Chemometrics

MA4605

Week 2. Lecture 3. The normal distribution

September 12, 2011

In previous lecture we saw that it is necessary to make repeated measurement in analytical experiments in order to reveal the presence of random errors.
Consider the 5 replicate titrations done by each student (A-D).

A	10.08	10.11	10.09	10.10	10.12
B	9.88	10.14	10.02	9.80	10.21
C	10.19	9.79	9.69	10.05	9.78
D	10.04	9.98	10.02	9.97	10.04

Two criteria used to compare results: average value and the degree of dispersion.

- Sample Mean: $\overline{x} = \frac{\Sigma x_i}{n}$ = 10.10 for student A
- Sample Standard Deviation: $s = \sqrt{\frac{\Sigma(x_i - \overline{x})^2}{n-1}}$ = 0.0158

Student	Mean	St.Dev.	Comment
A	10.10	0.0158	Precision, biased
B	10.01	0.1717	Imprecision, unbiased
C	9.9	0.2104	Imprecision, biased
D	10.01	0.0331	Precision, unbiased

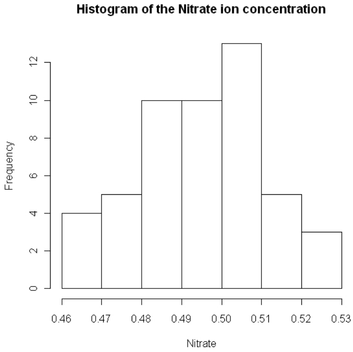
The empirical distribution of repeated measurements

Although the ST DEV gives a measure of the spread the data around the mean value, it does NOT indicate the shape of the distribution.
To illustrate this we consider the following example.
Example: 50 replicate determinations of the nitrate ion concentration in a particular water specimen are contained in *Table*2_1.

0.51	0.51	0.51	0.50	0.51	0.49	0.52	0.53	0.50	0.47
0.51	0.52	0.53	0.48	0.49	0.50	0.52	0.49	0.49	0.50
0.49	0.48	0.46	0.49	0.49	0.48	0.49	0.49	0.51	0.47
0.51	0.51	0.51	0.48	0.50	0.47	0.50	0.51	0.49	0.48
0.51	0.50	0.50	0.53	0.52	0.52	0.50	0.50	0.51	0.51

- Mean = 0.4998
- St Dev = 0.01647385

The distribution of the results can be most easily appreciated by drawing a histogram.



- This shows that the measurements are roughly symmetrical about the mean with most of the measurements clustered around the centre.
- This histogram can be easily approximated by a Normal density curve and we can use the Normal distribution to calculate probabilities associated with the intervals of interest.

The normal probability distribution has the following characteristics:

- the highest point on the normal curve is at the mean, which is also the median and mode of the distribution

The normal probability distribution has the following characteristics:

- the highest point on the normal curve is at the mean, which is also the median and mode of the distribution
- the mean of the distribution can be any numerical value: negative, zero or positive

The normal probability distribution has the following characteristics:

- the highest point on the normal curve is at the mean, which is also the median and mode of the distribution
- the mean of the distribution can be any numerical value: negative, zero or positive
- the normal probability distribution is symmetric: the shape of the curve to the left of the mean is a mirror image of the shape of the curve to the right of the mean

The normal probability distribution has the following characteristics:

- the highest point on the normal curve is at the mean, which is also the median and mode of the distribution
- the mean of the distribution can be any numerical value: negative, zero or positive
- the normal probability distribution is symmetric: the shape of the curve to the left of the mean is a mirror image of the shape of the curve to the right of the mean
- the standard deviation determines the width of the area under the curve. A distribution with a larger standard deviation will have a wider, flatter curve, showing more dispersion in the data

The normal probability distribution has the following characteristics:

- the highest point on the normal curve is at the mean, which is also the median and mode of the distribution
- the mean of the distribution can be any numerical value: negative, zero or positive
- the normal probability distribution is symmetric: the shape of the curve to the left of the mean is a mirror image of the shape of the curve to the right of the mean
- the standard deviation determines the width of the area under the curve. A distribution with a larger standard deviation will have a wider, flatter curve, showing more dispersion in the data
- the total area under the curve is 1

Probabilities of the normal variable.

Probabilities that the normal variable will have a value in a given interval or range are given by areas under the curve.
Probabilities for some commonly used intervals are:

- 68% of the time, a normal variable has a value within plus or minus one standard deviation of its mean
- 95% of the time, a normal variable has a value within plus or minus 1.96 standard deviations of its mean
- 99% of the time, a normal variable has a value within plus or minus 2.58 standard deviations of its mean

The Normal density function

The density function of a Normal distribution with mean μ and standard deviation σ is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp - \frac{(x-\mu)^2}{2\sigma^2} \tag{1}$$

The density function of the Normal **standard** distribution with mean μ=0 and standard deviation σ=1 is a simplified version of equation (1)

$$f(x) = \frac{1}{\sqrt{2\pi}} exp - \frac{x^2}{2} \tag{2}$$

Central Limit Theorem

Why is the normal density so important?
It describes the behaviour of the sample means regardless of the shape of the population from which we sample.

Central Limit Theorem:
When random samples of size **n** are drawn from any distribution, having mean μ and standard deviation σ, then if **n** is large, the sample means tend to form a normal distribution, with mean μ and standard error $\frac{\sigma}{\sqrt{n}}$.

Find a simulation that shows how the Central Limit Theorem works at:
http://onlinestatbook.com/stat_sim/sampling_dist/index.html