Data Type

Data represents measurements of either discrete or continuous quantities (variables). Discrete variables are those having discontinuous or individually distinct possible outcomes. Examples include:

- The flipping of a coin or rolling of dice.
- Counts of individual items or groups of items.
- Ordinal data: Ranked, but interval is not constant, e.g., Moh's hardness scale.
- Nominal data: Categorization or classification of measure, e.g. color of items.

Continuous variables are those having an uninterrupted range of possible outcomes (i.e., with no breaks). They have an infinite number of possible outcomes over a given range. Examples include:

- Percentage of an item (such data, which are closed and forced to a constant sum, sometimes require special care and attention).
- Temperature observations
- Fault strikes
- The earth's geopotential fields.

In addition, much data of interest to earth scientists 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 of these independent variables. Such data represent continuous time series data. They will also be referred to as signals, traces, records, and other names.
Going one step further, the data can also be classified according to how it is recorded for use: Analog signals are those signals which have been recorded continuously (even though one might argue that this is impossible). Discrete data are those which have been recorded at discrete intervals of the independent variable. All data which are represented on computers are discrete.

In either case, the data must (it will) be discretized before it is analyzed in any computational manner that we will employ.

Data limits

1. Data domain – because no phenomena is observed over all time or over all space, data have a limited domain.

2. Data range – it is equally true that no measuring technique can record (or transmit) values which 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 have internal noise). Dynamic Range \((DR)\) is the range over which the data can be measured or exists. This range is usually given on a logarithmic scale in decibels (dB):

\[
DR = 10 \times \log_{10} \left( \frac{\text{largest\_value}}{\text{smallest\_value}} \right) \text{ dB}
\]

One can see that every time the \(DR\) increases by 10, the ratio of the largest to smallest values increases by an order of magnitude. Strictly, this formula is to be applied to data represented as a power measurement (a squared quantity such as variance or square of the signal amplitude). If not, then the formula should be

\[
DR = 20 \times \log_{10} \left( \frac{\text{largest\_value}}{\text{smallest\_value}} \right) \text{ dB}
\]

For example, if the ratio between highest and lowest voltage (or current) was 10, then \(DR = 20 \times \log_{10} (10) = 20 \text{ dB}\). If these same data were represented in watts (power), then, since power is proportional to the ratio of the voltage squared, the ratio would be 100, and \(DR = 10 \times \log_{10} (100) = 20 \text{ dB}\). So, regardless of the manner in which we express our data, we get the same result. In most cases though, except for electrical data, the first formula given is the one to be used (so the data should be expressed as a power).

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

\[
t_0 < t < t_1 \text{ and } |f(t)| < M.
\]

Such functions are always integrable and manageable.

3) Data frequency – finally, most methods of data measurements cannot respond instantly to sudden change. The resulting data are thus said to be band limited. The data will not contain frequency information higher than that representing the fastest response of the recording device.

Noise

In almost all cases, real data such as that with which we will examine contains information other than that strictly desired ("desired" is a key word here since we all know the saying that one scientist's signal is another scientist's noise). Such data consist of samples of random variables. This does not imply that the data are totally random but instead 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, the imprecision of experiments or measuring devices, if we were to give an instrument the same input at two different times, we would get two different outputs because of noise at a later time. In analyzing data, one must never overlook or ignore the role of noise.

One of our main goals frequently desired in data analysis is to detect signal in noise or reduce the degree of noise contamination. We therefore try to enhance the signal-to-noise ratio ($SN$), defined (in decibels) as

$$SN = 10\log_{10}\left(\frac{\text{power of signal}}{\text{power of noise}}\right) \text{ dB}.$$ 

**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 give more or less the same value. 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 where the zero position has not been set 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!

**Analysis**

Analysis means to separate into 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 must have some sort of model of the data in mind (whether this is a conceptual, physical, intuitive, or other type of model is not important), we essentially need some sort of guideline to aid in our analysis. For example, it is not good to take a data set and simply compute the Fourier series because you know about Fourier analysis. One needs to have some 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 purpose of this class.

The following steps is part of most data analysis schemes:

1) Collect data  
2) Perform exploratory data analysis  
3) Reduce data to a few quantities which describe the relevant information  
4) Compare data to hypothesis using appropriate test

We will briefly discuss (1) while reviewing error analysis, which is the study and evaluation of uncertainty in measurements. The main point of step (2) is to familiarize ourselves with the data. This acquaintance is almost always best done by graphing the data. Only a fool will use some black-box technique to compile statistics from data and accept the validity of the 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!
Exploratory Data Analysis

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

1. Scatter plots
2. Schematic plots
3. Histograms
4. "Smoothing" of data
5. Residual plots

We will briefly discuss each of these 5 categories of exploratory techniques. For a complete treatment on EDA, see John Tukey’s *Exploratory Data Analysis* book.

1. Scatter plots

![Scatter plot showing all individual data points - the "raw" data.](image1)

Consider plotting every data value and let it all hang out. Such "scatter" plots show graphically the correlation between points, the orientation of the data, and the spread of clusters. We will later provide a more rigorous definition for what the correlation is; at this stage it is just a visual quantity.

2. Schematic plots

The main objective here is to summarize the data distribution using a simple graph. One very common method is the box and whisker diagram, which graphically presents 5 informative measures of the data. These 5 quantities are the: range of the data (minimum and maximum values), the median (the half-way point), and the hinges (or 25% and 75% quartiles) of a data distribution. Schematically, it can be illustrated as shown in Figure 1.2.

![An example of a "Box and Whisker" diagram. The 5 values gives a visual impression of how the data are distributed.](image2)
As an example of the 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 air and nitrogen produced by decomposition of a chemical compound. 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 1-1) to make a scatter plot and several schematic plots: The already mentioned box-and-whisker diagram and the bar graph.

We will first look at all the data using a scatter plot. It may look something like Fig 1.4.

<table>
<thead>
<tr>
<th>Date</th>
<th>Origin</th>
<th>Purifying Agent</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>29 Nov. '93</td>
<td>NO</td>
<td>Hot iron</td>
<td>2.30143</td>
</tr>
<tr>
<td>5 Dec. '93</td>
<td>&quot;</td>
<td>&quot;</td>
<td>2.29816</td>
</tr>
<tr>
<td>6 Dec. '93</td>
<td>&quot;</td>
<td>&quot;</td>
<td>2.30182</td>
</tr>
<tr>
<td>8 Dec. '93</td>
<td>&quot;</td>
<td>&quot;</td>
<td>2.29890</td>
</tr>
<tr>
<td>12 Dec. '93</td>
<td>Air</td>
<td>&quot;</td>
<td>2.31017</td>
</tr>
<tr>
<td>14 Dec. '93</td>
<td>&quot;</td>
<td>&quot;</td>
<td>2.30986</td>
</tr>
<tr>
<td>19 Dec. '93</td>
<td>&quot;</td>
<td>&quot;</td>
<td>2.31010</td>
</tr>
<tr>
<td>22 Dec. '93</td>
<td>&quot;</td>
<td>&quot;</td>
<td>2.31001</td>
</tr>
<tr>
<td>26 Dec. '93</td>
<td>N₂O</td>
<td>&quot;</td>
<td>2.29889</td>
</tr>
<tr>
<td>28 Dec. '93</td>
<td>&quot;</td>
<td>&quot;</td>
<td>2.29940</td>
</tr>
<tr>
<td>9 Jan. '94</td>
<td>NH₄NO₂</td>
<td>&quot;</td>
<td>2.29849</td>
</tr>
<tr>
<td>13 Jan. '94</td>
<td>&quot;</td>
<td>&quot;</td>
<td>2.29889</td>
</tr>
<tr>
<td>27 Jan. '94</td>
<td>Air</td>
<td>Ferrous hydrate</td>
<td>2.31024</td>
</tr>
<tr>
<td>30 Jan. '94</td>
<td>&quot;</td>
<td>&quot;</td>
<td>2.31030</td>
</tr>
<tr>
<td>1 Feb. '94</td>
<td>&quot;</td>
<td>&quot;</td>
<td>2.31028</td>
</tr>
</tbody>
</table>

Fig. 1.5. Schematic box-and-whisker plot of Lord Rayleigh's data.

