therefore follows that

$$q = q_0\left[1 \pm \left(\frac{\delta x}{|x_0|} + \frac{\delta y}{|y_0|}\right)\right] = q_0\left[1 \pm \frac{\delta q}{q_0}\right] \tag{2.16}$$

and that the fractional uncertainty for both products and quotients are the same.

REVIEW OF ERROR ANALYSIS

2.3.3 Uncertainty for general expressions

The stated uncertainties are the maximum values possible, but we suspect these are likely to be exaggerated. Later, we will show that if we assume our errors to be *normally* distributed (i.e., we have "Gaussian" errors) and our measurements are *independent*, then a better estimate of the uncertainty in sums and differences is

$$\delta s = \delta d = \sqrt{(\delta x)^2 + (\delta y)^2}, \tag{2.17}$$

and for products and quotients it becomes

$$\frac{\delta p}{p_0} = \frac{\delta q}{q_0} = \sqrt{\left(\frac{\delta x}{x_0}\right)^2 + \left(\frac{\delta y}{y_0}\right)^2}. \tag{2.18}$$

Note that in the case $s = nx$, where n is a constant, we must use $\delta s = n\delta x$ since all the x are the same and obviously *not independent* of each other. Similarly, the product $p = x_n$ will have the fractional uncertainty

$$\frac{\delta p}{p_0} = n\frac{\delta x}{x_0}, \tag{2.19}$$

since the (repeated) measurements are *dependent*. In conclusion, if $s = x + y + nz - u - v - mw$, then

$$\delta s = \sqrt{(\delta x)^2 + (\delta y)^2 + (n\delta z)^2 + (\delta u)^2 + (\delta v)^2 + (m\delta w)^2}. \tag{2.20}$$

Even if our assumption of independent measurements is incorrect, δs cannot exceed the ordinary sum

$$\delta s = \delta x + \delta y + n\delta z + \delta u + \delta v + m\delta w. \tag{2.21}$$

REVIEW OF ERROR ANALYSIS

Similarly, if

$$q = \frac{x \cdot y \cdot z^n}{u \cdot v \cdot w^m}, \tag{2.22}$$

then we find the fractional uncertainty to be

$$\frac{\delta q}{q_0} = \sqrt{\left(\frac{\delta x}{x_0}\right)^2 + \left(\frac{\delta y}{y_0}\right)^2 + \left(n\frac{\delta z}{z_0}\right)^2 + \left(\frac{\delta u}{u_0}\right)^2 + \left(\frac{\delta v}{v_0}\right)^2 + \left(m\frac{\delta w}{w_0}\right)^2}. \tag{2.23}$$

While this is the likely error for independent data, we can confidently say that the fractional uncertainty will be less than the linear sum

$$\frac{\delta q}{q_0} = \frac{\delta x}{x_0} + \frac{\delta y}{y_0} + n\frac{\delta z}{z_0} + \frac{\delta u}{u_0} + \frac{\delta v}{v_0} + m\frac{\delta w}{w_0}. \tag{2.24}$$

REVIEW OF ERROR ANALYSIS

Example 2–3. As most students who have taken an introductory physics lab in mechanics will know, measuring the period T of a pendulum of length ℓ lets one estimate the acceleration of gravity as

$$g = \frac{4\pi^2 \ell}{T^2}. \tag{2.25}$$

We wish to determine the uncertainty in this estimate given measurements of ℓ and T and their uncertainties. Being independent measurements we use (2.23) and find

$$\frac{\delta g}{|g_0|} = \sqrt{\left(\frac{\delta \ell}{\ell}\right)^2 + \left(2\frac{\delta T}{T}\right)^2}, \tag{2.26}$$

since the constants 4 and π have no uncertainty. Given the measurements $\ell = 92.95 \pm 0.10$ cm and $T = 1.936 \pm 0.004$ s, we obtain

$$g_0 = \frac{4\pi^2 0.9295 \text{ m}}{1.936^2 \text{ s}^2} = 9.79035 \text{ m s}^{-2}. \tag{2.27}$$

