# CHEM1047 - Week 8 Lecture 1 - Common statistical parameters

☐ *Chapter 14 of Monk and Munro, "Maths for Chemistry", 2nd edition.*

☐ *Sections 21.1-21.3 of Steiner, "The Chemistry Maths Book", 2nd edition.*

Statistical parameters are ubiquitous in chemistry: instrument accuracy, confidence intervals on experimental measurements, estimates for the probabilities of various malfunctions, equipment service life, *etc.* Satistical thermodynamics and quantum mechanics require statistical arguments in their fundamental definitions. Staticstics is a large subject; only a brief introduction is given here.

## 1. Histograms

A *histogram* is a type of plot that illustrates *distributions* of experimental measurement outcomes. The range of values spanned by the data is divided into "bins" and the data values that fall into each bin are counted. The number of instances is plotted, using a bar chart, as a function of the bin location:

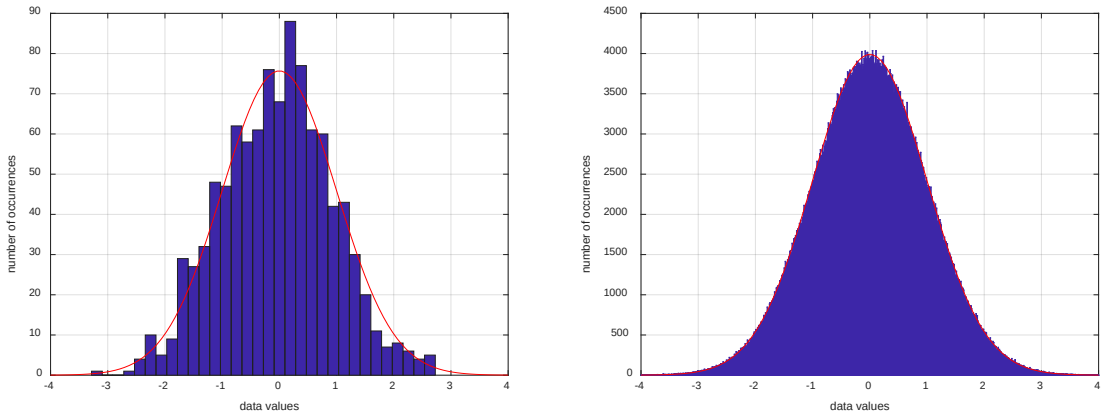


***Figure 1.*** *Typical histograms for data sets that span the range between –4 and +4, and contain $10^3$ measurements (left) and $10^6$ measurements (right). In any given measurement, values between –1 and +1 are a likely outcome; values above +3 or below –3 are unlikely. The red line is the underlying distribution function.*

Setting the number of bins to approximately the square root of the number of data values is usually a good choice. The function that the histogram converges to as the number of measurements becomes larger is called the *probability density function*. Histograms will be used in this lecture to illustrate the various statistical parameters.

## 2. Basic statistical parameters

Given a set of $N$ measurement results $\{a_k\}$, the following statistical parameters are commonly encountered in physical sciences:

1. *Arithmetic mean* – the average of all data points:

$$\langle a \rangle = \frac{1}{N}\sum_{k=1}^{N} a_k \tag{1}$$

The arithmetic mean is *additive* – when two statistically independent random variables are added together, their mean values also add:

$$\langle a+b \rangle = \langle a \rangle + \langle b \rangle \tag{2}$$

2. *Median value* – the value that has 50% of the data set above it and 50% below. The median does not have an analytical expression in terms of the data values of any finite measurement set; it is commonly determined as the value that occurs in the exact middle of a sorted data set. Median is often used instead of the mean for very broadly distributed parameters such as the household income (Figure 2).

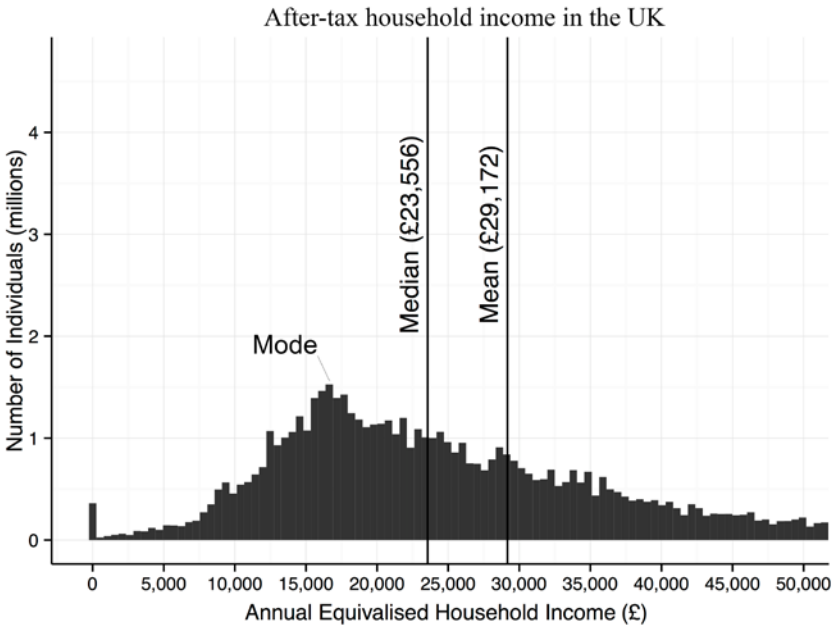


***Figure 2.*** *UK household income histogram for 2013/14. Source: Office of National Statistics. The histogram goes on for a very long time after the £50,000 cut-off value that is used in the figure.*

