CHAPTER 1

Introduction and Revision of Some Statistical Ideas

Now here, you see it takes all the running you can do, to keep in the same place.

Through the Looking Glass, LEWIS CARROLL

This chapter provides a quick revision of some basic ideas of statistics necessary for process control. These include dot diagrams, probability distributions, mean, variance, standard deviation, properties of averages, the central limit theorem, and the normal, Poisson, and binomial distributions.

1.1 NECESSITY FOR PROCESS CONTROL

If you have a house, you know that you must work hard to keep it habitable—the tiles on the roof, the paint on the walls, the washing machine, the refrigerator, the television, all need attention from time to time. A car, a friendship, and our own bodies must similarly be continually nurtured or they will not remain in shape very long. The same is true for industrial processes. If left to themselves, machines do not stay adjusted, components wear out, and managers and operators forget, miscommunicate, and change jobs. Thus a stable stationary state is an unnatural one and its approximate achievement requires a hard and continuous fight. Both process *monitoring* and process *adjustment* can help achieve this. Both are likely to be needed.

1.2 SPC AND EPC

Some 80 years ago Walter Shewhart introduced statistical process control (SPC) using charts. These quality control charts have been widely applied, especially in the parts industries, such as automobile manufacture, and have resulted in dramatic savings and important improvements in quality of products. Control charts are devices used to *monitor* quality characteristics and so keep them as close as possible

Statistical Control by Monitoring and Adjustment, Second Edition. By Box, Luceño, and Paniagua-Quiñones

to their target values for indefinite periods of time. Shewhart control charts used in SPC combine two separate ideas:

(a) The concept of charting and studying serial data to point to abnormal operation of the process
(b) The idea that control limit lines about the target can tell us when deviations are large enough to warrant seeking "assignable causes" of trouble that might then be eliminated

These ideas are concerned with monitoring ("debugging") the process. They originated in the *parts* industries where the objective was to reproduce individual items as accurately as possible. If you are making a particular part for an automobile, you would like the dimensions for each item to be nearly constant. It would be nice if this could be achieved by carefully adjusting the machine, once and for all, and letting it run. Unfortunately, this would rarely, if ever, result in the production of a uniform product. Extraordinary precautions are needed in practice to ensure that quality characteristics do not change or drift away from their target values.

By contrast, the process industries are often concerned with *adjusting* the process. They use techniques of engineering process control (EPC) to maintain close to target such responses as percentage conversion of chemicals and measures of purity that if uncontrolled might very easily drift away from the target values. Control is affected by automatically manipulating some compensating variable(s) using a system of feedback and/or feedforward control. Frequently, the only noncapital cost of such automatic process adjustment is the cost of being off target. This is different from the typical situation in the parts industry where other costs such as those induced by stopping a machine or the replacement of a tool often dominate.

Thus, if you had asked a statistical quality control practitioner and a control engineer what they meant when they spoke of process control, you would likely have received very different answers. On the one hand, the quality control practitioner would have talked about the uses of control charts for process monitoring. On the other hand, the control engineer would talk about such things as feedback control and automatic controllers for process adjustment. In the past, the differences between these two approaches have sometimes led to "turf wars" and acrimony between the groups practicing SPC and EPC. This was unfortunate because both ideas are important with long and distinguished records of practical achievement. They have a synergistic relationship and serious inefficiency can occur when these tools are not used together and appropriately coordinated. Automatic feedback control has sometimes been applied without removing major process "bugs" that could have been detected and eliminated with SPC methods. On the other hand, SPC control charts, intended to monitor the process, have sometimes been used for adjustment of the process (i.e., to perform feedback control); when used for this purpose, these charts can be very inefficient.

The sharply drawn lines dividing the parts industries and the process industries have begun to disappear. One reason is that some processes are hybrids, having certain aspects of the parts industry and others of the chemical industry. Another reason is that conglomerate companies, in which both kinds of manufacture occur, are now much more common. A third reason is that, because of the "quality revolution," a greater awareness of the importance of control has led each industry to adopt some control technologies of the other.

We begin by presenting a summary of some ideas[1] in statistics that are used in Chapter 2 for underpinning standard quality control methods for SPC. Experience with these methods has shown that in some applications they did not work very well. They are here regarded as providing a first approximation. To overcome these difficulties, a second approximation is developed in this book that shows the need for *combining* some ideas of process adjustment and process monitoring.

1.3 PROCESS MONITORING WITHOUT A MODEL

Anyone who has anything to do with a process or system knows about Murphy's law: "Anything that can go wrong will go wrong." So quality control and improvement are not easy. How can we learn to nullify Murphy's tricks and to deempower him? Figure 1.1 shows a graph where a measured process characteristic is plotted against time in what is called a *run chart*. Even when the process is in a state of ideal operation, we can expect variation about the target value T. Such variation is often referred to as *noise* and is said to be due to *common causes* (see Deming, 1986).

When at some point there is a large deviation from target, which sticks out from the noise, as in Figure 1.1, you naturally ask the question, "What happened there?"

Technically you would say that this deviation is probably a *special cause*: a change not due to noise that may help to discover its cause. Simple questions like "When did it occur?" and "Was anything special going on at that time?" can help to do this. If you can determine the reason for the special cause, it becomes an *assignable* cause, and you may be able to arrange that this defect can be removed and can never happen again. If you wanted to be more certain that the deviation was truly exceptional, you could look through past records to determine how often a deviation as large as this one had occurred before. Inspection of these records might show periods in which the process had been slightly high and periods where it had been slightly low. In this case you might choose to use the deviation from a local average to decide any action you might take.