Fortunately, in this example we know how to separate out the two data sets based on their origin. It appears that we are better off plotting the two data sets separately instead of as one 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 1-1. It would simply look like Figure 1.6 below:

Figure 1.6. Bar graph of the average values from Lord Rayleigh's Nitrogen data.

In this presentation, it is just barely clear that the weight of "nitrogen" extracted from the air is slightly heavier than nitrogen extracted from chemical compounds. The way it is 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 all the way from zero, while all the variation takes place in the 2.29–2.32 interval. By expanding the scale and choosing a box and whisker plot we concentrate on the differences and may end up with an illustration like the one shown in Figure 1.7:

Figure 1.7. Box-and-whisker diagram of data given in Table 1-1.
It is obvious that the second diagram allows a clearer interpretation than the bar graph. The diagram also benefits from the stretched scale which highlights the difference between the data groupings. In Rayleigh's case the convincing diagram was accepted as strong evidence for the existence of a new element (Argon), and a Nobel price followed later.

3. Histograms

![Histograms](./images/histograms.png)

Figure 1.8. A data set, here a function of distance, is converted into 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 provide an accurate impression of the data distribution even if it is multimodal. One breaks the range into sub-ranges and plots the frequency or occurrences for each range. Obviously, the width of the sub-range determines the level of detailed 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 reasonable, one can plot the individual values inside the histogram bars.

4. Smoothing

The purpose of smoothing is to highlight the general trend of the data and suppress high-frequency oscillations. We'll briefly look at two types of smoothing: Median filter and Hanning filter (A detailed introduction to filtering will be presented in part II of this two-semester series on data analysis). The median filter is typically a three-point filter and simply replaces a point with the median value of the point and its immediate neighbors. The filter then moves 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 2 or more (in that case a wider median filter, say 5-point, would be required). 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.,

\[
y_{\text{filtered}} = \frac{y_{i-1} + 2y_i + y_{i+1}}{4}
\]

Note that while such a filter works well for data that has 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 the median filter, followed by a treatment of the Hanning filter.

5. Residual plots

We can always make the assumption that our data can be decomposed into to parts: A smooth trend plus residuals. The simplest trend is just a straight line. One can easily define such a line by picking two representative points and compute the trend as
We then can form residuals $r_i = y_i - y_{\text{trend}}(x_i)$ (we will later learn more acceptable 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:

$$T: y^n, \ldots, y^1, y^{1/2}, \log(y), y^{-1/2}, y^{-1}, \ldots, y^n$$

The procedure we will follow is: 1) Take $\log(y)$ of the data and plot the values. 2) If the line is concave then choose a transformation closer to $y^n$. 3) If the data is convex then choose a transformation closer to $y^n$. While approximate, this method gives you a feel for how the data varies.

We will conclude this section with a quote from J. Tukey’s book "Exploratory Data Analysis" which basically says it all:

"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 statistics packages without comparing the values to your data. Very often such statistics are based on assumptions about the data distribution which may not be true in your case.

**REVIEW OF ERROR ANALYSIS**

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 most of 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, only make them as small as possible.

**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 other instruments can usually be bracketed between limits. E.g., we may know that $T$ is not less than 23 and no larger than 24. Hence, we shall report the value of $T$ as

$$T = 23.5 \pm 0.5$$

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 large values for $\delta x$ 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 $\delta x$. We need more detailed knowledge of the statistical laws that govern the process of measurement for this, and we will return to this issue later in this class.
Significant Figures

In general, the last significant figure on any stated answer should usually be of the same order of magnitude (same decimal position) as the uncertainty. An example shows how this works: If a measured distance is 6051.78 m with an uncertainty of ± 30 m, you should report the value as \( d = 6050 \pm 30 \) m. For intermediate calculations, use one extra decimal (if calculating by hand) or all available decimals (on calculators and computers).

Fractional Uncertainty

While the statement \( x = x_{\text{best}} \pm \delta x \) indicates the precision of the measurement, it is clear that such a statement will have different meanings depending on the value of \( x_{\text{best}} \) relative to \( \delta x \). Clearly, with \( \delta x = 1 \) m, we imply a different precision if \( x_{\text{best}} = 3 \) m or \( x_{\text{best}} = 1000 \) km. Thus, we should consider the fractional uncertainty,

\[
\frac{\delta x}{x_{\text{best}}}
\]

Note that the fractional uncertainty is a non-dimensional quantity. E.g., one can now report a length of 3 m with a precision of 6 cm as \( l = 3 \) m ± 2%.

Uncertainty in Derived Quantities

Consider the two values \( x \pm \delta x \) and \( y \pm \delta y \). What are the correct expressions for the uncertainties in

a) \( s = x + y \)

b) \( d = x - y \)

c) \( p = x \cdot y \)

d) \( q = x / y \)?

a) We see 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

\[
s = (x + y) \pm (\delta x + \delta y) = s_{\text{best}} \pm \delta s
\]

and similarly for differences (b):

\[
d = (x - y) \pm (\delta x + \delta y) = d_{\text{best}} \pm \delta d
\]

We use the approximate sign = since we may anticipate that the stated uncertainties \( \delta s \) and \( \delta d \) probably over-estimate the true uncertainties in the sum and difference.

c) For the product, we rewrite \( x \) and \( y \) as

\[
x = x_{\text{best}} \left( 1 \pm \frac{\delta x}{x_{\text{best}}} \right)
\]

\[
y = y_{\text{best}} \left( 1 \pm \frac{\delta y}{y_{\text{best}}} \right)
\]

Then the maximum value of \( p \) is
\[ p_{\text{high}} = x_{\text{best}} \left(1 + \frac{\delta x}{x_{\text{best}}} \right) y_{\text{best}} \left(1 + \frac{\delta y}{y_{\text{best}}} \right) = x_{\text{best}} y_{\text{best}} \left(1 + \frac{\delta x}{x_{\text{best}}} + \frac{\delta y}{y_{\text{best}}} + \ldots \right) \]

where the higher order term proportional to \( \delta x \cdot \delta y \) has been ignored. We note that

\[ p_{\text{best}} = x_{\text{best}} \cdot y_{\text{best}} \]

The minimum value is achieved by reversing the signs on \( \delta x, \delta y \). This gives us the uncertainty in \( p \) as

\[ \frac{\delta p}{p_{\text{best}}} \approx \frac{\delta x}{x_{\text{best}}} + \frac{\delta y}{y_{\text{best}}} \]

d) For quotients, the maximum value will be

\[ q_{\text{high}} = \frac{x_{\text{best}} \left(1 + \frac{\delta x}{x_{\text{best}}} \right)}{y_{\text{best}} \left(1 - \frac{\delta y}{y_{\text{best}}} \right)} = q_{\text{best}} \frac{1 + a}{1 - b} \]

Using the binomial theorem, we expand \((1-b)^{-1}\) as \(1+b+b^2+b^3+\ldots\), so

\[ \frac{1 + a}{1 - b} \approx (1 + a)(1 + b) = 1 + a + b + ab \approx 1 + a + b \]

where we again ignore higher-order terms assuming that \((\delta x/|x|) \ll 1\). Similarly, for the minimal value, find

\[ \frac{1 - a}{1 + b} \approx (1 - a)(1 - b) = 1 - a - b + ab \approx 1 - (a + b) \]

Thus

\[ q = q_{\text{best}} \left[1 \pm \left(\frac{\delta x}{x_{\text{best}}} + \frac{\delta y}{y_{\text{best}}} \right) \right] \]

The stated uncertainties are the maximum values possible, but we know these are likely to be too high. We will show later that if we assume that our errors are normally distributed (i.e., we have "Gaussian" errors) and our measurements are independent, a better estimate of uncertainties in sums and differences is

\[ \delta s = \delta d = \sqrt{\left(\delta x\right)^2 + \left(\delta y\right)^2} \]

and for products and quotients

\[ \frac{\delta p}{p_{\text{best}}} = \frac{\delta q}{q_{\text{best}}} = \sqrt{\left(\frac{\delta x}{x_{\text{best}}} \right)^2 + \left(\frac{\delta y}{y_{\text{best}}} \right)^2} \]