3. *Mode value* – the value that occurs most often in the data set. Some histograms may have many maxima, in that case the distribution is called *multimodal*. An example of a bimodal distribution is the fragment mass histogram of the fission process of $^{235}$U (Figure 3).

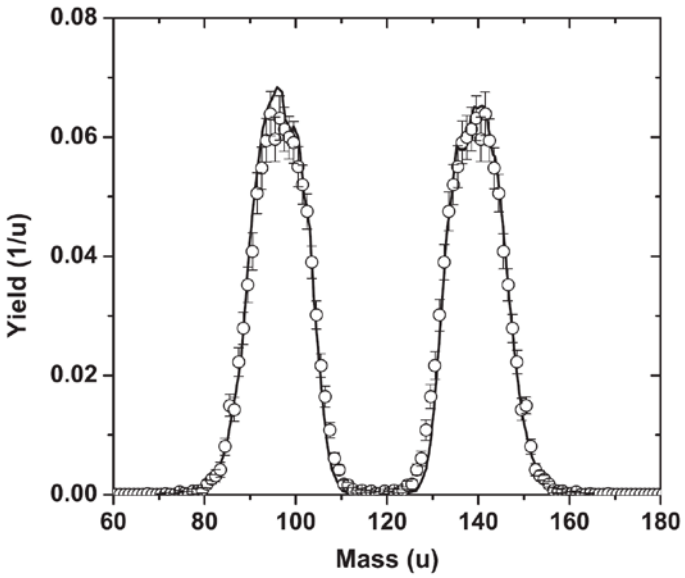


***Figure 3.*** *Distribution of $^{235}$U thermal preneutron emission masses. Source: 10.1103/PhysRevC.81.014607*

One of the two maxima is close to the mass of the radioactive $^{131}$I isotope, which consequently makes up nearly 3% of $^{235}$U fission products. This isotope is highly beta- and gamma-radioactive (the half-life is 8 days). When $^{131}$I is absorbed by the thyroid gland, the resulting radiation exposure can destroy the gland

– hence the iodide tablets in the radiation emergency kits: they contain the stable $^{127}$I isotope that provides competitive inhibition of $^{131}$I uptake by the thyroid. The other maximum on the fission mass distribution contains the notorious $^{90}$Sr isotope that can replace calcium in bones.

4. *Variance* – the average square of the distance between data values and their arithmetic mean:

$$\mathrm{Var}[a]=\frac{1}{N}\sum_{k=1}^{N}\left(a_k-\langle a\rangle\right)^2 \tag{3}$$

Opening the brackets yields a more convenient expression:

$$\begin{gathered}\mathrm{Var}[a]=\frac{1}{N}\sum_{k=1}^{N}\left(a_k-\langle a\rangle\right)^2=\\ =\frac{1}{N}\sum_{k=1}^{N}a_k^2-\frac{2\langle a\rangle}{N}\sum_{k=1}^{N}a_k+\frac{1}{N}\sum_{k=1}^{N}\langle a\rangle^2=\\ =\langle a^2\rangle-\langle a\rangle^2\end{gathered} \tag{4}$$

Although $\mathrm{Var}[a]$ does not have the same dimension as $a$, it is still additive. This property is counter-intuitive, but easy to prove. For statistically independent random variables:

$$\begin{gathered}\mathrm{Var}[a+b]=\left\langle(a+b)^2\right\rangle-\langle a+b\rangle^2=\left\langle a^2+2ab+b^2\right\rangle-\left(\langle a\rangle+\langle b\rangle\right)^2=\\ =\langle a^2\rangle+2\langle a\rangle\langle b\rangle+\langle b^2\rangle-\langle a\rangle^2-2\langle a\rangle\langle b\rangle-\langle b\rangle^2=\\ =\langle a^2\rangle-\langle a\rangle^2+\langle b^2\rangle-\langle b\rangle^2=\mathrm{Var}[a]+\mathrm{Var}[b]\end{gathered} \tag{5}$$

Note that the requirement for the two variables to be statistically independent is essential.

