CHEM1033 - Week 11 Lecture – Probability and statistics

Chapters 15 and 16 of Monk and Munro, "Maths for Chemistry", 2nd edition.
Sections 21.1-21.4 of Steiner, "The Chemistry Maths Book", 2nd edition.

Histograms

A histogram is a type of plot that illustrates distributions of numerical data. 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 points in each bin is then drawn, usually on a bar chart, as a function of the bin location:

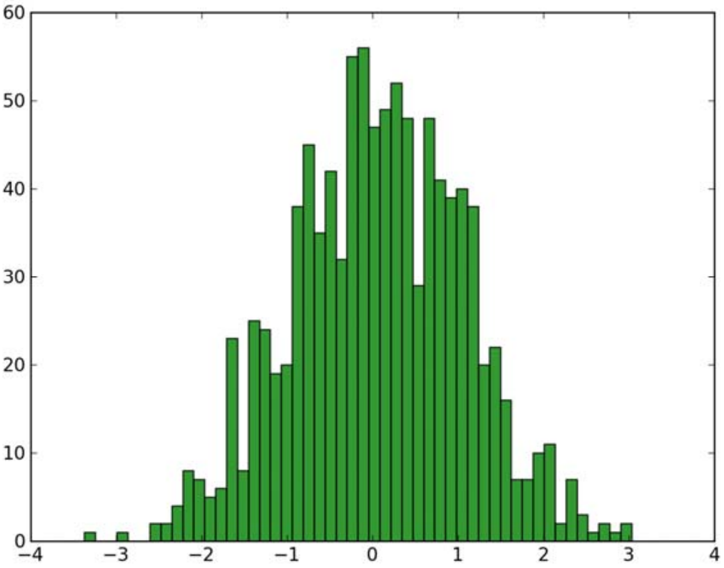

Figure 1. *A common-looking histogram for a data set that spans the range between –4 and 4, and contains 1000 measurements.*

Setting the number of bins to approximately the square root of the number of points is usually a good choice. Histograms are used for visualizing data distributions. They will also be used throughout this document to illustrate the meaning of the various statistical parameters and concepts.

Basic statistical parameters

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

1. <u>Mean value</u>:

$$\mu_a = \frac{1}{N}\sum_{k=1}^{N} a_k \tag{1}$$

2. <u>Median value</u>: the value that has 50% of the result set above it and 50% below. The median does not have an analytical expression in terms of the 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. Median post tax household income in the UK in 2013/14 was £23,556, the mean is £29,172.

3. <u>Mode value</u>: the value that occurs most often in the data set. Some histograms may have many maxima, in that case the distribution is called multimodal.

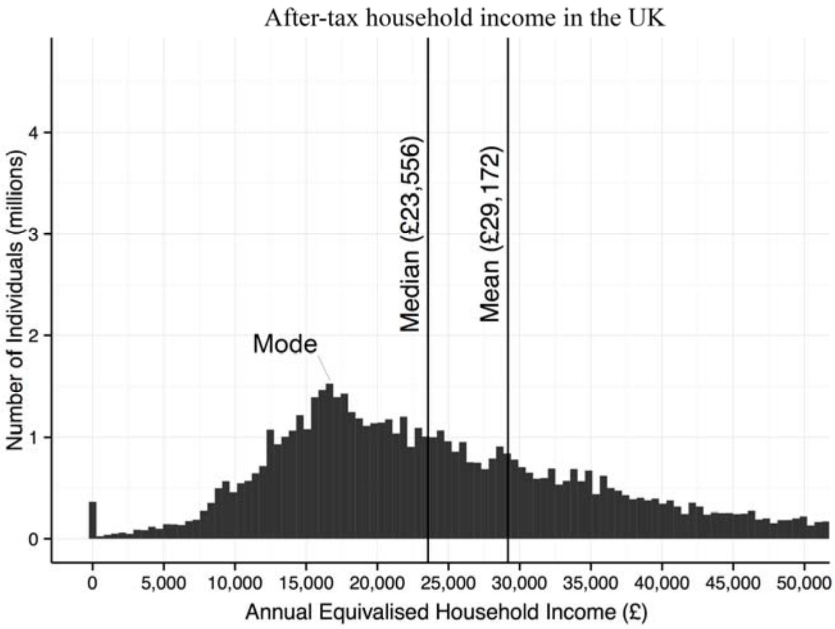

Figure 2. *UK household income histogram for 2013/14. Source: Office of National Statistics. The histogram goes on for a very long time after the £50,000 cut-off value that is used in the figure.*

4. Dispersion: also called variance, a measure of the spread in the data values around the mean value.

$$D_a = \frac{1}{N-1}\sum_{k=1}^{N}\left(a_k - \mu_a\right)^2 \tag{2}$$

Although D_a does not have the same dimension as the data values themselves, it has the advantage of being additive, *i.e.* when two random variables are added together, their variances also add up.