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
\[
\frac{\delta p}{p_{\text{best}}} = n \frac{\delta x}{x_{\text{best}}}
\]
since the measurement(s) are dependent. So, in summary, if \( s = x + y + nz - u - v - mw \), then
\[
\delta s = \sqrt{\left(\delta x\right)^2 + \left(\delta y\right)^2 + \left(n \delta z\right)^2 + \left(\delta u\right)^2 + \left(\delta v\right)^2 + \left(m \delta w\right)^2} \quad (1.4)
\]
In any case, even if our assumption of independent measurements is incorrect, \( \delta s \) cannot exceed the ordinary sum
\[
\delta s = \delta x + \delta y + n \delta z + \delta u + \delta v + m \delta w \quad (1.5)
\]
Similarly, if
\[
q = \frac{x \cdot y \cdot z^n}{u \cdot v \cdot w^m}
\]
we find the fractional uncertainty
\[
\frac{\delta q}{q_{\text{best}}} = \sqrt{\left(\frac{\delta x}{x_{\text{best}}}\right)^2 + \left(\frac{\delta y}{y_{\text{best}}}\right)^2 + \left(n \frac{\delta z}{z_{\text{best}}}\right)^2 + \left(\frac{\delta u}{u_{\text{best}}}\right)^2 + \left(\frac{\delta v}{v_{\text{best}}}\right)^2 + \left(m \frac{\delta w}{w_{\text{best}}}\right)^2} \quad (1.6)
\]
While this is the likely error for independent data, we can always say that the fractional uncertainty will be less than the sum
\[
\frac{\delta q}{q_{\text{best}}} = \frac{\delta x}{x_{\text{best}}} + \frac{\delta y}{y_{\text{best}}} + n \frac{\delta z}{z_{\text{best}}} + \frac{\delta u}{u_{\text{best}}} + \frac{\delta v}{v_{\text{best}}} + m \frac{\delta w}{w_{\text{best}}} \quad (1.7)
\]
As an example, by measuring the period \( T \) of a pendulum of length \( \ell \), one can estimate the acceleration of gravity as
\[
g = \frac{4\pi^2 \ell}{T^2}
\]
We find
\[
\frac{\delta g}{g_{\text{best}}} = \sqrt{\left(\frac{\delta \ell}{\ell_{\text{best}}}\right)^2 + \left(\frac{2 \delta T}{T}\right)^2}
\]
Given the values \( \ell = 92.95 \pm 0.10 \text{ cm} \) and \( T = 1.936 \pm 0.004 \text{ s} \), we find
\[
g = \frac{4\pi^2}{1.936^2} \cdot 92.95 \approx 9.79035 \text{ m s}^{-2}
\]
Now,
\[
\frac{\delta g}{g_{\text{best}}} = \sqrt{\left(\frac{0.1}{92.95}\right)^2 + \left(\frac{0.004}{1.936}\right)^2} \approx 0.4\%
\]
The answer, therefore, is
\[
g = 9.79 \pm 0.04 \text{ m s}^{-2}
\]
where we have only used the 2 significant decimals.
As another example, let us imagine measuring the length of a coastline segment using two different methods: (1) Set a compass to a fixed aperture $\Delta x = 1 \pm 0.025$ cm and march along the line, 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 walks so the line length is 50 cm. What is the uncertainty in length for the two methods? Well, let us first state that there will be an uncertainty for both methods that has to do with the under-sampling of short-wavelength "wiggles". That aside, we will see that the errors accumulate very differently. For the compass length-sum the errors are all dependent and we must use the summation rule to find the uncertainty $l = N \cdot 0.025$ 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 independent and hence the uncertainty is $\delta l = (0.1^2 + 0.1^2)^{1/2}$ 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 over-simplified but does illustrate the difference between the two types of errors.

### Uncertainty in a Function

$$y = f(x)$$

Figure 1.10 As $\delta x$ is very small, any well-behaved function can be approximated by a tangent at the point (x, y(x)); this is the first term in Taylor's expansion.

If $x$ is measured with uncertainty $\delta x$ and is used to evaluate the function $y = f(x)$, then the uncertainty $\delta y$ is

$$\delta y = \left| \frac{df}{dx} \right| \delta x$$

(1.8)

where the derivative is evaluated at $x$. As a simple example, let $y = \cos x$ and $x = 20 \pm 3^\circ$. Then

$$\delta y = \left| \frac{d\cos(x)}{dx} \right|_{x=20^\circ} \left( \frac{\pi}{180^\circ} \right)^3 \approx \left| \sin 20^\circ \right| \frac{3\pi}{180} = 0.02$$

where we have converted the angle from degrees to radians. The final answer, then, becomes

$$y = \cos(x) = 0.94 \pm 0.02.$$
Finally, for a function of several variables, \( f(x,\ldots,z) \),
\[
\delta f = \sqrt{\left( \frac{\partial f}{\partial x} \delta x \right)^2 + \cdots + \left( \frac{\partial f}{\partial z} \delta z \right)^2}
\] (1.9)
when \( x,\ldots,z \) are all random and independent. In any case, \( \delta f \) cannot exceed the ordinary sum
\[
\delta f \leq \left| \frac{\partial f}{\partial x} \right| \delta x + \cdots + \left| \frac{\partial f}{\partial z} \right| \delta z
\] (1.10)

Consider the function
\[
f(r, \theta, \phi) = \frac{1}{2} r^2 \cos^2 \theta \sin \phi
\] (1.11)
We measured the parameters and found \( r = 10 \pm 0.1 \), \( \theta = 60 \pm 1^\circ \), and \( \phi = 10 \pm 1^\circ \). Then, the uncertainty in the evaluated expression in (1.11) is found as
\[
\delta f = \sqrt{\left( r \cos^2 \theta \sin \phi \delta r \right)^2 + \left( r^2 \cos \theta \sin \phi \delta \phi \right)^2 + \left( \frac{1}{2} r^2 \cos^2 \theta \cos \phi \delta \phi \right)^2}
\] (1.12)
which means our final estimate of \( f(r, \theta, \phi) \) evaluates to
\[
f(r, \theta, \phi) = 2.17 \pm 0.26
\] (1.13)

**PROBABILITY BASICS**

In data analysis and hypothesis testing, we are concerned with separating the probable from the possible. First, let us have a look at possibilities. In many situations we can either list all the possibilities or say how many such outcomes there are. In evaluating possibilities we are often concerned with finding all the possible choices that are offered. Studying these choices leads us to the "multiplication of choices" rule:

If a choice consists of \( k \) steps, of which the first can be made in \( n_1 \) ways, \ldots, and the \( k_{th} \) in \( n_k \) ways, the total number of choices is \( \Pi n_i \)

This can often be seen most clearly with a tree diagram.

![Tree diagram](image)

Fig. 1.11. Tree diagram for illustrating choices.

The number of choices here are \( 3 \times 4 = 12 \)

**Permutations**

How many ways can we arrange \( r \) objects selected from a set of \( n \) distinct objects?
As an example, we have a tray with 20 rock samples. How many ways can you select 3 samples from the 20? The first rock can be any of 20, the 2nd will be any of 19, while the 3rd is one of 18. The total ways must therefore be $20 \times 19 \times 18 = 6840$ ways.

We can write the number of choices as $20 \times (20-1) \times (20-2)$, and in general we find

$$\text{ways} = n(n-1)(n-2)...(n-r+1) = _nP_r$$  \hspace{1cm} (1.14)

It is convenient to introduce the factorial $n!$ defined as

$$n! = \prod_{i=1}^{n} i$$  \hspace{1cm} (1.15)

In addition, we define 0! to equal 1. We can then rewrite (1.14) as

$$nP_r = \frac{n(n-1)(n-2)...(n-r+1)(n-r)...1}{(n-r)(n-r-1)...1} = \frac{n!}{(n-r)!}$$  \hspace{1cm} (1.16)

This quantity is called the number of permutations of $r$ objects selected from a set of $n$ distinct objects. Let us find how many different hands one can be dealt in poker. With $n = 52$ and $r = 5$, we find

$$52P_5 = \frac{52!}{(52-5)!} = \frac{52!}{47!} = 48 \cdot 49 \cdot 50 \cdot 51 \cdot 52 = 3 \cdot 10^8$$

This assumed that the order in which you received the cards are important.