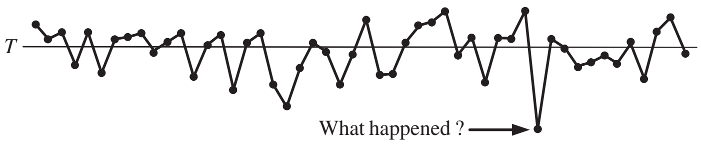

FIGURE 1.1 Values of a measured characteristic in state of ideal operation with suspicious observation.

[1] These ideas were developed, in particular, in Box et al. (1976).

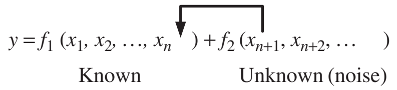
$y = f_1\,(x_1, x_2, \ldots, x_n\ \ ) + f_2\,(x_{n+1}, x_{n+2}, \ldots\ \ )$
Known
Unknown (noise)

FIGURE 1.2 Ideal operation of quality control chart to move previously unknown source of trouble x_{n+1} (special cause) from unknown to known.

1.4 DETECTING A SIGNAL IN NOISE

The intended mode of operation of a standard quality control chart is illustrated in Figure 1.2. The deviation y from the target is here represented as a function of a number of *known* operating variables $x_1, x_2, \ldots, x_n$ and a number of unknown factors $x_{n+1}, x_{n+2}, \ldots$ that together constitute the noise. A sufficiently extreme deviation can point to the existence of an assignable cause x_{n+1} that, as in Figure 1.2, may be moved from the unknown to the known. This kind of operation is called *process monitoring* or more colloquially continuous "*debugging*" of the process. This can have three desirable results:

(a) Sources of trouble may be identified and permanently eliminated.

(b) It may be possible to fix the newly found factor at its best level and so improve the process.

(c) The level of the residual noise will be reduced somewhat, making it easier to discover other assignable causes.

So a general definition of a quality control chart for monitoring would be that it is an efficient way of looking for *signals* in *noise*. But sometimes it is hard to distinguish between what might be noise and what might be a signal and so two questions come up: How big does a deviation need to be before we consider it too large to be easily explained as "noise" and a deviation from what?

Since process monitoring is or should be an efficient way of finding a signal in noise, the way to proceed depends on the nature of the signal and the nature of the noise. Assume for the moment that the signal consists of a single extreme value like that in Figure 1.1, which we will call a "spike." In this chapter we discuss three important distributions that might, in different conditions, represent the noise. These distributions are called the normal distribution, the binomial distribution, and the Poisson distribution. These have been used, respectively, for measurement data, data about proportions, and data about frequencies.

1.5 MEASUREMENT DATA

Suppose that we had data for the diameter, in micrometers, of holes from a drilling operation and the first 10 values were 94, 90, 94, 89, 100, 96, 97, 96, 92, and 98. These are plotted in Figure 1.3.

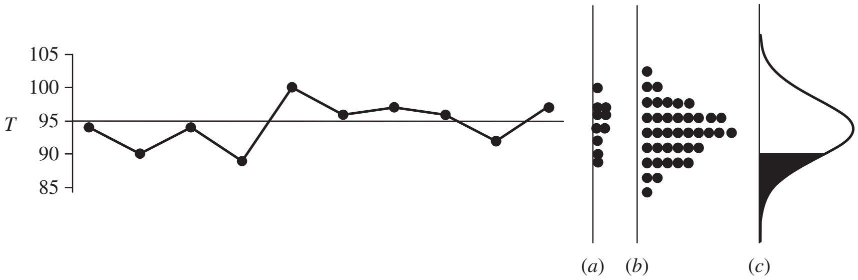

FIGURE 1.3 Diameter of holes, in micrometers, from drilling operation.

The data vary on both sides of the target value T, which is equal to 95 and is indicated by the central straight line. One way in which these 10 values can be appreciated in relation to each other is by means of a dot diagram, which consists of a horizontal scale with dots representing observations as in Figure 1.3*a*. But now imagine a dot diagram for, say, 200 observations obtained in this way with the dots piled one on top of the other. Then you would most likely obtain a diagram that looked like Figure 1.3*b*. Furthermore, you can imagine that if the number of observations were extremely large (theoretically infinite) then this diagram might be represented by a smooth curve like that in Figure 1.3*c*. This would be called the "population" *distribution* of the hole diameters and its mean would be denoted by the Greek letter μ (mu).

A quantity that has a distribution is called a *random variable*. Now suppose you *knew* this distribution and you wanted to know what proportion of the diameters were smaller than 90 μm (the shaded part of the distribution in Figure 1.3*c*). Then the answer to the problem would be obtained by finding out how large the shaded area was in relation to the total area. Or, alternatively, if you had scaled the distribution so that the area under the curve was 1, then the shaded area would give the *probability* of diameters of the holes being less than 90.

In this example say that the mean μ of the distribution is 96. In practice, you would not know this. However, you would know the average, 94.6, of the sample of 10 values that we will indicate by $\overline{y}$ (y bar). If you assumed that you had a *random* sample of 10 taken from the whole population, then $\overline{y}$ would provide an *estimate* of μ.

Exercise 1.1 How could you use the diagram in Figure 1.3*c* to find out the chance of a hole being between 90 and 100 μm in diameter?