We can now evaluate the fractional uncertainty as

$$\frac{\delta g}{|g_0|} = \sqrt{\left(\frac{0.1}{92.95}\right)^2 + \left(2\frac{0.004}{1.936}\right)^2} \approx 0.4\%. \tag{2.28}$$

The answer, therefore, is

$$g = 9.79 \pm 0.04 \text{ m s}^{-2}, \tag{2.29}$$

where we have only used two significant decimals.

REVIEW OF ERROR ANALYSIS

Another case study revisits the debate that raged in the 19th century regarding the age of the Earth. Observing the slow process of erosion, Charles Darwin had implied that perhaps the Earth might be as old as 300 million years. Lord Kelvin, the preeminent physicist of his times, strongly objected and set out to calculate the age using the conductive cooling of the Earth (this and many other fascinating stories from the development of the geological sciences are portrayed in the classic book, *Great Geological Controversies* by A. Hallam [Oxford University Press]).

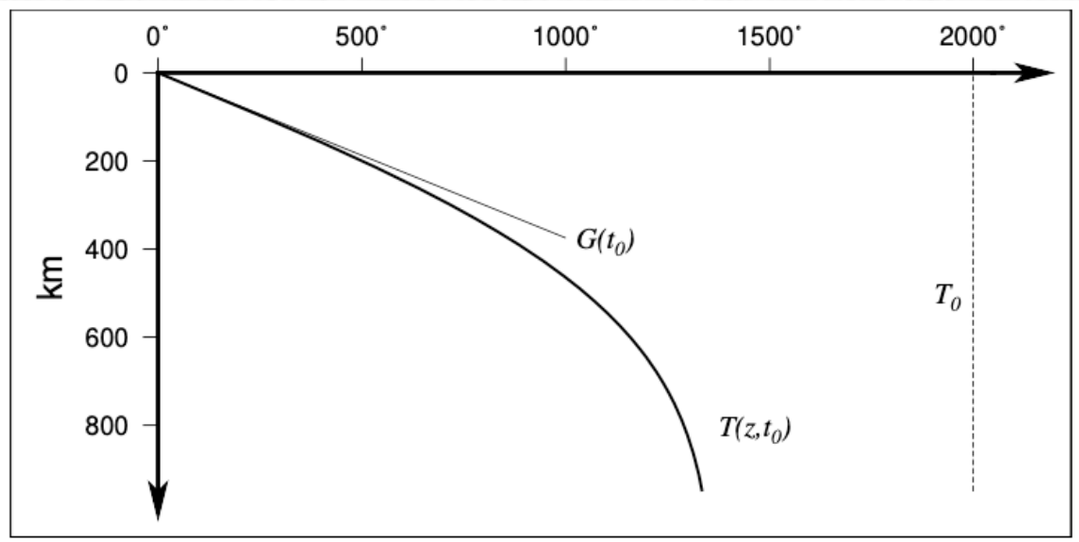

Figure 2.1: Lord Kelvin's model for the vertical temperature profile of the Earth (the *geotherm*) at a time t_0 since its initial formation at a constant temperature T_0 (dashed vertical geotherm). The tangent to the curve at the surface represents the vertical temperature gradient (G), which could be estimated from temperature measurements in mines.