**Combinations**

Many times, however, we don't care about the exact ordering of the $r$ objects, i.e., abc is the same choice as acb for our purpose. In general, $r$ objects can be arranged in $r!$ different ways \(_nP_r = r!\). Since we are only concerned about which $r$ objects have been selected and not their order, we can use \(_nP_r\) but must divide the result by $r!

$$nC_r = \frac{n^P_r}{r!} = \frac{n!}{r!(n-r)!} = \binom{n}{r}$$  \hspace{1cm} (1.17)

The quantities \(_nC_r\) are called the binomial coefficients. For example, in how many ways can you pick three tide gauge records from 10 stations?

$$10C_3 = \binom{10}{3} = \frac{10!}{3!7!} = \frac{8 \cdot 9 \cdot 10}{1 \cdot 2 \cdot 3} = 8 \cdot 3 \cdot 5 = 120$$

After picking the $r$ objects, $n-r$ objects are left, so consequently, there are as many ways of selecting $n-r$ objects from $n$ as there are for selecting $r$ objects, i.e.,

$$\binom{n}{r} = \binom{n}{n-r}$$  \hspace{1cm} (1.18)

**Probability**

So far we have studied only what is possible in a given situation. We have listed all possibilities or determined how many possibilities there are. However, to be of use to us we need to be able to judge which of the possibilities are probable and which are improbable.
The basic concept of probability can be stated: If there are $n$ possible outcomes or possibilities, and $s$ of these are regarded as favorable (or "success"), then the probability of a success is given by

$$P = \frac{s}{n}$$

This classical definition applies only when all possible outcomes are equally likely. As an example, what are the probability of drawing an ace from a deck of cards?

Answer: \( P = \frac{4}{52} = \frac{1}{13} = 7.7\% \).

How about getting a 3 or a 4 with a balanced die?

Answer: \( s = 2 \) and \( n = 6 \), so \( P = \frac{2}{6} = 33\% \).

While equally likely possibilities are found mostly in games of chance, the classical probability concept also applies to random selections. E.g., If 3 of 20 water samples have been contaminated and you select 4 random samples, what are the probability of picking one of the bad samples?

Answer: We have \( \binom{20}{4} = 4845 \) ways of making the selection of our 4 samples. The number of "favorable" outcomes is \( \binom{17}{3} \)[we pick 3 good samples of the 17 good ones] x \( \binom{3}{1} \)[we pick one of the three bad samples] = 2040. It then follows that the probability is

$$P = \frac{s}{n} = \frac{2040}{4845} = 42\%.$$  

Here we used the rule of multiplicative choices.

Obviously, the classical probability concept won't be useful when some outcomes are more likely than others. A better definition would then be

*The probability of an event is the proportion of the time that events of the same kind will occur in the long run.*

So, when the Weather Service says that the chance of rain on any day in June is 0.2, it is based on past experiences that on average June had 6 days of rain. Another important probability theorem is the law of large numbers, which states

*If a situation, trial, or experiment is repeated again and again, the proportion of successes will tend to approach the probability that any one outcome will be a success.*

which is basically our probability concept in reverse.

Some rules of Probability

In statistics, the set of all possible outcomes of an experiment is called the sample space, usually denoted by the letter \( S \). Any subset of \( S \) is called an event. An event may contain more than 1 item. Sample spaces may be finite or infinite. Two events that have no elements in common are said to be mutually exclusive, meaning they cannot both occur at the same time.

- There are only positive (or zero) probabilities, symbolically written \( P(A) \geq 0 \) for any event \( A \).
- Every sample space has probability 1; \( P(S) = 1 \) where \( P = 1 \) means absolute certainty.
If 2 events are mutually exclusive, the probability that one or the other will occur equals the sum of their probabilities

\[ P(A \cup B) = P(A) + P(B) \]  

(1.19)

Regarding notation, \( \cup \) means Union (read as "OR"), \( \cap \) means intersection ("AND"), and ' means complement ("NOT").

We can furthermore state

\[ P(A) \leq 1, \text{ since absolute certainty is the most we can ask for.} \]
\[ P(A) + P(A') = 1, \text{ since it is certain that an event either will or will not occur} \]

Probabilities and odds

If the probability of an event is \( p \), the odds for its occurrence are

\[ a : b = p / (1 - p) \]

The inverse relation gives

\[ p = \frac{a}{a + b} \]

Additional Addition Rules

The addition rules demonstrated above only holds for mutually exclusive events. Let us now consider a more general case

Fig. 1.12. Venn diagram for probabilities of finding hydrocarbons.

The sketch in Figure 1.12 is a Venn diagram, a handy graphical way of illustrating the various combinations of possibilities and probabilities. The diagram illustrates the probabilities associated with finding hydrocarbons during a hypothetical exploration campaign. We have from the diagram that

\[ P(\text{Oil}) = 0.18 + 0.12 = 0.3 \]
\[ P(\text{Gas}) = 0.24 + 0.12 = 0.36 \]
\[ P(\text{Oil} \cup \text{Gas}) = 0.18 + 0.12 + 0.24 = 0.54 \]

Now, if we used the simple addition rule (1.19) we would find

\[ P(\text{Oil} \cup \text{Gas}) = P(\text{Oil}) + P(\text{Gas}) = 0.3 + 0.36 = 0.66. \]

This value overestimates the probability because finding Oil and finding Gas are not mutually exclusive since we are might find both. We can correct the equation by writing

\[ P(\text{Oil} \cup \text{Gas}) = P(\text{Oil}) + P(\text{Gas}) - P(\text{Oil} \cap \text{Gas}) = 0.3 + 0.36 - 0.12 = 0.54. \]

The general addition rule for probabilities thus becomes
\[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \] (1.20)

Note that if the events are mutually exclusive then \( P(A \cap B) = 0 \) and we recover the original rule.

**Conditional Probability**

We must sometimes evaluate the probability of an event given that another event already has occurred. We write the probability that \( A \) will occur given that \( B \) already has occurred as

\[ P(A | B) = \frac{P(A \cap B)}{P(B)} \] (1.21)

In our exploration example, we can find the probability of finding oil given that gas already has been found as

\[ P(\text{Oil} | \text{Gas}) = \frac{P(\text{Oil} \cap \text{Gas})}{P(\text{Gas})} = \frac{0.12}{0.36} = \frac{1}{3} \]

We can now derive a general multiplication rule from (1.21) by multiplying it by \( P(B) \) which gives

\[ P(A \cap B) = P(B)P(A | B) \]

\[ P(A \cap B) = P(A)P(B | A) \] (1.22)

which states that the probability of both events \( A \) and \( B \) occurring is given by the probability of one event multiplied by the probability that the other event will occur given that the first one already has occurred (occurs, or will occur). This is also called the Joint Probability or Bayes Basic Theorem.

Now, if the events \( A \) and \( B \) are independent events, then the probability that \( A \) will take place is not influenced whether \( B \) has taken place or not, i.e.

\[ P(A|B) = P(A) \]

Substituting this expression into (1.22) we obtain

\[ P(A \cap B) = P(A)P(B) \]

That is, the probability that both \( A \) and \( B \) will occur equals the product of their probabilities. In general, for \( n \) independent events with individual probability \( p_i \), the probability that all \( n \) events occur is

\[ P = \prod_{i=1}^{n} p_i \] (1.23)

E.g. What are the probability of rolling 3 ones in a row with a die? Answer: With \( n = 3 \) and \( p = 1/6 \),

\[ P = \frac{1}{6} \cdot \frac{1}{6} \cdot \frac{1}{6} = 0.005 \]

While \( P(A|B) \) and \( P(B|A) \) may look similar, they can be vastly different. E.g. Let \( A \) be the event of a death on the Bay Bridge, and \( B \) the event of a magnitude 8 earthquake in San Francisco. Then, \( P(A|B) \) is the probability of a fatality on the Bay Bridge given that a large earthquake has taken place nearby, while \( P(B|A) \) is the probability that we’ve had a magnitude 8 quake given a death on the bridge. Clearly \( P(A|B) \) seems more likely than \( P(B|A) \) since we know the former to have happened in the past but can find many other causes of death on the freeway than earthquakes (traffic accidents, heart attacks, gang violence, etc.).
We can arrive at a relation between \( P(B|A) \) and \( P(A|B) \) by equating the two expressions in (1.22). We obtain \( P(A) \cdot P(B|A) = P(B) \cdot P(A|B) \) or

\[
P(B|A) = \frac{P(B) \cdot P(A|B)}{P[A]} \tag{1.24}
\]

This is a useful relation since we may sometimes know one conditional probability but are interested in the other. For example, we may know that salt domes often are associated with large curvatures in the gravity field. However, we may be more interested in the converse: Given that large curvatures in the gravity field exist, what are the probability that salt domes will be the cause of the anomalies?

