Review of normal probability distribution

(Part II)

How do we get all these information?

First, let us look at the relationship between probabilities and areas under the density curve.

Equation (1) above shows the probability density function for a normal distribution of mean μ and standard deviation σ or variance σ^2. When we integrate this probability density function, we actually obtain the related cumulative distribution function, giving the probability of a value *up to* x occurring. This is being shown as the area under the curve.

Mathematically, the area is written as an integral, so the relationship between the probability and area can be written as

$$P(a < X < b) = \int_a^b f(x)dx \qquad a \leq b \tag{2}$$

That means probabilities are non-zero for any interval since $f(x) > 0$ for all x, whilst any single point, say a, has probability zero:

$$P(X = a) = \int_a^a f(x)dx = 0 \tag{3}$$

For integrating with $x = -\infty$ to $x = +\infty$, we have an interesting result:

$$P(-\infty < X < \infty) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]} dx = 1.00 \tag{4}$$

In short, if we assume that $x_1, ..., x_n$ are drawn according to the density

(equation (1)), for any such sample we can compute the same mean $\bar{x}$ and the sample standard deviation s; As n increases, the sample mean will approach μ and the sample deviation will approach σ. This explains why μ and σ are called the mean and standard deviation of the normal distribution.

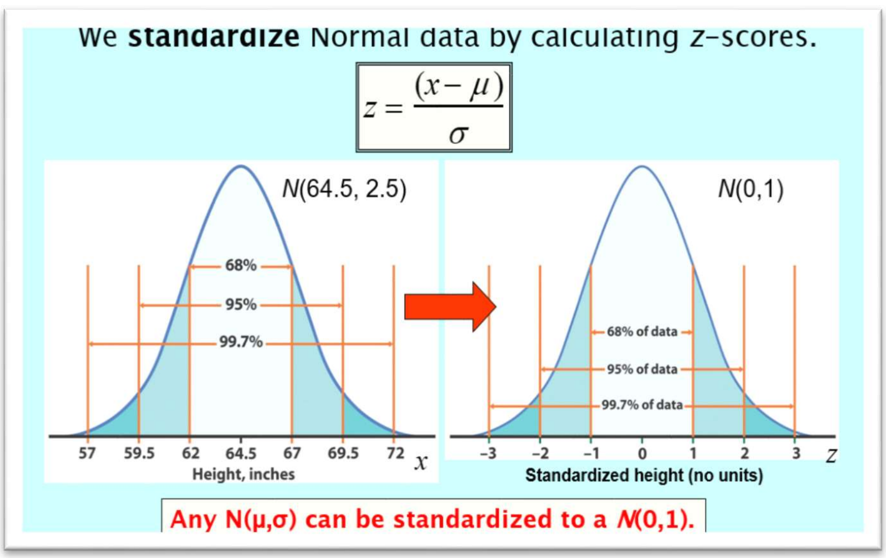
We say that a variable X is normally distributed – or Gaussian – with mean μ and standard deviation σ if equation (2) is true for any $a \leq b$ where the function f is defined by equation (1). Then, we write $X \sim N(\mu, \sigma)$.

Furthermore, when the mean μ is *zero* (0) and standard deviation σ is 1, its normal distribution $N(0,1)$ is called the **standard normal distribution** and has density:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} e^{\left(-\frac{z^2}{2}\right)} \qquad (5)$$

where, $z = \frac{(x-\mu)}{\sigma}$ with $\mu = 0$, $\sigma = 1$.

Equation (5) implies that for whatever values of μ *and* σ that we have for a normal distribution, we can easily transform this distribution to a *standard normal distribution*. We can therefore make all probabilities in the $N(\mu, \sigma)$ distribution to be computed from $\Phi(z)$, as shown in a slide example in Figure 4.



Probabilities concerning an $N(0,1)$ distributed variable are easily expressed in terms of the continuous distribution function:

$$P(a < Z < b) = \int_a^b \phi(z)dz = \int_{-\infty}^b \phi(z)dz - \int_{-\infty}^a \phi(z)dz = \Phi(b) - \Phi(a) \qquad (6)$$

We can look up the values of Φ in statistical tables or computed by a computer program such as MS Excel® or a scientific calculator.

The following Table 1 shows selected values of Φ, the cumulative function for the standard normal distribution. We can use the Excel function "=NORM.DIST(z,0,1,TRUE)" to calculate these Φ values.