Example 2–4. Lord Kelvin assumed the whole Earth was once at a uniform temperature T_0 and had since cooled at the surface to $\sim 0°C$ (as indicated in Figure 2.1). Then, the physics of heat conduction in solids dictates that

$$t_0 = \frac{T_0^2}{\pi \kappa G^2}, \tag{2.30}$$

with initial temperature $T_0 \approx 2000 \pm 200$ °K, thermal diffusivity $\kappa = 1 \pm 0.25$ mm^{2}s^{-1}, and observed near-surface temperature gradient $G = 25 \pm 5$ °K km^{-1}. Given the procedures established earlier, we first determine

$$\frac{\Delta t_0}{t_0} = \left[\left(2\frac{\Delta T_0}{T_0} \right)^2 + \left(\frac{\Delta \kappa}{\kappa} \right)^2 + \left(2\frac{\Delta G}{G} \right)^2 \right]^{\frac{1}{2}}. \tag{2.31}$$

REVIEW OF ERROR ANALYSIS

Inserting the estimated parameters, we find

$$\frac{\Delta t_0}{t_0} = \left[\left(2\frac{200}{2000} \right)^2 + \left(\frac{0.25}{1} \right)^2 + \left(2\frac{5}{25} \right)^2 \right]^{\frac{1}{2}} \approx 51\%. \tag{2.32}$$

We evaluate Kelvin's estimate of the age of the Earth to be

$$t_0 = \frac{(2000\ ^\circ\mathrm{K})^2}{\pi \cdot 10^{-6}\mathrm{m}^2\mathrm{s}^{-1}(25\ ^\circ\mathrm{K} \cdot 10^{-3}\mathrm{m}^{-1})^2} \approx 2 \cdot 10^{15}\ \mathrm{s}\ = 65\ \mathrm{Myr}. \tag{2.33}$$

Given the fractional uncertainty, we obtain $t_0 = 65 \pm 33$ Myr. As the debate raged on, positions hardened and Lord Kelvin continued to revise his estimates downwards, finally settling on 25 Myr. Modern science estimates that the Earth is closer to 4.6 *billion* years old. Where did Kelvin go wrong?

As a final case, let us imagine we are measuring the length of the coastline segment in Figure 2.2 using two different methods: (1) Set a compass to a fixed aperture $\Delta x = 1 \pm 0.025$ cm and march along the line counting the steps, and (2) use a digitizing tablet and sample the line approximately every $\Delta x = 1 \pm 0.1$ cm. Let us assume that it took $N = 50$ clicks or steps so the measured line length in both cases is 50 cm. What is the uncertainty in the length for the two methods? First, let us state that there will be an uncertainty for both methods that has to do with the undersampling of short-wavelength coastline "wiggles" (in the continuation of this volume we will learn about the *fractal* nature of coastlines and that perhaps our simple approach here is a bit naive).