If there are more than one event \( B_i \) (all mutually exclusive) that are conditionally related to event \( A \), then \( P(A) \) is simply the sum of the conditional probabilities of the \( B_i \), i.e.

\[
P(A) = \sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)
\]

Substituting (1.25) into (1.24) gives

\[
P(B_i|A) = \frac{P(B_i) \cdot P(A|B_i)}{\sum_{i=1}^{n} P(A|B_i) \cdot P(B_i)}
\]

This is the general Bayes Theorem. As an example, let us assume that an unknown fossil fragment was found in a stream bed in Kansas, and a paleontologist would like to send a student field party out to search for a more complete specimen. Unfortunately, the source of the fragment cannot be identified uniquely since it was found below the junction of two dry stream tributaries. The drainage basin \( B_1 \) of the larger stream covers 180 km\(^2\); the other \((B_2)\) only 100 km\(^2\). Based on this we might expect that the probabilities that the fragment came from one of the basins are

\[
P(B_1) = \frac{180}{280} = 0.64
\]

\[
P(B_2) = \frac{100}{280} = 0.36
\]

based solely on the percentage of basin area to total area. However, inspecting a geological map reveals that 35% of the outcropping rocks in the large basin \( B_1 \) are marine, while almost 80% of the outcrops are marine in basin \( B_2 \). We can now state two conditional probabilities

\[
P(A|B_1) = 0.35 \quad \text{Probability of a marine fossil, given it was derived from basin } B_1
\]

\[
P(A|B_2) = 0.80 \quad \text{Probability of a marine fossil, given it was derived from basin } B_2
\]

With these probabilities and Bayes theorem (1.26) we can find the conditional probability that the fossil came from basin \( B_1 \) given that the fossil is marine

\[
P(B_1|A) = \frac{P(A|B_1) \cdot P(B_1)}{P(A|B_1) \cdot P(B_1) + P(A|B_2) \cdot P(B_2)} = \frac{0.35 \cdot 0.64}{0.35 \cdot 0.64 + 0.80 \cdot 0.36} = 0.44
\]

Similarly, the probability of the fossil coming from the smaller basin \( B_2 \)

\[
P(B_2|A) = 0.56.
\]

It therefore seems somewhat more likely that the smaller basin was the source of the fossil. However, \( P(B_1|A) \) and \( P(B_2|A) \) are not very different and depends on the assumptions used to select \( P(B_i) \) and \( P(A|B_i) \).
When discussing exploratory data analysis we mentioned that it is useful to be able to present large data sets using just a few parameters. The box-and-whisker diagram graphically summarized a data distribution. However, it is often necessary to represent a data set by means of a single number which, in its way, is descriptive of the entire data set. We will see there are several ways to select this "representative" value. We will mostly be concerned with measures which somehow describes the center or middle of the data set. These are called estimates of central location.

Population and Samples

If a data set consists of all conceivably possible (or hypothetically possible) observations of a certain phenomenon, we call it a population. A population can be finite or infinite. Any subset of the population is called a sample. Thus a series of 12 coin-tosses is a sample of the potentially unlimited number of tosses in the population. We will most often find that we are analyzing samples of a much larger population, and our aim will be to learn something about the population by studying the smaller sample set.

Measures of location (mean, median, mode)
The best known estimate of central location is called the arithmetic mean, defined as

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$  \hspace{1cm} (1.27)

The mean is also loosely called the "average." When reporting the mean, always say "mean" and not "average" so that people know what you have done. We call $\bar{x}$ the sample mean to distinguish it from the true mean of the population

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$$  \hspace{1cm} (1.28)

The mean has many useful properties, which explains its common use:

- It can always be calculated for any numerical data: It always exists.
- It is unique and straight-forward to calculate.
- It is relatively stable and do not fluctuate much from sample to sample from the same population.
- It lends itself to further statistical treatment: several $\bar{x}$ can later be combined into an overall mean.
- It takes into account every data value.

However, the last property can sometimes be a liability. Should a few data deviate excessively from the body of the data then it doesn't make sense to include them in the sample. A better estimate for the central location may then be the sample median

$$\tilde{x} = \begin{cases} 
  x_{(n/2+1)}, & n \text{ is odd} \\
  \frac{1}{2}(x_{(n/2+1)} + x_{(n/2)}), & n \text{ is even}
\end{cases}$$  \hspace{1cm} (1.29)

Here, the data must be sorted into ascending (or descending) order. We then choose the middle value (or mean of the two middle values for even $n$) as our median.

E.g., consider the sample of sandstone densities (2.30, 2.20, 2.35, 2.25, 2.30, 23.0, 2.25). The median density can be found to be $x = 2.30$, a reasonable value, while the mean $\bar{x} = 5.24$, which is a useless estimate since it is clearly way outside the bulk of the data. For this reason we say that the median is a robust estimate of central location. Here, is rather obvious that the value 23.0, probably a typo, threw off the mean, and we could correct for that and find $\tilde{x} = 2.28$ instead. However, in many cases our data set is very large and we must anticipate that some values might be bad.

The disadvantage of the median is the need to sort the data which can be slow. However, like the mean, the median always exists and is unique.

Our final traditional estimate for central location will be the mode. The mode is defined as the observation that occurs the most frequently. For finding the central location the mode is at a disadvantage since it may not exist (no two values are the same) or may not be unique (Our densities have two modes). Of course, if our data set is expected to have more than one "peak," modal estimates are important, and we will return to that later. The mode is often denoted as $x$. The mean, median and mode of a distribution are related as indicated in Fig. 1.15.
Returning to the mean, it is occasionally the case that some measurements are considered more important than others. It could be that some observations were made with a more precise instrument, or simply that some values are not as well documented as others. These are examples of situations where we use a weighted mean

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

(1.30)

where \(w_i\) is the weight of the \(i\)'th value. If all \(w_i = 1\) we recover the original definition for the mean (1.27). This equation is also convenient when we need to compute the overall, or grand mean based on the means from several data sets. The grand mean based on \(m\) data sets may be written as

$$\bar{x} = \frac{\sum_{i=1}^{m} n_i \bar{x}_i}{\sum_{i=1}^{m} n_i}$$

(1.31)

where the \(n_i\) take the place of the weights in (1.30).

**Measures of variation**

While measures of central location give us an important attribute of our data, it says little about how the data are distributed. We need some way of representing the variation of our observations about the central locations. In the EDA section, we used the range and hinges to indicate data variability. Another way to define the variability would be to compute the deviations form the mean,

$$\Delta x_i = x_i - \bar{x}$$

(1.32)

and take the average of the sum of deviations, \(\frac{1}{n} \sum_{i=1}^{n} \Delta x_i\). However, it turns out that this sum is always 0, which makes it rather useless for our purposes. A more useful quantity might be the mean of the absolute value (\(AD\)) of the deviations:

$$AD = \frac{1}{n} \sum_{i=1}^{n} |\Delta x_i|$$

(1.33)

Because of the absolute value sign this function is non-analytic and often completely ignored by statisticians. You will find very superficial treatment of medians and absolute deviations in most elementary statistics books. However, when dealing with real data that include occasional bad values, the \(AD\) is useful, just as the median is more useful than the mean. However, the most common way to describe variation of a population is to define it as the average squared deviation, i.e., the population variance is...
\[
\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2
\]

and the population standard deviation
\[
\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2}
\]

Most often we will be working with samples rather than the entire population. It turns out that the sample variance \(s^2\) and sample standard deviation \(s\) is given by
\[
s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

Note that we are dividing by \(n - 1\) rather than \(n\). This is because \(\bar{x}\) must be estimated from the samples rather than being a parameter of the population like \(\mu\) and \(N\). This reduces the degrees of freedom by 1; hence we divide by \(n - 1\).