5. Standard deviation: defined as the square root of the variance.

$$\sigma_a = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N}\left(a_k - \mu_a\right)^2} \tag{3}$$

This is a more intuitive measure of the variability in the data, however this parameter is not additive – when two independent variables are added together, standard deviations should be converted to variances, variances should be added up and the result converted back into standard deviations:

$$\sigma_{a+b} = \sqrt{\sigma_a^2 + \sigma_b^2} \tag{4}$$

In the case when a variable is multiplied by an exactly known constant k, its standard deviation is multiplied by the same constant:

$$\sigma_{ka} = k\sigma_a \tag{5}$$

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. It may be demonstrated that the standard deviation of the mean is related to the standard deviation in the following way:

$$\sigma_{\mu_a} = \sigma_a \big/ \sqrt{N} \tag{6}$$

7. Correlation coefficient: a measure of statistical covariation of two parameters. A high correlation coefficient indicates that the two variables tend to move in the same direction simultaneously, but does not, in general, imply any causal relationship between the two parameters.

$$\rho_{a,b} = \frac{1}{\sigma_a \sigma_b} \frac{1}{N} \sum_{k=1}^{N} \left(a_k - \mu_a\right)\left(b_k - \mu_b\right) \tag{7}$$

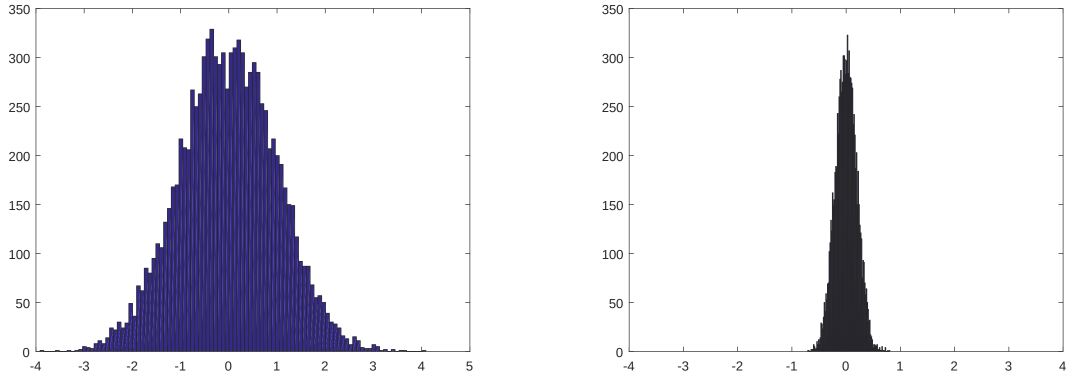

Figure 3. *Examples of histograms of data with large (left) and small (right) standard deviations. Data on the left has σ = 1.0, data on the right has σ = 0.1.*

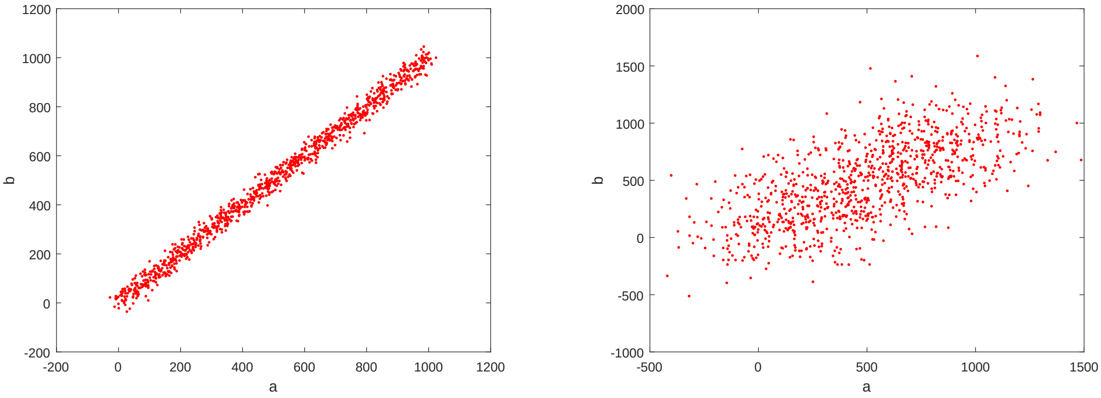

Figure 4. *Data sets showing a high degree of correlation between variables (left) and a low degree of correlation (right). Data on the left has ρ = 0.995, data on the right has ρ = 0.662.*

