Chapter 4

Visualising and Quantifying the Properties of Data

4.1 Histograms

It is often very convenient to describe data in a succinct way, either through a visual representation or through a brief quantitative description. Given a sample of data, for example the results of an ensemble of coin flip experiments like those described in Section 3 it can be more instructive to view the results tabulated in a histogram (See figure 4.1) of the data. A histogram is a graphical representation where data are placed in discrete bins. Normally the content of each bin is shown on the vertical axis, and the value of the bin can be read off of the horizontal axis. In this particular coin-flipping experiment there were 27H (heads-up) and 23T (tails-up) obtained. The benefit of drawing a histogram over providing the corresponding data in a table is that one can quickly obtain an impression of the shape of the data as a function of the different bins as well as the relative bin content. For the coin flipping experiment this is not a significant benefit, however the benefits of using histograms becomes more apparent when considering more complex problems.

4.2 Mode, Median, Mean

Continuous data variables are discretized when represented by a histogram, in doing this one does loose information, so there has to be a trade off between the width of bins and the number of entries in a given bin. Indeed the width of all of the bins need not be constant when trying to find a balance between bin content and bin width.

If we take n repeated measurements of some observable x, it is useful to try and quantify our knowledge of the ensemble in terms of a single number to represent the value measured, and a second number to represent the spread of measurements. So our knowledge of the observable x will in general require some central value of the observable, some measure of the spread of the observable, and the units that the observable is measured in. The spread of measurements is discussed in Section 4.3 below, here we discuss the representative value measured from an ensemble of data.

The mode of an ensemble of measurements is the most frequent value obtained. If the measurement in question is of a continuous variable, one has to bin the data in terms of a histogram in order to quantify the modal value of that distribution.

The median value of the ensemble is the value of x where there are an equal number of measurements above and below that point. If there is an odd number of measurements, then this is straight forward and there are

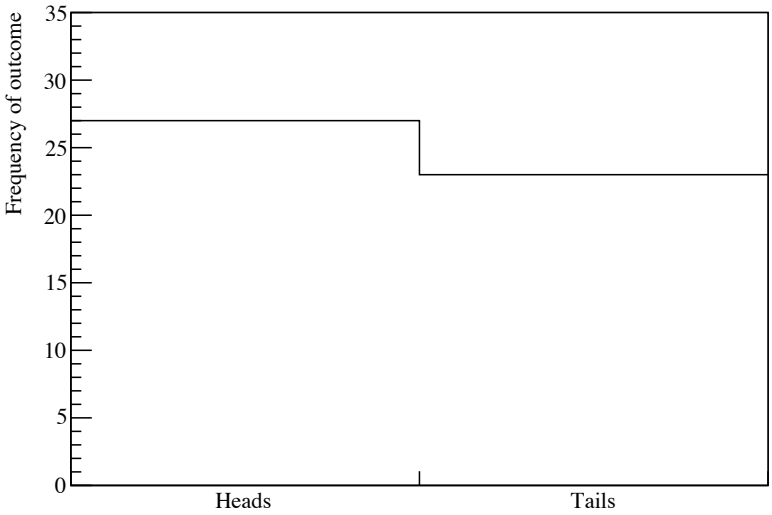

Figure 4.1: The outcome of an ensemble of coin flipping experiments resulting in either Heads or Tails.

$(n-1)/2$ points above and below the median value. If there is an even number of measurements, then the median value is taken as the midpoint between the two most central values. The median value can be useful when it is necessary to rank data (for example in computing upper limits or some correlation coefficients, described later in the course).

A better way to quantify the value measured is to take an arithmetic average of the individual measurements. The arithmetic mean value (usually this is just called the mean for short) of a set of data, is the average value of x computed from the ensemble. The mean value is denoted either by $\overline{x}$ or $\langle x \rangle$ and is given by

$$\overline{x} = \langle x \rangle = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{4.2.1}$$

where x_i is the i^{th} measurement of x. The mean value of a function $f(x)$ can be calculated in the same way using

$$\overline{f} = \frac{1}{n} \sum_{i=1}^{n} f(x_i). \tag{4.2.2}$$

If the function in question is a continuous one, then the average value of the function between $x = a$ and $x = b$ is simply

$$\overline{f} = \frac{1}{b-a} \int_{x=a}^{b} f(x)dx.$$

It is possible to compute the average of a set of binned data, however if rounding occurs in the binning process, then some information is lost and the resulting average will be less precise than obtained using the above formulae.

Exercise: Modify the above formulae (Eq. 4.2.1 and 4.2.2) so that they can be used to compute the arithmetic average of an ensemble of binned data.