We can now show one property of the mean: It is clear that \(s^2\) depends on the choice for \(\bar{x}\). Let us find the value for \(\bar{x}\) in (1.36) that gives the smallest value for \(s^2\):
\[
f(\bar{x}) = s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2
\]

A function \(f\) has a minimum when \(df/d\bar{x} = 0\) and \(d^2f/d\bar{x}^2 > 0\), so
\[
\frac{df}{d\bar{x}} = \frac{\sum_{i=1}^{n} -2(x_i - \bar{x})}{n-1} = \frac{-2}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}) = 0
\]

which gives
\[
\sum_{i=1}^{n} (x_i - \bar{x}) = 0
\]
or
\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

Since
\[
\frac{d^2f}{dx^2} = \frac{2n}{(n-1)} > 0
\]
\(f\) has a minimum. Thus, we find that the value \(\bar{x}\) that minimizes the standard deviation equals the mean we defined earlier. This is a very useful and important property of the mean. Because \(\bar{x}\) minimizes the squared "misfit", it is also called the least-squares estimate of central location (or \(L_2\) estimate for short). When computing the mean and standard deviation on a computer we do not normally use (1.36) since it requires 2 passes through the data: One to compute \(\bar{x}\) and one to solve (1.36). Rather, we rearrange (1.36) to give
\[
s = \sqrt{\frac{n \sum x_i^2 - 2n \bar{x} \sum x_i + n \bar{x}^2}{n(n-1)}} = \sqrt{\frac{n \sum x_i^2 - (\sum x_i)^2}{n(n-1)}}
\]
OK, so the arithmetic mean is the value that minimizes the sum of squared deviations from the central value. Can we use the same arguments on the mean absolute deviation and find what the best value for \( x \) is? Let's try:

\[
\frac{d}{d\bar{x}} \left( \frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}| \right) = -\frac{1}{n} \sum_{i=1}^{n} \frac{|x_i - \bar{x}|}{x_i - \bar{x}} = 0 \tag{1.38}
\]

The term inside the summation can only take on the value \( \pm 1 \) and 0. Thus, the only \( x \) that can satisfy (1.38) is the one value chosen such that half the \( x_i \) is smaller (giving -1) and half the \( x_i \) is larger (giving +1). Thus, we have proved that the median is the location estimate that minimizes the mean absolute deviation. We will discuss more aspects of robust methods shortly. The median is also called the L_1-estimate of central location.

**Robust Estimation**

The discussion of mean and median brings up the issue of robust estimation: How to calculate a stable and reasonable estimate of central location in the presence of contaminated data. As an indicator of how robust a method is, we will introduce the concept of "breakdown point." It is the smallest fraction of the observations that have to be replaced by outliers to throw the estimator outside reasonable bounds.

We have already seen that even one bad value is enough to throw the mean way off. In our densities of sandstone, we had \( \rho = \{2.2, 2.25, 2.25, 2.3, 2.3, 2.35, 23.0\} \) with \( n = 7 \). If we realized that 23.0 should be 2.3, we find \( \rho = 2.28 \pm 0.05 \), while if we used the \( \rho_i = 23.0 \) we would find \( \rho = 5.24 \pm 7.8 \). The second estimate is obviously far outside the 2.2–2.35 range first found. We can therefore say that the least squares estimate (the mean) has a breakdown value of \( 1/n \); it only takes one outlier to ruin our day. On the other hand, note that the median is 2.3 in both cases, well inside the acceptable interval. It is found that the breakdown point of the median is 50%; We would have to replace half the data with bad outliers to move the estimate of the median outside the range of the original data values.

Apart from the central location estimator, we also want a robust estimate of the spread of the data. Clearly, the classical standard deviation is worthless since only one bad value will make it explode due to the \( x^2 \) effect. From the success of taking the median of a string of numbers rather than summing them up, could we do something similar with the deviations? For example, what value of \( x \) minimizes median \( \{|x_i - x|\} \)? You can probably see for yourselves that \( x \) must equal our old friend the median. Because of the robustness of the median operator, we will often use the quantity the median absolute deviation (MAD) as our robust estimate of "spread" or variation:

\[
MAD = \frac{1}{1.4826} \text{median} |x_i - \bar{x}| \tag{1.39}
\]

where the factor 1.4826 is a correction term that makes the MAD equal to the standard deviation for a normal distribution. Like the median, the MAD has a breakdown point of 50%. The MAD for our example was 0.07 and it remained unchanged by using the contaminated value.

Having robust estimates of location and scale, we can attempt to identify outliers. We may compute the robust standard units

\[
z_i = \frac{x_i - \bar{x}}{\text{MAD}} \tag{1.40}
\]

and compare them to a cutoff value: If \( |z_i| > z_{cut} \) we say we have detected an outlier. The choice for \( z_{cut} \) is to a certain extent arbitrary. It is, however, quite normal to choose \( z_{cut} = 3 \). Chances that \( z_i \) exceeds that of \( z_{cut} \) is very small if the \( z_i \)'s came from a normal distribution. Our normalized densities (with contaminated value) using the mean and \( s \) to compute \( z \) gives
\[ z_{ls} = \{-0.39, -0.38, -0.38, -0.377, -0.377, -0.37, 2.28\} \]

where none of the values qualify as an outlier. Using the median and MAD we find
\[ z_{lm} = \{-1.35, -0.68, -0.68, 0.0, 0.0, 0.68, 280.0\} \]

and we see that the bad sample gives a huge \( z \)-value 2 orders of magnitude larger than any other. Clearly, the least-squares technique alone is not trustworthy when it comes to detecting bad points. The outlier detecting scheme presents us with an elegant technique: First find and remove the outliers from the data, then use classical least-squares techniques on the remaining data points. The resulting statistics are often called the least trimmed squares estimates (LTS). We will return to the concept of robustness while discussing regression.

Inferences about the Mean

How well does our sample mean compare to the true population mean \( \mu \)? An important theorem, called the central limit theorem, states

If \( n \) (the sample size) is large, the theoretical sampling distribution of the mean can be approximated closely with a normal distribution.

This is rather important since it justifies the use of the normal distribution in a wide range of situations. It simply states that the sample mean \( \bar{x} \) is an unbiased estimate of the population mean and that the scatter about \( \mu \) is normally distributed. It can be shown that the standard deviation of the sampling mean, \( s_\bar{x} \) is related to the population deviation \( \sigma \) by:

\[ s_\bar{x} = \frac{\sigma}{\sqrt{n}} \quad (1.41) \]

or

\[ s_\bar{x} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}} \quad (1.42) \]

depending on whether the population is infinite (1.41) or finite (1.42) of size \( N \). Thus, as \( n \) grows large, \( s_\bar{x} \to 0 \). Furthermore, the sampling variance \( s^2 \) has the mean value \( \sigma^2 \) with the standard deviation

\[ \sigma^2_s = \sigma^2 \sqrt{\frac{2}{n-1}} \quad (1.43) \]

which also \( \to 0 \) for large \( n \).

Covariance and Correlation

We found earlier that the sample variance was defined as

\[ s^2_x = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})}{n-1} \]

It is often the case that our data set consists of pairs of properties like sets of (pressure, depth), (temperature, depth), or concentrations of two elements. Denoting the two properties by \( x \) and \( y \), we can compute the variance of each quantity separately. i.e. for \( y \) we find
We can now define the covariance between \( x \) and \( y \) in a similar way:

\[
S_{xy}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}
\]  

(1.44)

While \( s_x \) and \( s_y \) tell us about how the \( x \) and \( y \) values are distributed individually, \( s_{xy} \) tells us how the \( x \) and \( y \) values vary together.