1.6 TWO IMPORTANT CHARACTERISTICS OF A PROBABILITY DISTRIBUTION

The *mean* μ is a measure of location, which determines where the distribution is centered. The *standard deviation* σ (Greek sigma) is a measure of dispersion, which determines how widely the distribution is spread. Another important measure of

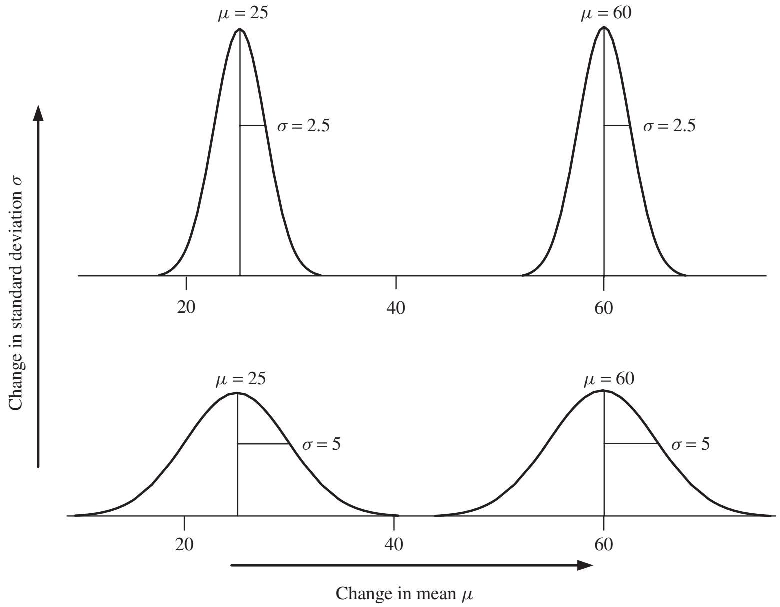

FIGURE 1.4 Distributions with means $\mu = 25$, 60 and standard deviations $\sigma = 2.5$, 5.

spread is the *variance* σ^2, the square of the standard deviation. For illustration Figure 1.4 shows distributions with different locations and spreads with $\mu = 25$, 60 and with $\sigma = 2.5$, 5.

1.7 NORMAL DISTRIBUTION

A distribution that has often been used to represent the population distribution is the *normal* distribution, shown in Figure 1.5. This is a symmetric distribution defined by two constants or "parameters": its mean μ and its variance σ^2 or equivalently its standard deviation σ. For the normal distribution σ, as is illustrated in Figure 1.5, is the distance from the *point of inflection* of the curve to the mean μ. The point of inflection, in the figure, is the point where the gradient (slope) of the curve stops increasing and starts decreasing.

Exercise 1.2 What are the variances of the distributions in Figure 1.4?

1.8 NORMAL DISTRIBUTION DEFINED BY μ AND σ

An important characteristic of distributions[2] is the so-called *central limit effect*. It turns out that, regardless of the shape of the distribution of the original

[2] Except for distributions that are mathematical curiosities.

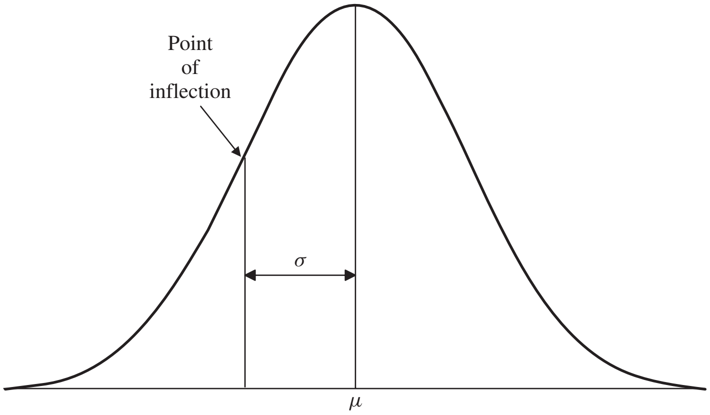

FIGURE 1.5 Mean (measure of location), standard deviation (measure of spread), and point of inflection of normal curve.

independent[3] measurements, the shape of the distribution of *averages* looks more and more like that of a normal distribution as the sample size increases. This central limit effect is illustrated in Appendix 1A. Because of this effect, it would often be safe to assume, even for samples of size $n = 4$, that the distribution of sample averages was very approximately normal.

A second important fact is that, whatever the distribution of the data having variance σ^2, the variance of *averages* of *n independent* observations is σ^2/n.

1.9 PROBABILITIES ASSOCIATED WITH NORMAL DISTRIBUTION

Look at the normal distribution in Figure 1.6. The unshaded area, within the range of $\mu \pm \sigma$, represents a proportion 0.683 of the total area under the curve. Thus the chance of a quantity that is normally distributed lying within plus or minus one standard deviation of its mean is 68.3%, or about two-thirds. Equivalently, the probability of such a quantity falling outside these limits is 31.7%, or about one-third.

The unshaded and lightly shaded areas within the range $\mu \pm 2\sigma$ together represent a proportion 0.954 of the total area under the curve. So the chance for a quantity that is normally distributed lying within two standard deviations of the mean is 95.4%, or about $\frac{19}{20}$. Equivalently, the probability of such a quantity falling outside these limits is 4.6%, or about $\frac{1}{20}$.

Finally, the proportion of the distribution within the range $\mu \pm 3\sigma$ represents 0.9973 of the total area. The chance of lying within three standard deviations of the mean is 99.73%, or about $\frac{399}{400}$. Equivalently, the probability of falling outside these

[3] That is, observations where each data value does not affect the probability distribution of any of the others.