Figure 4.2 shows the representation of a sample of data as plotted in a histogram. This figure has two arrows to indicate the mean and mode values. For this particular sample of data the mean is 5.1, and the mode is 6.5. The fact that the mode is greater than the mean is an indication that the data are asymmetric about the mean. We usually refer to such a distribution as being skewed, and in this case the data are skewed to the right. The skewness of a distribution is discussed further in Section 4.4.

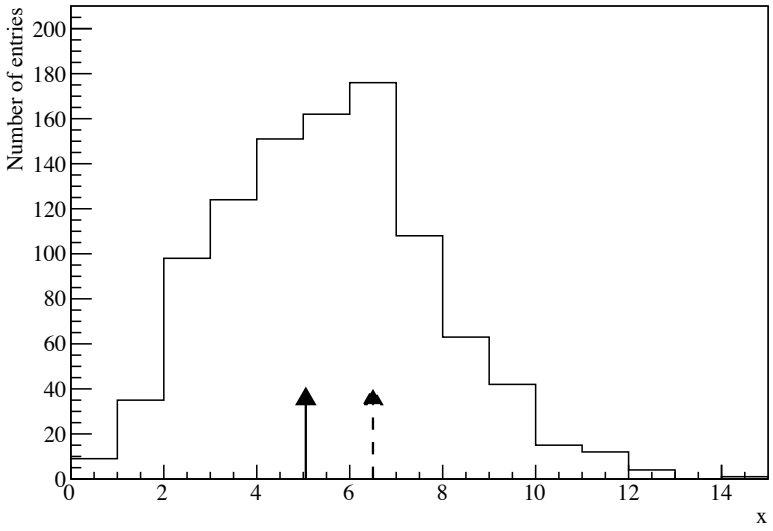

Figure 4.2: A sample of data represented by a histogram, with the mean (solid arrow) and mode (dashed arrow) indicated.

4.3 Quantifying the spread of data

4.3.1 Variance

The mean of an ensemble of data given by $\overline{x}$ doesn't provide any information on how the data are distributed. So any description of a set of data just quoting a value for $\overline{x}$ is incomplete. We need a second number in order to quantify the spread or dispersion of data about the mean value. The average value of the deviations from the mean value is not a useful quantity as by definition this will be zero for a symmetrically distributed sample of data. We can consider the average value of the deviations from the mean squared as a measure of

the spread of our ensemble of measurements. This is called the variance $V(x)$

$$\begin{aligned} V(x) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2, \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2x_i\overline{x} + \overline{x}^2, \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i=1}^{n} 2x_i\overline{x} + \frac{1}{n}\sum_{i=1}^{n}\overline{x}^2, \\ &= \overline{x^2} - 2\overline{x}^2 + \overline{x}^2, \\ &= \overline{x^2} - \overline{x}^2. \end{aligned} \tag{4.3.1}$$

So the variance is given by

$$V(x) = \overline{x^2} - \overline{x}^2.$$

The quantity $(x_i - \overline{x})$ is sometimes referred to as the residual[1] of x.

4.3.2 Standard Deviation

The root of the mean-squared (root-mean-squared) deviation is called the standard deviation, and this is given by:

$$\begin{aligned} \sigma(x) &= \sqrt{V(x)}, \\ &= \sqrt{\overline{x^2} - \overline{x}^2}. \end{aligned} \tag{4.3.2}$$

The standard deviation quantifies the amount by which it is reasonable for a measurement of x to differ from the mean value $\overline{x}$. In general we would expect to have 31.7% of measurements to deviate from the mean value by more than 1σ, 4.5% of measurements to deviate by more than 2σ, and 0.3% of measurements to deviate by more than 3σ. If we performed a set of measurements where our results were more broadly distributed than this, we would worry about what might have gone wrong with the experiment.

The definition of standard deviation given in Eq. 4.3.2 is known to be biased. The level of bias is given by a factor of $(n-1)/n$, so one often multiplies Eq. 4.3.2 by the *Bessel Correction* factor of $n/(n-1)$ in order to remove this bias. The corresponding form for the standard deviation is

$$\sigma(x) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{4.3.3}$$

It can be seen that both forms are identical for $n \to \infty$, and it is better to use the second form for small samples of data. Physicists prefer using the standard deviation rather than variance when describing data as the former has the same units as the observable being measured.

[1]The residual of an observable is the degree to which a constraint equation or other system is satisfied by measurement errors or approximate parameters. In this case it the residual is the difference between the i^{th} measurement of an observable and the arithmetic mean determined from an ensemble of measurements of that observable.