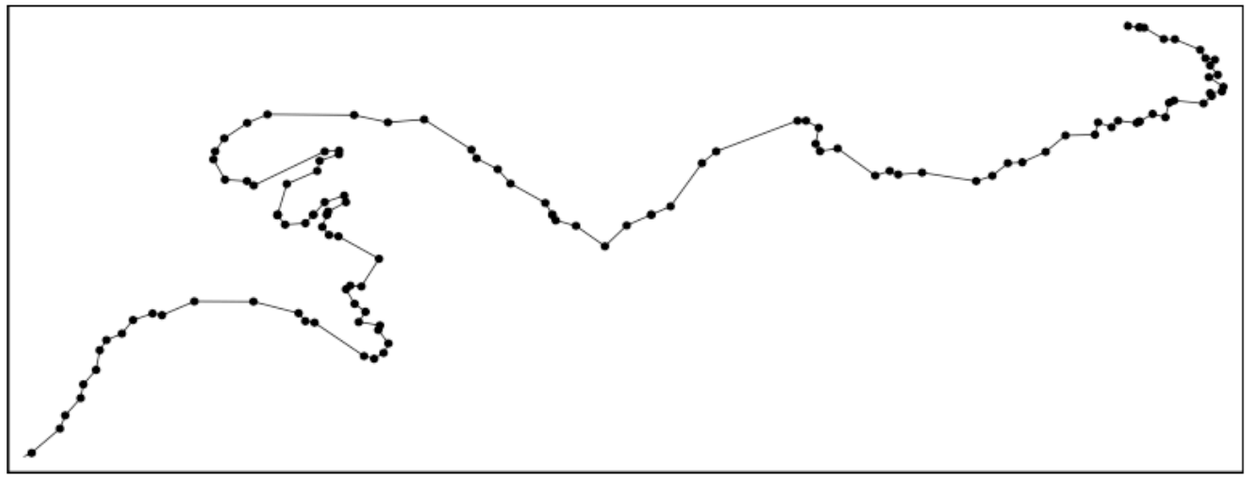

Figure 2.2: Example of a coastline segment whose length we attempt to estimate using both a compass and via digitizing.

REVIEW OF ERROR ANALYSIS

That aside, we can see that the errors accumulate very differently. For the compass length-sum the errors are all dependent (since the aperture is fixed) and we must use the summation rule to find the uncertainty $\delta l = N \cdot 0.025$ cm = 1.25 cm. For the digitizing operations all the uncertainties associated with points 2 through 49 largely cancel and we are left with the uncertainty of the endpoints. Those are clearly independent and hence the uncertainty is $\delta l = (0.12 + 0.12)1/2$ cm = 0.14 cm. The systematic errors using the compass accumulate while the errors in digitizing only affect the end-points. This discussion of digitizing errors is a bit oversimplified, but it does illustrate the difference between the two types of errors and how they accumulate.

2.3.4 Uncertainty in a function

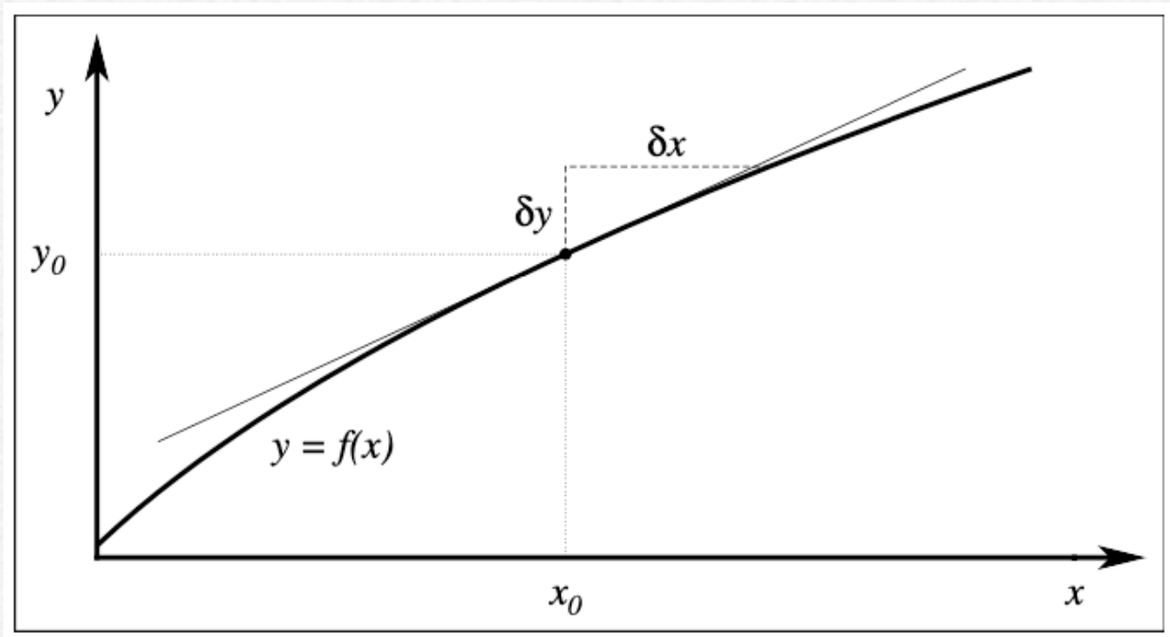

Figure 2.3: As δ*x* becomes very small, the derivative of any well-behaved function can be approximated by a *tangent* at the point (x_0,y_0 = $y(x_0)$); this is the second term in Taylor's expansion.