Because the value of the covariance clearly depends on the units of \( x \) and \( y \), it is difficult to state what a covariance has to be to be meaningful. This difficulty is overcome by defining the correlation coefficient \( r \) which normalizes the covariance to fall in the \([-1,1]\) range:

\[
r = \frac{s_{xy}}{s_x s_y}
\]

(1.45)

If \( |r| \) is close to 1 then the variables are strongly correlated. Values of \( r \) close to 0 mean that there is little significant correlation between the data attributes. Fig. 1.16 shows some examples of data pairs and their correlation:

We see that in general, \( r \) will tell us how well the data are "clustered" in some direction. Note in particular example f), which present data that are clearly correlated (all pairs be on a circle), yet \( r = 0 \) This is because \( r \) is a measure of linear relationships between values; a non linear relationship may not register. Thus we must be careful with how we use \( r \) to draw conclusions about the interdependency of two values. E.g. If our \( x-y \) data are governed by a \( y = \sqrt{x} \) law, we will find a fairly good correlation between \( x \) and \( y \) but will be wrong to conclude that \( x \) and \( y \) have a linear relationship (plotting \( y \) versus \( \sqrt{x} \) would give a linear relationship and a much higher value of \( r \)). We will return to correlation under the curve fitting and multiple regression later in the semester.
Moments

Returning to the L₂ estimates, we will briefly introduce the concept of moments. In general, the \( r \)'th moment is defined as

\[
m_r = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^r
\]

From this it can be seen that the mean and variance are the first and second moment, respectively. We will look at some higher order moments one may encounter in the literature. The first is called the skewness (\( SK \)) and it is the third moment:

\[
SK = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s} \right)^3
\]

where we normalize by \( s \) to get dimensionless values for \( SK \). The skewness is used to indicate if our data distribution is asymmetric about the mean. A positive \( SK \) means we have a longer tail to the right of the mean than to the left, and vice versa.

![Examples of data distributions with positive and negative skewness.](image)

Unfortunately, if the data have outliers then the \( SK \) will be very sensitive to these values and consequently be of little use to us. A more robust estimate of skewness is the Pearsonian coefficient of skewness

\[
SK_p = \frac{3(\bar{x} - \hat{x})}{s}
\]

where we basically compare the mean and the median. An even higher order moment is the kurtosis

\[
K = \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s} \right)^4 \right\} - 3
\]

The -3 makes \( K = 0 \) for a normal distribution, which we will discuss shortly. \( K \) attempts to represent the distribution's "sharpness" (\( K > 0 \)) or "flatness" (\( K < 0 \)) of a data distribution. However, for most real data \( K \) can be almost infinite, and should be used only with "well-behaved" data.
Fig 1.18. Examples of distributions with different kurtosis. Distributions with negative $K$ are called platykurtic; a positive $K$ is called leptokurtic. You will of course be pleased to learn that an intermediate case is called mesokurtic.

Probability Distributions

An important concept in statistics and probability is the Probability Distribution. It is a function $P(x)$ which indicate the probability that the event $x$ will take place. $P(x)$ can be a discrete or continuous function. As an example, consider the function $P(x)$, $x=1, 2, ..., 6$, that gives the probability of throwing an $x$ with a balanced die.

$$P(x) = \frac{1}{6}, \ x = 1, 2, ..., 6$$

or for flipping a coin,

$$P(x) = \frac{1}{2}, \ x = \{H, T\}$$

Here we can relate $P(x)$ to the area under the curve.

Two important properties shared by all discrete probability distributions are

$$0 \leq P(x) \leq 1 \quad \text{for all } x$$

$$\sum_{i=1}^{n} P(x_i) = 1$$

Often we are more interested in knowing the probability of a certain outcome after $n$ repeated tries, like "what are the probability of receiving junk mail 3 days a week." To derive such a function, we will assume each event is independent and has the probability $p$. Then, the probability that an event does not occur is the complement, $q = 1 - p$. Thus, the probability of getting $x$ successes (and $n - x$ failures) in $n$ tries is

$$P(x) = p^x q^{n-x}$$

This probability applies to a specific order of all possible outcomes. Since we don't care about the order in which the successful $x$ events occurred, we must scale $P(x)$ by the number of possible combinations of $x$ successes in $n$ tries. We know this amount is $\binom{n}{x}$, so our function becomes

$$P(x) = \binom{n}{x} p^x q^{n-x} = \binom{n}{x} p^x (1-p)^{n-x}, x = 0, 1, ..., n$$  (1.50)
This is known as the binomial probability distribution or simply the binomial distribution and is used to predict the probability that \( x \) events out of \( n \) tries will be successful, given that each independent \( x \) has the probability \( p \) of success.

E.g. What are the chances of drawing 3 red cards in 6 tries from a deck (putting the card back into the deck after each try)? Here \( p = 1/2 \), so

\[
P(3) = \frac{6!}{3!3!} \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^3 = 0.31
\]

One might think that getting half red and half black cards would have higher probability, but remember that we require exactly 3 reds. If we compute the probability of getting 1, 2, or 3 reds separately and used the summation rule to compute the probability that we would draw 1, 2, or 3 red cards then \( P \) would be much higher.

For example, we wish to invest money in a 5-well wildcat exploration program in a virgin basin, where the probability of success per hole is guesstimated to be \( p = 0.1 \). What are the chances we will lose our investment?

\[
P = \binom{5}{0} 0.1^0 (1 - 0.1)^5 = 0.59
\]

Continuous Probability Distributions

While many populations are of a discrete nature (e.g, outcomes of coin tosses, number of micro fossils in a core, etc.), we are very often dealing with observations of a phenomenon that can take on any of a continuous spectrum of values. We may sample the phenomenon at certain points in space-time and thus have discrete observations; nevertheless the underlying probability distribution is continuous.

![Continuous Probability Density Function](image)

Continuous distributions can be thought of as the limit for discrete distributions when the "spacing" between events shrink to zero. Hence we must replace the summations with integrals

\[
\int_{-\infty}^{\infty} p(x) \, dx = 1
\]

(1.51)

Because of their continuous nature, functions such as \( p(x) \) in (1.51) are called probability density functions (PDF). The probability of an event is still defined by the area under the curve, only that now we must integrate to find the area and hence the probability.

E.g., the probability that a random variable will take on a value between \( a - \Delta \) and \( a + \Delta \) is

\[
P(a \pm \Delta) = \int_{a - \Delta}^{a + \Delta} p(x) \, dx
\]

As \( \Delta \to 0 \) we find that the probability goes to zero. Thus the probability of getting exactly \( x = a \) is nil.
The cumulative probability distribution gives the probability that an observation less than or equal to \( a \) will occur. We obtain the integral expression for this distribution by replacing the lower limit by \(-\infty\) and the upper limit by \( a \):

\[
P_c(a) = \int_{-\infty}^{a} p(x) \, dx
\]

(1.52)

Obviously, as \( a \to \infty \), \( P_c(a) \to 1 \).

The normal distribution

So far the function \( p(x) \) has been arbitrary: any continuous function with an area = 1 under the curve would qualify. We will now turn our attention to the best known and most frequently used PDF which goes under the name the "Normal Distribution." Its study dates back to the 18th century investigations into the nature of experimental error. It was found that repeated measurements of the same quantity displayed a surprising degree of regularity. In particular, K. F. Gauss played a major role in developing the theoretical foundations for the normal distribution curve, hence its other name, the Gaussian distribution. It is given by

\[
p(x) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(x-\mu)^2}{2\sigma^2} \right)
\]

(1.53)

where \( \mu \) and \( \sigma \) have been defined previously. The constants before the exponential makes the area under the curve equal to 1. It is often convenient to transform ones \( x \)-values into so-called standard units:

\[
z_i = \frac{x_i - \mu}{\sigma}
\]

(1.54)

in which case (1.53) reduces to

\[
p(z) = \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{z^2}{2} \right)
\]

(1.55)

which has zero mean and unit standard deviation and area.

Fig. 1.21. A normally distributed data set will have almost all of its points within \( \pm 3\sigma \) of the mean.