4.3.3 Full Width at Half Maximum: FWHM

Often instead of quantifying a distribution using the variance or standard deviation, scientists will quote the Full Width at Half Maximum (FWHM). This has the advantage that any extreme outliers of the distribution do not contribute to the quantification of the spread of data. As the name suggests, the FWHM is the width of the distribution (the spread above and below the mean) read off of a histogram of data at the points where the distribution falls to half of the maximum. The FWHM can be compared to the standard deviation of a Gaussian distribution by noting that

$$FWHM = 2.35\sigma.$$

however some thought should be put into deciding if such a comparison is meaningful or sensible for a given set of data.

4.4 Skew

In the previous discussion, we have formalized a way to quantify a distribution of data resulting from an ensemble of measurements with only two numbers. One number giving a single value describing the outcome of our measurement the *arithmetic mean*, and one number describing the spread of measurements the *standard deviation*. In the definition of standard deviation, we have implicitly assumed that the data are similar above and below the mean. This can be seen in Eq. 4.3.1 as we compute the sum of squared deviations from the mean. Our quantification of the data doesn't tell us about any possible asymmetry about the mean value. We can try and quantify such asymmetries using the skew or skewness of the data. Skew is a quantity derived from the third power of x and is given by

$$\begin{aligned}
\gamma &= \frac{1}{n\sigma^3}\sum_{i=1}^{n}(x_i - \overline{x})^3, \\
&= \frac{1}{\sigma^3}\overline{(x - \overline{x})^3}, \\
&= \frac{1}{\sigma^3}\left(\overline{x^3} - 3\overline{x}\,\overline{x^2} + 2\overline{x}^3\right),
\end{aligned} \tag{4.4.1}$$

where the steps in going from the second to last line have been omitted.

Exercise: Derive the form given for γ starting from $\frac{1}{n\sigma^3}\sum_{i=1}^{n}(x_i - \overline{x})^3$.

As with the standard deviation, there are alternate forms for the skew of a distribution. A simple form is given by Pearson's Skew which is

$$\text{Skew} = \frac{mean - mode}{\sigma},$$

which is zero by definition when the distribution of data is symmetric and the mean and mode coincide.

Skew can occasionally be useful. Quantities derived from higher order moments exist, but they are not commonly used in physical sciences.

4.5 Measurements of more than one observable

4.5.1 Covariance

The previous discussion only considered measuring some observable quantity x, and how we can simply quantify an ensemble of n such measurements. Often we are faced with the more complicated problem of making measurements that depend on more than one observable. We can construct a variance between two observables x and y, commonly called the covariance $\mathrm{cov}(x, y) = V_{x,y}$ which is given by

$$\begin{aligned} \mathrm{cov}(x, y) &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}), \\ &= \overline{xy} - \overline{x}\,\overline{y}. \end{aligned} \tag{4.5.1}$$

If the values of x and y are independent, then the covariance will be zero. If however large values of x tend to occur with large values of y, and similarly small values of x with small values of y then the covariance will be non-zero. We can extend our notation to an arbitrary number of dimensions, where we can denote the i^{th} data point as $\underline{x}_i$, and $\underline{x}_i = (x_{(1)}, x_{(2)}, \ldots x_{(M)})$. Here the subscript i refers to the data set, and the subscripts with parentheses refer to the particular dimension of the data point. For example in a two-dimensional problem $x_{(1)} = x$ and $x_{(2)} = y$. Using this notation we can write an $M \times M$ covariance matrix (also called the error matrix) for an M dimensional problem with elements given by

$$\begin{aligned} \mathrm{cov}(x_{(i)}, x_{(j)}) &= \frac{1}{n}\sum_{i=1}^{n}(x_{(i)} - \overline{x_{(i)}})(x_{(j)} - \overline{x_{(i)}}), \\ &= \overline{x_{(i)}x_{(j)}} - \overline{x_{(i)}}\,\overline{x_{(j)}}, \\ &= V_{i,j}, \end{aligned}$$

where i and j take values from one to M. The diagonals of this matrix, where $i = j$ are simply the variance of the i^{th} observable.