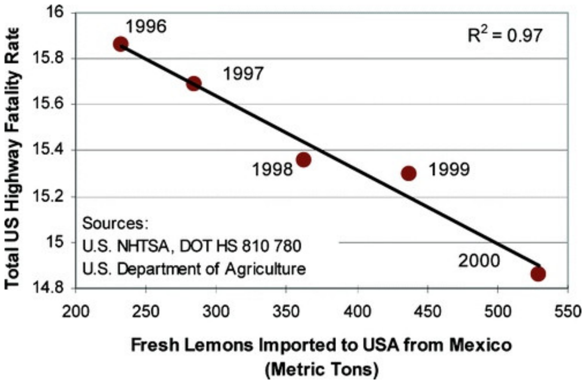

Figure 5. *An illustration to the very important principle that correlation does not imply causation.*

Linear propagation of errors

A problem that is often encountered in physical sciences is establishing the uncertainly in a function of uncertain parameters. More precisely, given the arguments $\{x, y, z, \ldots\}$ with standard deviations $\{\sigma_x, \sigma_y, \sigma_z, \ldots\}$, the task is to determine the standard deviation σ_f of the function $f\left(x, y, z, \ldots\right)$. This problem has no general analytical solution. However, in a special case when the function is approximately linear in the vicinity of the point $\{x, y, z, \ldots\}$, a multi-dimensional Taylor expansion may be used, followed by Equations (4) and (5):

$$\sigma_f \approx \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial f}{\partial y}\right)^2 \sigma_y^2 + \left(\frac{\partial f}{\partial z}\right)^2 \sigma_z^2 + \ldots} \tag{8}$$

In situations where the Taylor series is not a good approximation, the only consistent way of determining the standard deviation of a function of uncertain parameters is the *Monte-Carlo method*: the parameters are varied within their statistical distributions and the function is recomputed until sufficient statistics is accumulated to determine the distribution parameters of the function.

Manipulating probabilities

The probability $P(E)$ of an event E that belongs to a set of events F is defined as the measure of likeness that an event would occur. Probabilities vary from 0 (impossible event) to 1 (certain event). The following rules are used when manipulating probabilities:

1. If the probability of the event A occurring is $P(A)$, then the probability of the event not occurring is

$$P(\neg A) = 1 - P(A) \tag{9}$$

2. For independent events $A, B \in F$ the joint probability of both events occurring is

$$P(A \cap B) = P(A)P(B) \tag{10}$$

For example, the probability of having two heads in two independent coin tosses is $(1/2)(1/2) = 1/4$.

3. For mutually exclusive events $A, B \in F$, the probability of one of the events occurring is

$$P(A \cup B) = P(A) + P(B) \tag{11}$$

If the events are not mutually exclusive, then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{12}$$

For example, the probability of rolling 5 or 6 on a twenty-sided dice is $(1/20) + (1/20) = 1/10$.

Example: in one really competitive job market, the probability of landing a specific type of job for a reasonably qualified candidate who has applied is just 10%. Assuming that twenty applications have all been independently and diligently prepared by the same individual, find the probability of landing that type of job after twenty applications.

Solution: if the probability of success in a single application is 0.10, then the probability of failure (Rule 1) is 0.9, then the probability of 20 independent failures is 0.90^{20} = 0.12, then the probability of 20 sequential failures not happening is 1 – 0.12 = 0.88, *i.e.* there is an 88% chance of at least one application succeeding.

Basic Bayesian analysis

The following rule was formulated by Bayes for conditional probability $P(A|B)$, which is defined as the probability of event A given certain knowledge that event B has occurred:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{13}$$

Example: the population of flufficles on the third planet of Sirius has a peculiar problem – their prison population does not appear to be representative of their general population. Specifically, of those flufficles that are in prison, 13.7% are shiny and the rest are dull. The fraction of shiny flufficles in the general population of the planet, however, is only 2.9%. This is worrying some politicians. Using Bayesian analysis, estimate the coefficient α in the following relation:

$$P(\text{criminal} \mid \text{shiny}) = \alpha P(\text{criminal})$$

and interpret the outcome in practical terms.

Solution: using Equation (13), we obtain

$$P(\text{criminal} \mid \text{shiny}) = \frac{P(\text{shiny} \mid \text{criminal})}{P(\text{shiny})} P(\text{criminal})$$

The probability of a criminal flufficle being shiny is $P(\text{shiny} \mid \text{criminal}) = 0.137$ and the probability of any flufficle being shiny is $P(\text{shiny}) = 0.029$, meaning that $\alpha = 4.7$. Statistically speaking, a shiny flufficle on Sirius 3 is almost five times more likely to be up to no good than your average flufficle!

Week 15 workshop exercises

Steiner, 2nd edition: section 21.12, problems 1-10.

Extra difficulty exercises for the brave

Steiner, 2nd edition: section 21.12, problems 12-15.