Given the functional form of \( p(z) \) we can evaluate the cumulative probability that an observation \( a \) will be \( \leq z \):

\[
P(z \leq a) = \int_{-\infty}^{a} p(z) \, dz = \int_{-\infty}^{0} p(z) \, dz + \int_{0}^{a} p(z) \, dz = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_{0}^{a} \exp \left( -\frac{z^2}{2} \right) \, dz
\]
Let
\[ u^2 = \frac{z^2}{2} \quad \text{and} \quad dz = \sqrt{2} du \]

Then
\[
P\left( z \leq a \right) = \frac{1}{2} + \frac{1}{\sqrt{\pi}} \int_{0}^{a} e^{-u^2} du = \frac{1}{2} + \frac{1}{\sqrt{\pi}} \sqrt{\frac{a}{2}} \text{erf}\left( \frac{a}{\sqrt{2}} \right) = \frac{1}{2} \left( 1 + \text{erf}\left( \frac{a}{\sqrt{2}} \right) \right)
\]

It follows that, for any value \( z \), the cumulative distribution is
\[
P_c(z) = \frac{1}{2} \left( 1 + \text{erf}\left( \frac{z}{\sqrt{2}} \right) \right) \quad (1.56)
\]

Furthermore, the probability that \( z \) falls between \( a \) and \( b \) must be
\[
P_c(a \leq z \leq b) = P_c(z \leq b) - P_c(z \leq a) = \frac{1}{2} \left( \text{erf}\left( \frac{b}{\sqrt{2}} \right) - \text{erf}\left( \frac{a}{\sqrt{2}} \right) \right) \quad (1.57)
\]

Example: Investigations on the strength of olivine has given estimates of Young's modulus \( E \) that follow a normal distribution given by \( \mu = 1.0 \times 10^{11} \) Nm and \( \sigma = 1.0 \times 10^{10} \) Nm. What are the probability that a sample estimate \( E \) will be in the interval \( 9.8 \times 10^{10} \) Nm < \( E < 1.1 \times 10^{11} \) Nm? We convert the limits to normal scores and find it corresponds to the interval \(-0.2 < z < 1.0\). Using these values for \( a \) and \( b \) we find the probability (appendix 2.2) to be 0.4436.

The normal distribution is also a very good approximation to the binomial distribution for large \( n \) (actually, when \( np \) and \( (1 - p) n \) > 5). The mean and standard deviation of this distribution becomes
\[
\mu = np \quad \text{and} \quad \sigma = \sqrt{np(1 - p)}
\]

Leading to
\[
P_b(\lambda) = \frac{1}{\sqrt{2\pi np (1 - p)}} e^{\frac{(\lambda - np)^2}{2np(1 - p)}} \quad (1.58)
\]

Example: What is the probability that at least 70 of 100 sand grains will be larger than 0.5 mm if the probability \( p = 0.75 \) that any single grain is that large? Using the approximation (6.1) we find \( \mu = np = 75 \) and \( s = \sqrt{np(1 - p)} = 4.33 \). Converting 69.5 (halfway between 69 and 70) to \( z \) scores gives -1.27, and we find in a table that the probability becomes 0.898.

The Poisson distribution

An approximation to the binomial distribution is used when the probability \( p \) for one event is small and \( n \) is large. Such events are called rare, and the distribution is given as
\[
P(\lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, ..., n \quad (1.59)
\]

where \( \lambda = np \) is the rate of occurrence. The Poisson distribution can be used to evaluate the probabilities for rare events such as large earthquakes, volcano eruptions, and reversals in the geomagnetic field. I.e., the number of floods occurring in a 50-year period has been shown to
follow a Poisson distribution with $\lambda = 2.2$. What are the probability that we will have at least one flood in the next 50 year period? Here, $P = 1 - P_\theta$, the probability of having no flood. Plugging in for $x = 0$ and $\lambda = 2.2$ we find $P_\theta = 0.1108$ so $P = 0.8892$.

The exponential distribution

Another important probability distribution is the exponential distribution. It is given by

$$P_e(x) = \lambda e^{-\lambda x}$$

for some constant $\lambda$. Most of the time we will see it used as a cumulative distribution:

$$P_e(x) = 1 - e^{-\lambda x}$$

Eq. (1.61) gives the probability that the observation $x$ will be in the range $0 \leq x \leq x$. As an example of its use, it has been reported that the heights of Pacific seamounts follow an exponential distribution defined as

$$P(h) = 1 - e^{-h/340}$$

which gives the probability that a seamount is shorter than $h$ meters. As an example, (1.62) then predicts that we might expect that $P(1000) = 1 - e^{-1000/340} \approx 95\%$ of them are less than 1 km tall.

Log-normal Distribution

Many data sets like grain-size of sediments, geochemical concentrations, etc. have a very skewed and long-tailed distribution. In general, such distributions arise when the observed quantities depend on products rather than sums. It therefore follows that the logarithm of the data may be normally distributed. Hence, taking the logarithm of your data may make the transformed distribution look normal. If this is the case, you can apply standard statistical techniques applicable to normal distributions on your log distribution and convert the results (e.g., mean, standard deviation) back to get the proper units.

Inferences about Means

The central limits theorem states that the sample mean drawn from any distribution will be normally distributed even if the data themselves are not normally distributed, and furthermore that the sample mean is an unbiased estimator of the population mean. We can then use our knowledge of the normal distribution to quantify our faith in how precise our sample mean is. We already know that $s_x = s/\sqrt{n}$, so we can state with probability $1-\alpha$ that $\bar{x}$ will differ from $\mu$ by at most $E$, which is given by

$$E = z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

where we have used $s$ as an estimate of $\sigma$. I. e., the chance that $x$ exceeds the $z_{\alpha/2}$ confidence level is $\alpha$. 

These error estimates apply to large samples $n \geq 30$ and infinite populations. In those cases we can use our sample standard deviation $s$ in place of $\sigma$ which we usually do not know. (1.63) can be inverted to yield the sample size necessary to be confident that the error in our sample mean is no larger than $E$. We find

$$n = \left( \frac{z \cdot s}{E} \right)^2$$

(1.64)

The normal score for our sample mean is

$$z = \frac{\bar{x} - \mu}{s / \sqrt{n}}$$

Since this statistic is normally distributed we know that the probability is $1 - \alpha$ that $z$ will take on a value in the interval

$$-z_{\alpha/2} < z < z_{\alpha/2}$$

Plugging in for $z$,

$$-z_{\alpha/2} < \frac{\bar{x} - \mu}{s / \sqrt{n}} < z_{\alpha/2}$$

or

$$\bar{x} - z_{\alpha/2} \cdot \frac{s}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

(1.65)

Eq. (1.65) shows the confidence interval on $\mu$ at the $1 - \alpha$ confidence level. Very often our confidence levels will be $95\%$ ($\sim 2\sigma$) or $99\%$ ($\sim 3\sigma$).

Smaller Samples

The previous section dealt with large ($n \geq 30$) samples where we could assume that $x$ was normally distributed as dictated by the central limits theorem. For smaller samples we must assume that the population we are sampling is normally distributed. We can then base our inferences on the statistics
\[ t = \frac{x - \mu}{\sqrt{\frac{s}{n}}} \]

whose distribution is called the Student \( t \)-distribution. It is similar to the normal distribution but the shape depends on the number of degrees of freedom, \( \nu = n - 1 \). For large \( n \) (and hence \( \nu \)) the \( t \) statistics becomes the same as \( z \) statistics. As with \( z \) statistics, one can find tables with \( t \) values for various combinations of confidence levels and degrees of freedom \(^\dagger\).

**Example:** Given our density estimates from before \{2.2, 2.25, 2.25, 2.3, 2.3, 2.3, 2.35\}, what is the 95\% confidence interval on the sample mean?

**Solution:** We have \( \bar{x} = 2.28 \) with \( s = 0.05 \), \( \alpha = 1 - 95\% = 0.05 \). The degrees of freedom \( \nu = n - 1 = 6 \). A table of \( t \) statistics (e.g., Appendix 2.4) gives \( t_{df = 2} = t_{0.025} \) for \( \nu = 6 \) to be 2.447. Using 6.8 we find our sample mean brackets the population mean thus (with \( t_{df = 2} \) instead of \( z \) and \( s \) instead of \( \sigma \)).

\[
2.28 - 2.447 \cdot \frac{0.05}{\sqrt{7}} < \mu < 2.28 + 2.447 \cdot \frac{0.05}{\sqrt{7}}
\]

or

\[
2.234 < \mu < 2.326
\]

or

\[
\mu = 2.280 \pm 0.046
\]

\(^\dagger\) For insight into what \( t \)-distribution and others really are, get *Numerical Recipes* by Press et al. This excellent book gives both theory and computer code (Choose between C, Fortran, and Pascal).