4.5.2 Correlation

The covariance is a dimensional quantity. It is useful to work with a quantity that is a dimensionless measure of the dependence of one variable on another. The Pearson's correlation coefficient ρ_{xy} is one such variable where the covariance is normalised by the product if standard deviations of x and y:

$$\begin{aligned} \rho_{xy} &= \frac{\mathrm{cov}(x, y)}{\sigma_x \sigma_y}. \\ &= \frac{1}{n\sigma_x\sigma_y}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}). \end{aligned} \tag{4.5.2}$$

Again this can be generalized to an $M \times M$ matrix called the correlation matrix. The elements of the

correlation matrix are given by

$$\begin{aligned}\rho_{ij} &= \frac{\mathrm{cov}\big(x_{(i)}, x_{(j)}\big)}{\sigma_{(i)}\sigma_{(j)}}. \\ &= \frac{V_{i,j}}{\sigma_{(i)}\sigma_{(j)}}.\end{aligned}$$

It can be seen from Eq. 4.5.2 that ρ_{xy} is zero if x and y are independent (following on from the previous discussion with respect to covariance). If x is completely dependent on y, then the possible values of ρ_{xy} are ± 1. In the case that x increases with increasing y $\rho_{xy} = +1$, and for the case where x decreases for increasing y $\rho_{xy} = -1$.

As with the other variables mentioned, there is more than one type of correlation. If the type is not specified then it is assumed that this corresponds to the Pearson correlation coefficient discussed here.

4.5.3 Removing correlations between variables

If we find that two variables are correlated, we can try and remove the correlation by performing a simple translation (equivalent to a rotation about the origin by some angle θ in a two-dimensional space). The solution to this problem can be readily seen by considering the covariance matrix between n variables $\underline{x}$. We wish to map the $\underline{x}$ space onto some un-correlated parameters $\underline{x}'$ in such a way that the covariance matrix V' is diagonal. Another way to express this is that we have to diagonalize the covariance matrix V and apply the corresponding rotation onto all events in the data set $\Omega(\underline{x})$ to produce the data set of rotated elements $\Omega'(\underline{x}')$.

Matrix diagonalisation is a well known problem in linear algebra that one uses a number of methods of increasing rigor to perform. Given a non-singular matrix V, one can write down the eigen-value equation:

$$(V - \lambda I).\underline{r} = 0, \tag{4.5.3}$$

where λ is an eigenvalue, I is the identity matrix, and $\underline{r}$ is a vector of coordinates. The eigenvalues can be determined by solving $det(V - \lambda I) = 0$, and given these it is possible to compute the corresponding eigenvectors. If this method fails to work, there are other methods including single-value-decomposition (SVD) that may suffice. The derivation of the transformation matrix for $n = 2$ is given in Cowan's 'Statistical Data Analysis'.

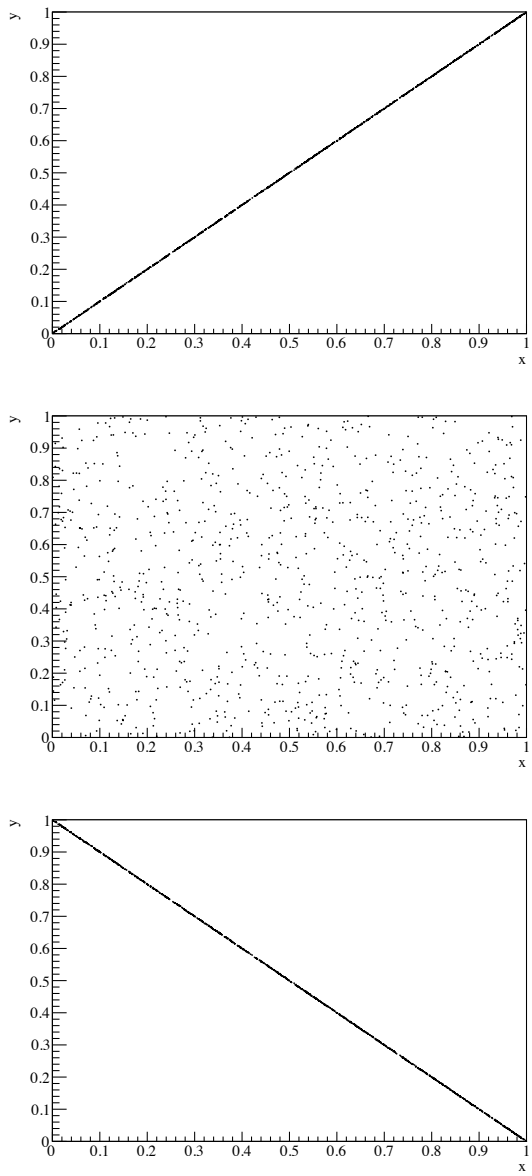

Figure 4.3: From top to bottom, different samples of data with Pearson correlation coefficients of $\rho_{x,y} = -1, 0$ and $+1$.