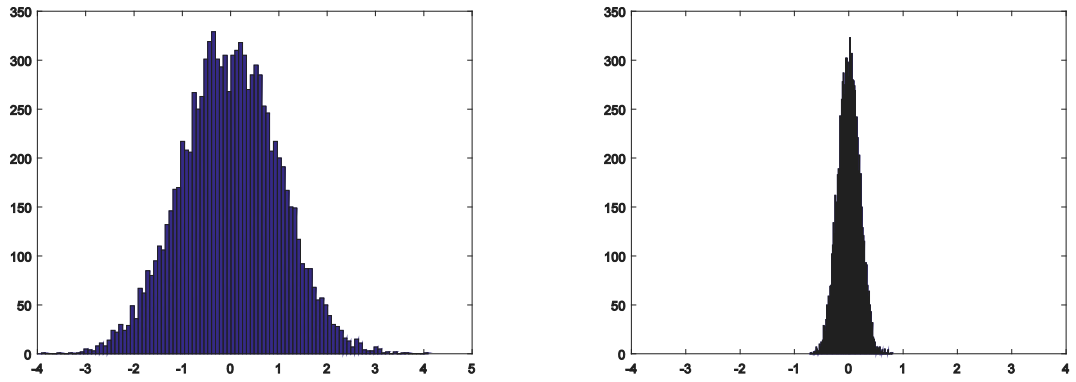


***Figure 4.*** *Examples of histograms of data with large (left) and small (right) standard deviations. Data on the left has $\sigma = 1.0$, data on the right has $\sigma = 0.1$.*

5. *Standard deviation* – the square root of the variance:

$$\sigma_a=\sqrt{\langle a^2\rangle-\langle a\rangle^2} \tag{6}$$

This is a more intuitive measure of the variability in the data because it has the same dimension as the data itself, however this parameter is not additive.

When two independent variables are added together, standard deviations should be converted to variances by squaring them. Variances should be added and the result converted into standard deviation:

$$\sigma_{a+b} = \sqrt{\sigma_a^2 + \sigma_b^2} \tag{7}$$

Examples of distributions with large and small standard deviations are shown in Figure 4.

6. *Standard deviation of the mean* – due to statistical cancellation of errors, the accuracy of the average value of the data set is greater than the accuracy of each individual point. The standard deviation of the mean is related to the standard deviation in the following way:

$$\sigma_{\langle a \rangle} = \sigma_a / \sqrt{N} \tag{8}$$

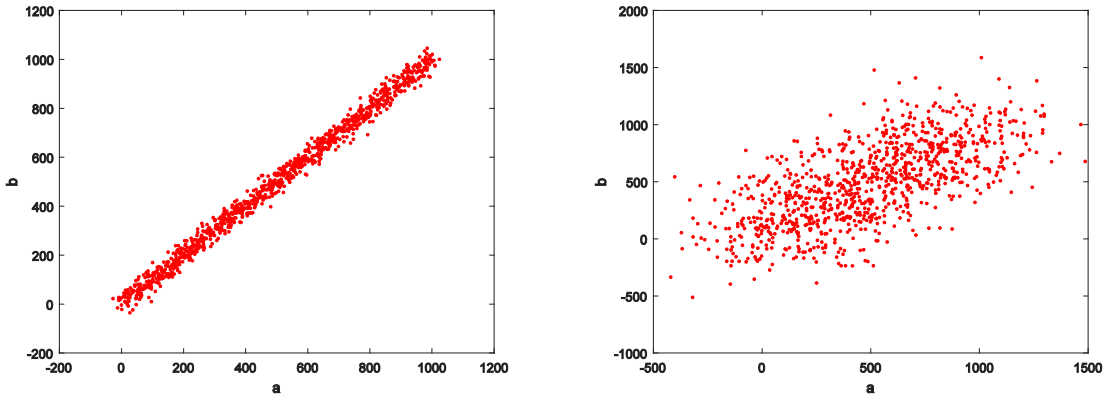


***Figure 5.*** *Data sets showing a high degree of correlation between variables (left) and a low degree of correlation (right). Data on the left has $\rho = 0.995$, data on the right has $\rho = 0.662$.*

7. *Correlation coefficient* – a measure of statistical covariation of two parameters:

$$\rho_{a,b} = \frac{\langle ab \rangle - \langle a \rangle \langle b \rangle}{\sigma_a \sigma_b} \tag{9}$$

A high correlation coefficient indicates that the two variables tend to move in the same direction simultaneously (Figure 5). However, a high correlation coefficient does not imply any causal relationship between the variables – this is illustrated in Figure 6.

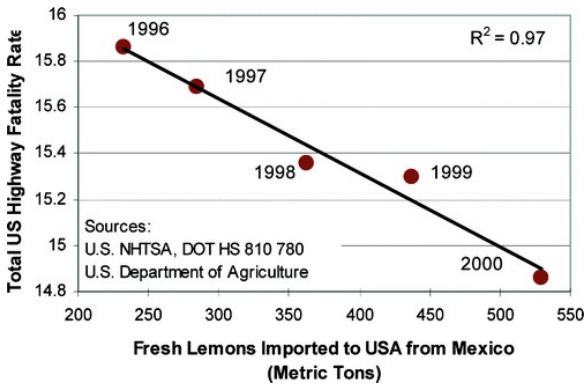


***Figure 6.*** *An illustration to the important principle that correlation does not imply causation.*