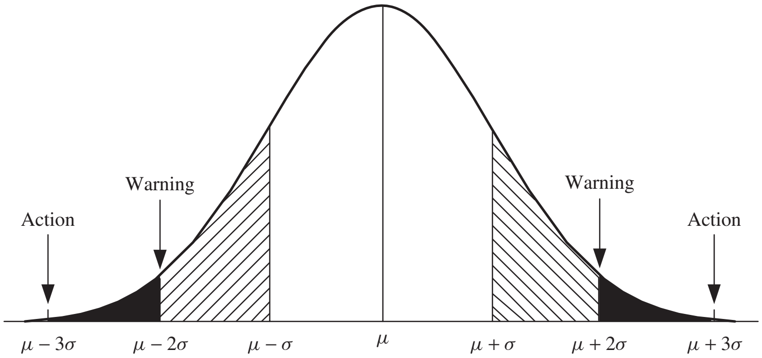

FIGURE 1.6 Normal distribution showing limits for $\mu \pm \sigma$, $\mu \pm 2\sigma$, and $\mu \pm 3\sigma$.

limits is 0.27%, or about $\frac{1}{400}$. But notice that in order to make use of these facts we would need to know the mean μ and the standard deviation σ or in practice how to estimate them. In a quality control chart the limits at $\pm 2\sigma$ are sometimes called *warning* limits and the limits at $\pm 3\sigma$ are called *action* limits.

1.10 ESTIMATING MEAN AND STANDARD DEVIATION FROM DATA

The data from the drilling operation may be used to exemplify the calculation of the sample mean $\overline{y}$ and sample standard deviation s:

Hole number	1	2	3	4	5	6	7	8	9	10
Hole diameter, μm	94	90	94	89	100	96	97	96	92	98

The average $\overline{y}$ of the sample data supplies an estimate of the mean μ of the hypothetical large distribution of values from which it is imagined to be drawn:

$$\hat{\mu} = \overline{y} = \frac{1}{10}(94 + 90 + 94 + 89 + 100 + 96 + 97 + 96 + 92 + 98) = 94.6$$

where, here and throughout this book, the "hat" (applied in this case to μ) means "an estimate of."

The standard deviation σ measures the *root-mean-square deviation* of the data from its mean and is an important measure of spread whether or not the data are normally distributed. Thus,

$$\sigma = \sqrt{\frac{1}{n}\sum (y - \mu)^2}$$

where here the symbol Σ (capital Greek sigma) indicates a sum taken over the whole population.

The sum of squared deviations from the *sample mean*, 94.6, in this example is

$$(-0.6)^2 + (-4.6)^2 + (-0.6)^2 + (-5.6)^2 + (5.4)^2 + \cdots + (-3.4)^2 = 115.2$$

By substituting that sample mean for the true mean in this expression we have cheated a bit, but it can be shown that this can be exactly allowed for by dividing the sum of squares not by $n = 10$ but by $n - 1 = 9$. Thus the *sample standard deviation*, the estimate $\hat{\sigma} = s$ of σ obtained entirely from the sample data itself, is

$$\hat{\sigma} = s = \sqrt{\frac{115.2}{9}} = 3.58$$

and the corresponding *sample variance* is

$$\hat{\sigma}^2 = 12.8$$

If, in general, we denote a sample of n data values by $y_1, y_2, y_3, \ldots, y_n$, then the sample mean is given as

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

where the index i varies from 1 to n. So, the operation to obtain the sample standard deviation can be written as

$$\hat{\sigma} = s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n-1}}$$

An equivalent formula for the sum of squares $\sum(y - \overline{y})^2$ is $\sum y^2 - n\overline{y}^2$.

Exercise 1.3a Obtain the sample mean and sample standard deviation for the following data: 6, 4, 3, 8, 4.

1.11 COMBINING ESTIMATES OF σ^2

Suppose you have three separate estimates of the same variance σ^2 based on n_1, n_2, and n_3 observations. Then the combined estimate $\hat{\sigma}_G^2$ (say) would be

$$\hat{\sigma}_G^2 = \frac{(n_1 - 1)\,\hat{\sigma}_1^2 + (n_2 - 1)\,\hat{\sigma}_2^2 + (n_3 - 1)\,\hat{\sigma}_3^2}{(n_1 - 1) + (n_2 - 1) + (n_3 - 1)}$$

or alternatively

$$\hat{\sigma}_G^2 = \frac{\sum_{i=1}^{n_1}(y_i - \overline{y}_1)^2 + \sum_{j=1}^{n_2}(y_j - \overline{y}_2)^2 + \sum_{k=1}^{n_3}(y_k - \overline{y}_3)^2}{n_1 + n_2 + n_3 - 3}$$

Using these ideas, Shewhart produced control charts for data that appear as measurements. We discuss such charts in Section 2.1.

Exercise 1.3b List any assumptions which have been made in Sections 1.5–1.11.

1.12 DATA ON FREQUENCIES (EVENTS): POISSON DISTRIBUTION

Sometimes you need to consider the *frequency* of events, for example, the number of injuries per week occurring at a particular workplace. If some rather unlikely assumptions were true, the frequencies of no event, one event, two events, and so on, occurring in a week would have a distribution called the Poisson distribution. One important assumption would be that the probability of an injury in any small interval of time (say 1 min) was constant. We later discuss the difficulties that occur when this and other assumptions are not true.

