Review of normal probability distribution

(Part I)

Introduction

In the newly adopted ISO/IEC 17025 accreditation standards, a new requirement has caught most laboratory personnel by surprise, i.e. the requirement of applying *decision rule* in making a conformity statement in the test report issued, if required. Based on the measurement result obtained, we can be asked by your clients to decide for them whether it indicates compliance or non-compliance with a specification or a regulatory limit.

For example, the trace concentration of a pesticide residue in a vegetable sample submitted for analysis can be used to assess compliance with an upper food safety limit stipulated in the national Food Regulations.

In such situation, we need to take the measurement uncertainty into account. We can use its expanded uncertainty, $\pm U$ to infer the probability density function for the measured value, showing whether there is a larger probability of the value lying near the center of the expanded uncertainty interval than near the ends. However, there is always a risk associated with making a wrong decision on conformity acceptance.

The normal probability distribution function (also known as the Gaussian distribution function) is arguably the most commonly used distribution in statistics. This is partly because the normal distribution is a reasonable description of how many continuous variables are distributed in reality, from industrial process variation to laboratory testing data. It is indeed an important tool to describe randomness of continuous data variation in laboratory analysis. In fact, it turns out to describe many types of data very well, not the least biological data.

Secondly, under specific conditions, we may assume that sampling distributions of statistics, such as the sample means are normally distributed, even if the samples are drawn from populations that are not normally distributed. This is credited to "the Central Limit Theorem", stating that averages are approximately normally distributed, (almost) no matter the properties of the original variables.

This theorem lays down the statistical properties of sample means as follows:

a. **The sample mean is an unbiased estimate.** That is, sample means on average "hit" the right population mean. That is to say by taking a very large number of different samples, the mean of the sample means will be the population mean;
b. The sample mean is a consistent estimate. That is, sample means get more and more precise as the sample size increases, while the standard deviation of population mean decreases.

What is the normal distribution?

Laboratory analysts can testify that replicated measurement results can often be expected to show more values nearer the mean, and the results are distributed quite evenly about the mean, with about half falling greater than the mean and half falling less than the mean.

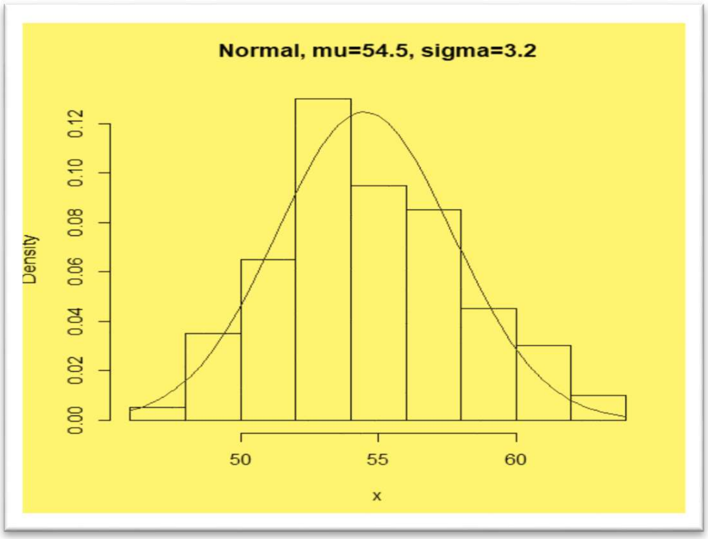
Data that follows this random distribution about a mean can be described by a normal, or Gaussian, probability density function (pdf):

$$y = f(X \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]} \qquad (1)$$

Note that the pdf is a function of x – the values that can be taken by the data. And, the notation for a normal distribution is written as $N(\mu, \sigma^2)$. For more detailed explanation, see below.

A probability density function is defined in terms of its area; the probability of finding a result between two values of x (say, a and b) is the area under the pdf between a and b. The characteristic shape of this pdf is the familiar "bell-shaped curve". Figure 1 below shows a histogram of 100 repeated results with a mean value μ of 54.5 and standard deviation σ of 3.2, and a fitted curve.

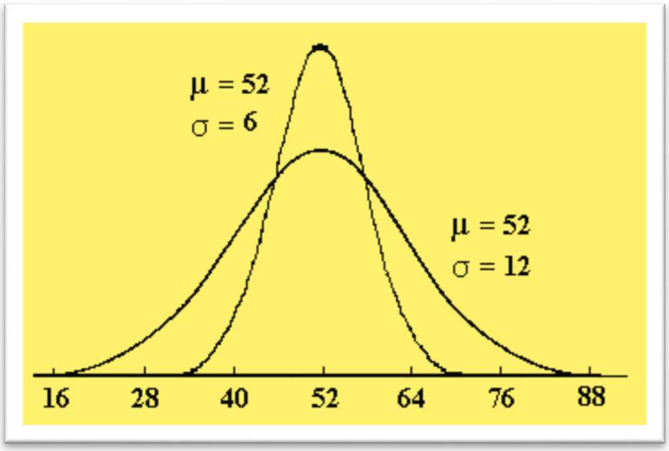
Figure 1 A histogram fitted with normal distribution curve



There is no need to remember this complicated formula (1), but some of its general properties are important:

a. The curve is symmetrical around the population mean μ, so values smaller than μ are just as likely as values larger than μ. And, the greater the population standard deviation σ the greater the spread of the curve. In other words, the larger σ the more likely are observations far from μ. See an illustration as shown in Fig 2.

Figure 2 Two normal distribution curves with same mean value μ =52, but σ = 6 and 12, respectively.

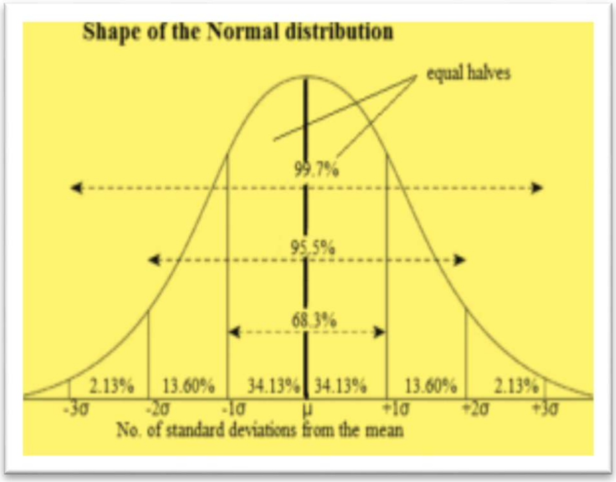


b. Whatever the values of mean μ and standard deviation σ, the normal distribution exhibits the following properties:

- Approximately 68% of the population values lie within $\underline{+}1\sigma$ of the mean
- Approximately 95% of the population values lie within $\underline{+}2\sigma$ of the mean
- Approximately 99.7% of the population values lie within $\underline{+}3\sigma$ of the mean

These properties are illustrated in Figure 3.

Figure 3 A property of the normal distribution: different percentages of areas under the curve with $\underline{+}1\sigma$, $\underline{+}2\sigma$ and $\underline{+}3\sigma$.



(to be continued….)