Table 1 Selected Φ values calculated for various z inputs

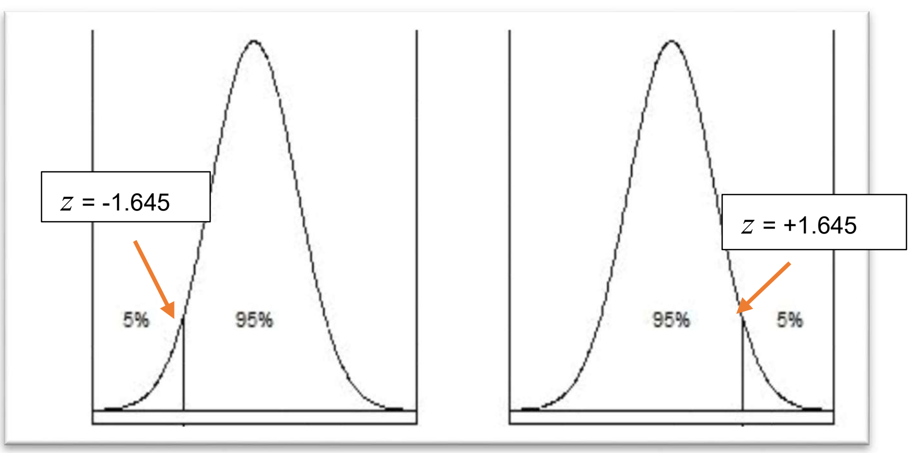
z	$\Phi(z)$	z	$\Phi(z)$
-3.000	0.001	1.282	0.900
-2.000	0.023	1.645	0.950
-1.000	0.159	1.960	0.975
0.000	0.500	2.326	0.990
1.000	0.841	2.576	0.995
2.000	0.977	3.090	0.999
3.000	0.999	3.291	1.000

How to interpret and apply these Φ values?

The interpretation of the density is like this: *for an interval* (a, b)*, where* $a < b$*, the area under the curve from* a *to* b *is the probability that a random observation falls within the interval.* The total area under the density is 1.

The figure z = 1.645 in Table 1 shows the area under the left side of the curve expressed as $\Phi(1.645)$ is 95% of the total area, or 0.950 as a coefficient of the total area. See Figure 5 below:

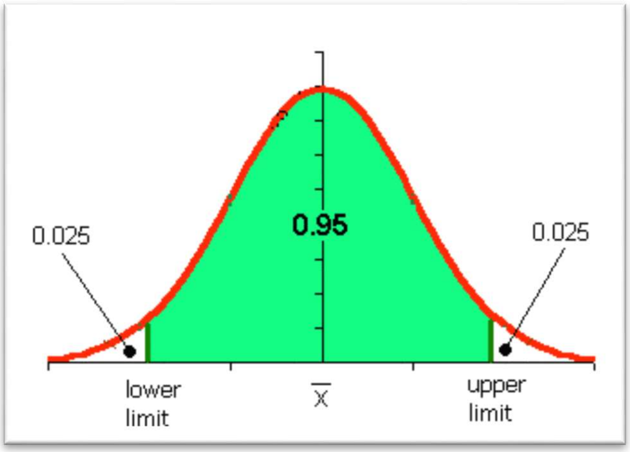
Figure 5 Areas under the normal curve for $z = \pm 1.645$



When the area under the curve is 0.95 bounded by both lower $-z$ and upper $+z$ limits, expressed as $P(-z \leq X \leq +z)$, the Φ–value is to be evenly divided, with value of 0.025 instead of 0.05, as shown in Figure 6 below.

In this case, we have $\Phi(z)$ = −1.96 and +1.96, respectively. This also explains the use of the coverage factor k = 2 or 1.96 for 95% confidence in the evaluation exercise of expanded measurement uncertainty.

Figure 6 The normal distribution graph showing $\Phi(z) = \pm 1.645$



Further comments

a. Figures 5 and 6 show the graphs of Φ. Notice that Φ(z) is always in the interval of N(0,1) and that Φ is increasing. In Figure 5, we have selected Φ(1.645) for illustrations:

$$\Phi(-\ 1.645) = P(Z \leq -\ 1.645) = 0.05$$

$$\Phi(1.645) = P(Z \leq 1.645) = 0.95$$

Since the points −1.645 and 1.645 by themselves have zero probability (see equation (3)), it follows that

$$P\ (Z > 1{,}645) = 0.050$$

$$P\ (-1.645 < Z < 1.645) = 1 - 0.050 - 0.050 = 0.900$$

b. In most of our routine laboratory analysis, we often perform duplicates or the most, triplicates. We seldom carry out larger number of replicated testing than 3 unless we are involved in new method development and method validation exercises. However, it is safe to assume normal distribution characteristics of our test data.

c. There are many test statistics to check if your test data are normally distributed. Some of such tests are listed below:

- Shapiro-Wilk normality test
- Anderson-Darling test
- Kolmogorov-Smirnov normality test