1.12.1 Calculations for Poisson Distribution

On the *Poisson assumption*, the chance of 0, 1, 2, 3, ..., y, ... injuries occurring in any given week is

$$\Pr(y) = e^{-\mu}\mu^{y}/y! \tag{1.1}$$

where μ is the mean number of injuries per week and $y!$ is the factorial $y \times (y-1) \times (y-2) \times \cdots \times 2 \times 1$.

For example, if the mean number of injuries was 2.1 a week, then knowing only this mean you could find the probability (proportion of weeks) in which there should be no injuries, one injury, two injuries, and so on. Thus:

Number of injuries, y	0	1	2	3	4	5	6
Probability of y injuries in any given week	0.12	0.26	0.27	0.19	0.10	0.04	0.01

These form the probability distribution plotted in Figure 1.7*a*. Thus, on the Poisson assumption, in a hundred weeks you would expect that there would be about 12 weeks (100 × 0.12) with no injuries, 26 weeks (100 × 0.26) with one injury, 27 weeks (100 × 0.27) with two injuries, and so on. As a second example, if the mean number μ of injuries were 10 per week, the appropriate Poisson distribution would be that shown in Figure 1.7*b*.

Exercise 1.4 If the mean number μ of injuries per week were 10:

(a) Calculate the probabilities for 0, 1, 2, ..., 20 injuries in any given week.

(b) Do your results agree with Figure 1.7*b*?

(c) Calculate the probability that there would be less than five accidents.

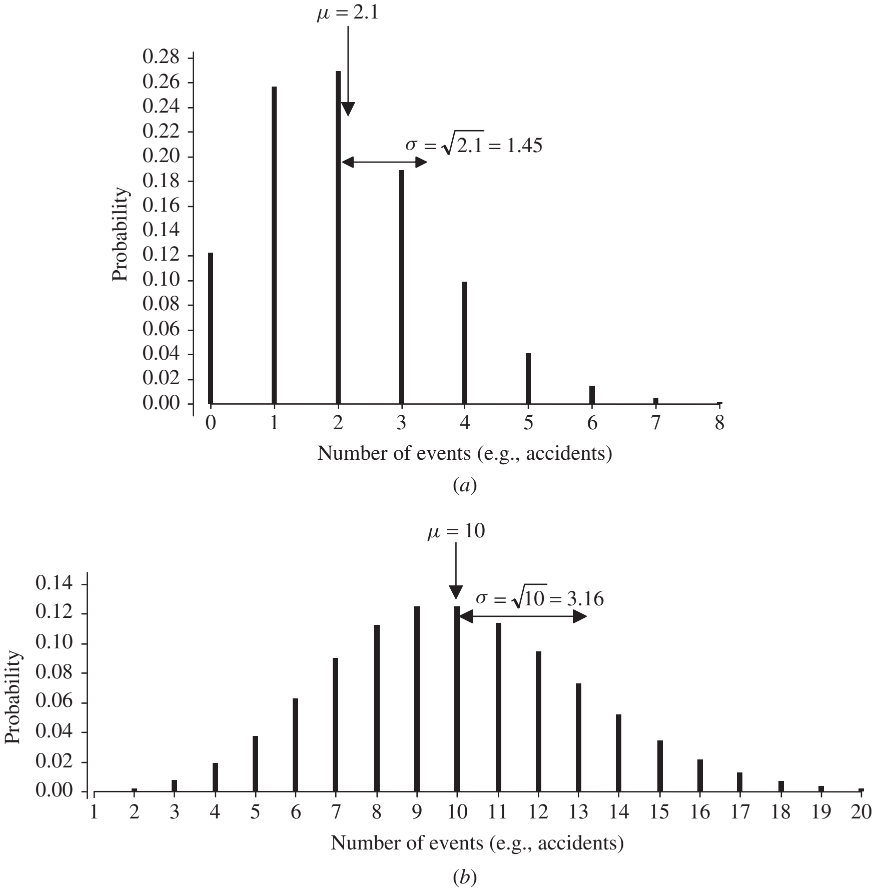

FIGURE 1.7 Poisson probability distribution: (*a*) $\mu = 2.1$; (*b*) $\mu = 10$.

1.12.2 Mean of Poisson Distribution

The mean μ (the point of balance) of this distribution, which takes values $y = 0, 1, 2, 3, \ldots$ with probability $\Pr(y)$, is given by summing the product $\Pr(y) \cdot y$, that is,

$$\mu = \sum \Pr(y) \cdot y \tag{1.2}$$

Look again at the Poisson distribution in Figure 1.7*a* generated from Equation (1.1). The mean for this distribution is, according to Equation (1.2),

$$\mu = (0.12 \times 0) + (0.26 \times 1) + (0.27 \times 2) + (0.19 \times 3) + (0.10 \times 4)$$
$$+ (0.04 \times 5) + \cdots = 2.1$$

As expected the calculation says that the mean is equal to $\mu = 2.1$, the value used to generate it. In general we call the mean a measure of location since it tells us, for example, that the y's are distributed about 2.1 and not somewhere else.

1.12.3 Variance of Poisson Distribution

The variance of a Poisson distribution with mean μ is the mean value of the squares of the deviations $y - \mu$. Thus

$$\sigma^2 = \sum \Pr(y)(y - \mu)^2 \tag{1.3}$$

Accordingly, the variance in this Poisson example can be calculated as

$$\sigma^2 = 0.12 \times (0 - 2.1)^2 + 0.26 \times (1 - 2.1)^2 + 0.27 \times (2 - 2.1)^2 + \cdots = 2.1$$

This illustrates the remarkable and seductive fact that, *on the assumption that the number of weekly events follows the Poisson distribution*, the variance σ^2 would be exactly equal to the mean μ. Thus, if the assumptions were true, you would need to know just one number, namely, the mean μ of a Poisson distribution, and you would know everything about the distribution.