Many solutions to scientific or engineering problems require the evaluation of functions with our uncertain mea- surements as arguments. If *x* is measured with uncertainty δ*x* and is used to evaluate the function $y = f(x)$, then the uncertainty δ*y* is related to the *derivative* of the function at *x*, i.e.,

REVIEW OF ERROR ANALYSIS

$$\delta y = \left|\frac{df}{dx}\right|_{x_0} \cdot \delta x, \tag{2.34}$$

where the derivative is evaluated at *x*0 (e.g., Figure 2.3).

Example 2–5. Let $y(x) = \cos x$ and $x = 20 \pm 3°$. Following (2.34),

$$\delta y = \left|\frac{d\cos(x)}{dx}\right|_{x=20°} \left(\frac{\pi}{180°}\right) 3° = |-\sin 20°| \frac{3\pi}{180} = 0.342 \cdot 0.0524 \approx 0.02, \tag{2.35}$$

where we have converted the angle from degrees to radians (why?). The final answer then becomes

$$y = \cos(x) = 0.94 \pm 0.02. \tag{2.36}$$

Finally, for a function of multiple variables, *f* (*x*, ..., *z*), we extend our analysis to find

$$\delta f = \sqrt{\left(\frac{\partial f}{\partial x}\delta x\right)^2 + \cdots + \left(\frac{\partial f}{\partial z}\delta z\right)^2} \tag{2.37}$$

when *x*, ..., *z* are all random and independent. As before, δf cannot exceed the ordinary sum

$$\delta f \leq \left|\frac{\partial f}{\partial x}\right| \delta x + \cdots + \left|\frac{\partial f}{\partial z}\right| \delta z, \tag{2.38}$$

which is suitable for dependent measurements.

REVIEW OF ERROR ANALYSIS

Example 2–6. Consider the spherical function

$$f(r,\theta,\phi) = \frac{1}{2} r^2 \cos^2\theta \sin\phi. \tag{2.39}$$

We measured the parameters and found $r = 10 \pm 0.1, \theta = 60 \pm 1°$, and $\phi = 10 \pm 1°$. Using (2.37), the uncertainty in the evaluated expression in (2.39) is found as

$$\delta f = \sqrt{(r\cos^2\theta\sin\phi\delta r)^2 + (-r^2\cos\theta\sin\theta\sin\phi\delta\theta)^2 + \left(\frac{1}{2}r^2\cos^2\theta\cos\phi\delta\phi\right)^2}, \tag{2.40}$$

which means our final estimate of $f(r,\theta,\phi)$ evaluates to

$$f(r,\theta,\phi) = 2.17 \pm 0.26. \tag{2.41}$$

While the simple expressions for uncertainty in a derived quantity (2.21 or 2.23) generally apply, one must be careful with expressions where one or more of the measurements appear in different *subgroups* within the expression. In such cases one must treat the entire expression as a function of several variables and apply the general expression given in (2.37), as our next example illustrates.

Example 2–7. Given the multivariate function

$$f(a,b) = 2ab^2 + \pi b + 1, \tag{2.42}$$

we wish to find its value and uncertainty for $a = 0.3 \pm 0.02$ and $b = 1 \pm 0.01$. Because we have more than one term that depends on b we cannot easily employ the rules in (2.21) and (2.23) but must use (2.37) instead. We first evaluate $f(0.3, 1)$ to be 5.7416... and next evaluate the uncertainty using (2.37):

$$\delta f = \sqrt{(2b^2\delta a)^2 + [(4ab + \pi)\delta b]^2} = 0.05903.... \tag{2.43}$$

Consequently, this yields a final estimate of

$$f = 5.74 \pm 0.06, \tag{2.44}$$

where we have rounded the answer to two decimals only.