Exercise 1.5 Use Equation (1.3) to show that the standard deviation of the distribution in Figure 1.7*b* is $\sqrt{10}$.

1.13 NORMAL APPROXIMATION TO POISSON DISTRIBUTION

Figure 1.8*a* shows a Poisson distribution with mean 20.6 and hence, on the Poisson assumption, with standard deviation $\sqrt{20.6} = 4.54$. Figure 1.8*b* shows a normal distribution with the same mean and standard deviation as the Poisson. As you can see by looking at Figures 1.8*a* and 1.8*b*, the shape of the Poisson distribution approximates the normal distribution with the same mean, and the standard deviation the square root of the mean. It turns out that if $\mu = 12$ or more the Poisson distribution is well approximated by the normal distribution. So if $\mu > 12$ and the Poisson assumptions were true you could use the normal distribution with mean μ and standard deviation $\sqrt{\mu}$ to approximate the Poisson distribution.

1.14 DATA ON PROPORTION DEFECTIVE: BINOMIAL DISTRIBUTION

Sometimes it is the *proportion* of manufactured articles that contain some defect that is of interest. For example, condoms are sometimes tested by inflating a number of randomly chosen test specimens to a very high fixed pressure and noting the proportion p that burst.

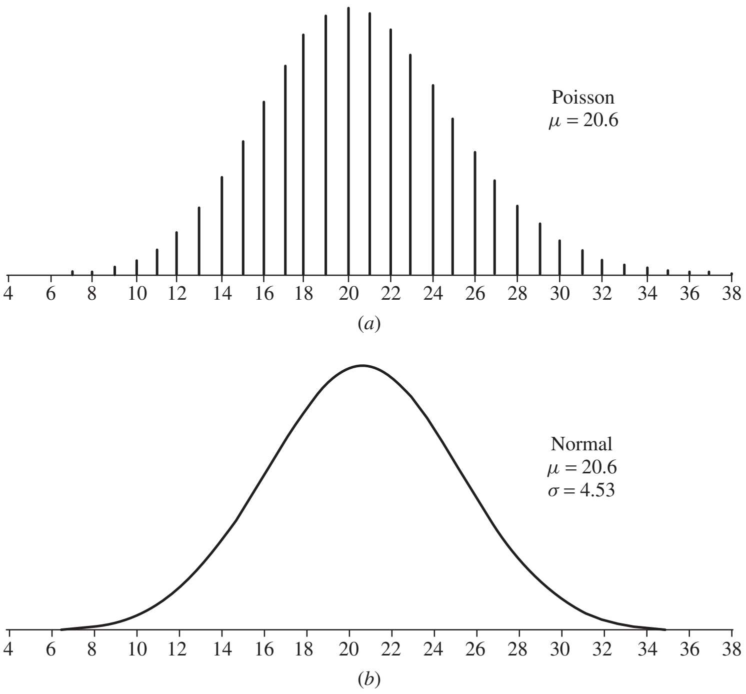

FIGURE 1.8 (*a*) Poisson distribution with mean 20.6 representing probabilities of different frequencies of accident. (*b*) Approximating normal distribution with mean 20.6 and standard deviation $\sqrt{20.6} = 4.54$.

1.14.1 Calculations for Binomial Distribution

If we test n items, then, again, *on some rather unlikely assumptions and, in particular, the assumption that p remains fixed*, the binomial distribution tells you the proportion of times that you should get 0, 1, 2, ... , y, ... , n defectives.

For a probability p and a sample of size n the chance of exactly y failures is

$$\Pr(y) = \frac{n!}{y!(n-y)!} p^y q^{n-y} \quad \text{where } q = 1 - p \tag{1.4}$$

For example, suppose you had a biased penny with the property that it came down heads with probability $p = 0.8$ (and hence tails with a probability $q = 0.2$) and suppose you made $n = 5$ throws. Then knowing only p and using Equation (1.4) you could find the chance of getting 0 heads, 1 head, 2 heads, ..., 5 heads as follows:

Number of heads, y	0	1	2	3	4	5
Pr(y)	0.0003	0.0064	0.0512	0.2048	0.4096	0.3276

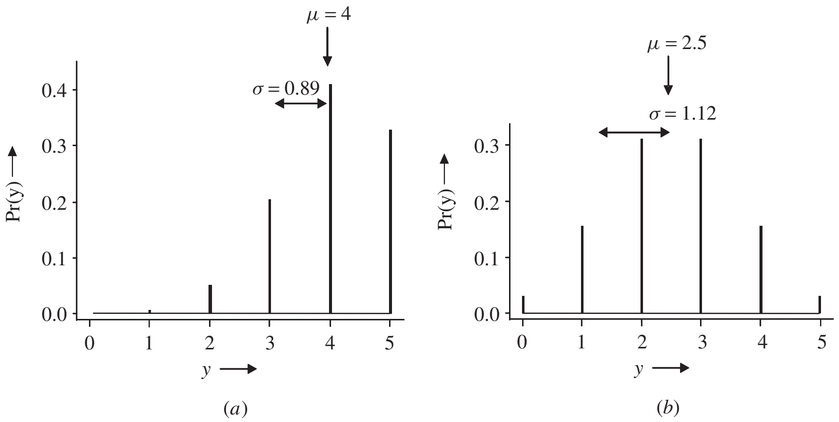

FIGURE 1.9 Binomial distribution for different choices of p: (a) $p = 0.8$, $n = 5$; (b) $p = 0.5$, $n = 5$ (distribution for fair penny).

where, for example,

$$\Pr(y = 3) = \frac{5!}{3!2!}0.8^3 \times 0.2^2 = 0.2048$$

The binomial probability distribution for this biased penny is shown in Figure 1.9*a*. For a fair penny, with $p = 0.5$, the corresponding probability distribution is shown in Figure 1.9*b*.

Exercise 1.6 Calculate $\Pr(y)$ where y is the number of heads for $n = 5$ throws of the fair coin for $y = 0, 1, 2, 3, 4, 5$. Do your calculations produce the distribution in Figure 1.9*b*?

1.14.2 Mean and Variance of Binomial Distribution

For the binomial distribution the number of defectives from a sample of n has mean np and variance npq. Thus the proportion y/n of the defectives has mean p and variance pq/n.

Exercise 1.7 Carry through the calculations to obtain the means and variances for the binomial distributions in Figures 1.9*a* and 1.9*b*. What do you find? Do the values agree with the expressions $\mu = np$ and $\sigma^2 = npq$?

1.15 NORMAL APPROXIMATION TO BINOMIAL DISTRIBUTION

As with the Poisson distribution, the binomial distribution may be approximated by a normal distribution if n is moderately large and *p is not small*. A general rule is that you should not use a normal approximation if $n < 12(1 - p)/p$.

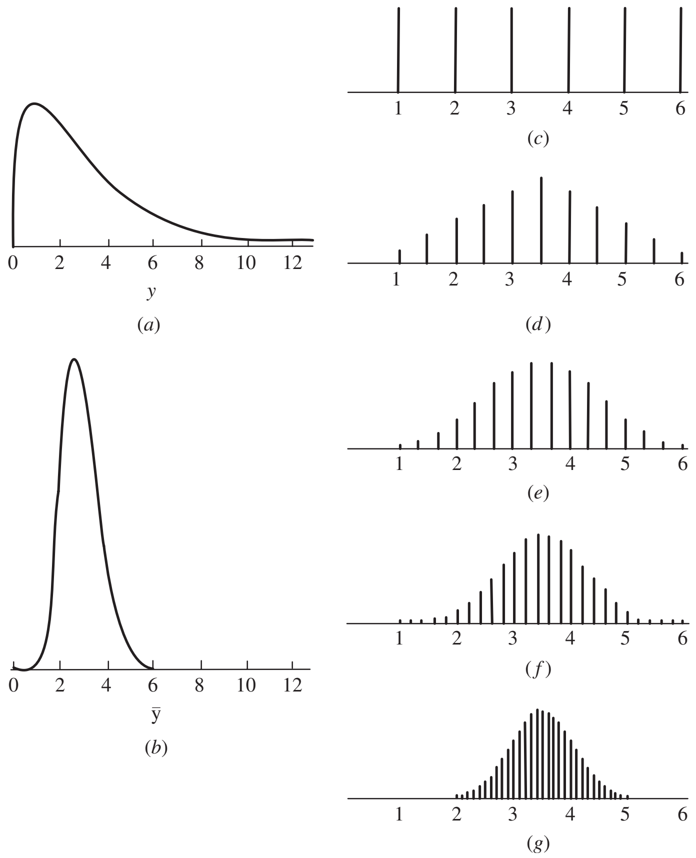

FIGURE 1A.1 (*a*) Skewed distribution for y. (*b*) Distribution of average $\overline{y}$ of 10 observations randomly sampled from skewed distribution. (*c*) Distribution of number of pips from single throw of fair dice. Distribution of average of (*d*) 2, (*e*) 3, (*f*) 5, and (*g*) 10 throws.

APPENDIX 1A: CENTRAL LIMIT EFFECT

From Figure 1A.1 you see that the distribution of an average is to be more nearly normal than the "parent" distribution from which the data are drawn. Why should random errors tend to have a distribution that is approximated by the normal rather than by some other distribution? An important rationale relies on the central limit effect.

If the overall error e is an aggregate of a number of component errors, $e = e_1 + e_2 + \cdots + e_m$, such as sampling errors, measurement errors, or manufacturing variation, in which no one component dominates, then almost irrespective of the distribution of the individual components, the distribution of the aggregated error e will tend to the normal as the number of components gets larger.

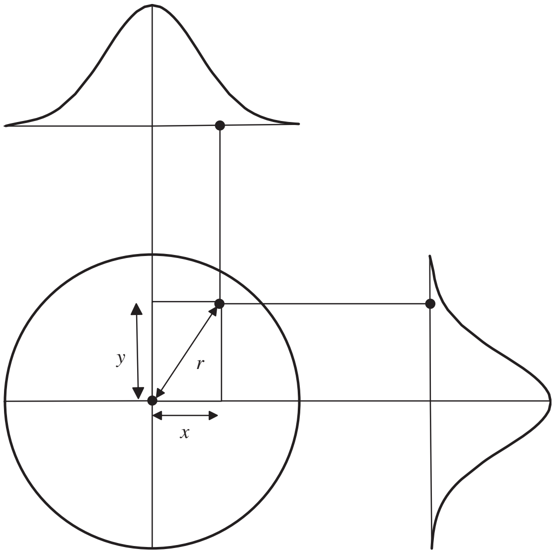

FIGURE 1A.2 Illustration of Clark Maxwell's derivation of normal distribution.

A Remarkable Argument

Clark Maxwell was concerned with proving that the velocities of individual molecular particles must be normally distributed. We will illustrate this argument but, for simplicity, will change the context a little and pose the problem in terms of a sharpshooter firing at a target. Suppose that neither the sharpshooter nor the rifle has biases of any kind. Then, looking at Figure 1A.2, the following two postulates seem reasonable:

1. The probability of a shot hitting the target at any distance r from the bull's-eye is the same for every point at the same distance r. That is, the probability is constant on every circumference of radius r centered at the bull's-eye.
2. The probability of the horizontal coordinate x having any particular value is completely independent of the probability of the vertical coordinate y having any particular value.

The remarkable fact is that, as soon as these two postulates are admitted, x and y *must* be individually normally distributed.

For those interested in mathematical matters, a very rough sketch of the proof is as follows: From postulate 1, the probability distribution of x and y must be of the form

$$p(x, y) = p(r^2) = p(x^2 + y^2)$$

But from postulate 2,

$$p(x, y) = p(x) \cdot p(y)$$

The only function for which $p\,(x) \cdot p(y) = p\,(x^2 + y^2)$ is the exponential function, so that $p(y)$ must be of the form ae^{by^2}, where a is a positive constant and b is a negative constant. Likewise for $p(x)$, so that

$$p(x, y) = p(x)p(y) = a^2 e^{b(x^2+y^2)}$$

But the probability becomes less as x and y get larger, so that b must be negative. If we call it $-c^2$, then $p(y) = ae^{-c^2y^2}$.

If the reader will graph this function for any values of a and c, a normal curve will be obtained of the shape we have already seen in Figures 1.4 and in 1.5.

The constants may be expressed in a more commonly used form as follows. The constant c is conveniently reexpressed in terms of the standard deviation σ, that is, the root-mean-square error of y. It can be shown that this requires that $c^2 = 1/2\sigma^2$. Also, since y *must* be somewhere, the constant a must be chosen to make the total area (probability) under the curve equal to 1. This requires that $a = 1/\sqrt{2\pi\sigma^2}$. Finally, the data can be centered about some mean μ other than zero by substituting $y - \mu$ for y. Putting all this together we obtain the general form of the normal distribution:

$$p(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2}$$

The normal distribution is important not only because it is a probability curve that frequently approximates the distribution of random errors but also because it provides, as we have seen, an approximation for other distributions such as the Poisson and the binomial.

Problems

1.1 What do you understand by "synergistic control"?

1.2 What is meant by white noise and statistically independent?

1.3 How could you tell that, to an adequate approximation, a given model explains the data?

1.4 Show how process monitoring can gradually eliminate lurking variables?

1.5 Suppose that the number of errors per page in a certain book can be modeled by a Poisson distribution with mean $\mu = 1.2$. Find the probability that a page chosen at random will contain:

(a) At least one error
(b) One error
(c) Two errors

1.6 The following data represent the number of calls entering into a phone server per minute for 25 periods of 1 min. Assume the data can be modeled by the Poisson distribution with mean 18 phone calls per minute. Plot the data with warning and action limits. Do you suspect the means have changed? Why?

Period	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
Calls	18	17	18	22	15	13	15	15	19	19	11	17	13	21	21	19	23	19	27	19	25	17	26	25	22

1.7 A microchip manufacturer has data following the Poisson distribution for the number of defects per microchip. If the mean number of defects per microchip is 0.2, how many microchips with (a) 0 defects, (b) 1 defect, and (c) two or more defects can you expect in 1000 microchips?

1.8 Assume that the proportion of defective items produced by a stable process is $p = 0.01$. A monitoring scheme samples eight items every 2 h and declares the process in control only if no defective items are found.

(a) Calculate the probability of declaring the process in control the next sampling period.

(b) What is the probability of observing exactly one defective item the next sampling period?

(c) What is the probability of finding two defective items the next sampling period?

1.9 The percentage of registered voters in favor of certain candidate is about 40%. Assuming that the binomial model approximates the distribution of the number of voters in a sample of 10 in favor of the candidate, what is the probability that (a) less than 3 were in favor of the candidate and (b) between 6 and 9 voters, inclusive, were in favor of the candidate?

1.10 Suppose the concentration of a certain substance is approximately normal with mean $\mu = 99.2$ and standard deviation $\sigma = 0.2$. What is the chance of observing a concentration (a) less than 99, (b) bigger than 99.7, and (c) between 99.2 and 99.9?

1.11 A random variable has a normal distribution with mean 60 and a standard deviation of 5. What is the probability of observing values (a) less than 50, (b) between 55 and 75, and (c) greater than 62.

1.12 (a) Use the normal distribution to approximate the probability that a Poisson variable with mean 35 exceeds 30.

(b) Use a calculator or a computer and calculate the mentioned probability using Equation (1.1).

1.13 (a) Use the normal distribution to approximate the probability that a Poisson variable with mean 80 is less than 90.

(b) Use a calculator or a computer and calculate the mentioned probability using Equation (1.1).

1.14 (a) Use the normal distribution to approximate the probability that a binomial variable with $p = 0.6$ and $n = 15$ is less than 13.

(b) Use a calculator or a computer and calculate the mentioned probability using Equation (1.4).

1.15 (a) Use the normal distribution to approximate the probability that a binomial variable with $p = 0.3$ and $n = 15$ is less than 22.

(b) Use a calculator or a computer and calculate the mentioned probability using Equation (1.4).