3 Statistical Models or Quality Control and Improvement

Chapter Outline

Learning Objectives

1. Construct and interpret visual data displays, including the stem-and-leaf plot, the histogram, and the box plot
2. Compute and interpret the sample mean, the sample variance, the sample standard deviation, and the sample range
3. Explain the concepts of a random variable and a probability distribution
4. Understand and interpret the mean, variance, and standard deviation of a probability distribution
5. Determine probabilities from probability distributions
6. Understand the assumptions for each of the discrete probability distributions presented
7. Understand the assumptions for each of the continuous probability distributions presented
8. Select an appropriate probability distribution for use in specific applications
9. Use probability plots
10. Use approximations for some hypergeometric and binomial distributions

3.1 Describing Variation

Stem-and-Leaf Display

TABLE 3.1

Cycle Time in Days to Pay Employee Health Insurance Claims

Claim	Days	Claim	Days	Claim	Days	Claim	Days
1	48	11	35	21	37	31	16
2	41	12	34	22	43	32	22
3	35	13	36	23	17	33	33
4	36	14	42	24	26	34	30
5	37	15	43	25	28	35	24
6	26	16	36	26	27	36	23
7	36	17	56	27	45	37	22
8	46	18	32	28	33	38	30
9	35	19	46	29	22	39	31
10	47	20	30	30	27	40	17

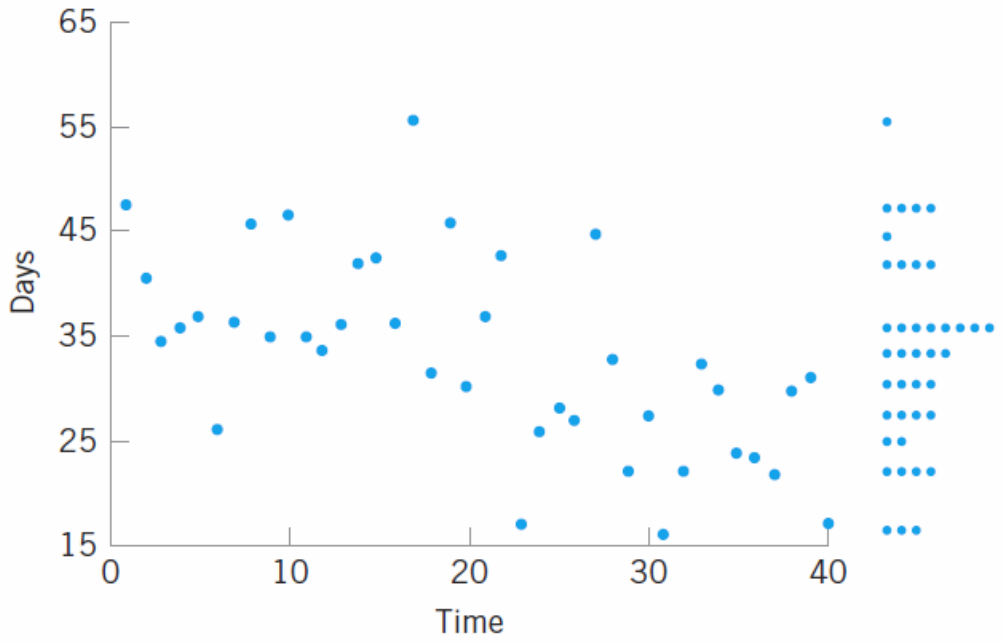
Stem-and-Leaf Display: Days

```
Stem-and-leaf of Days
N = 40
Leaf Unit = 1.0
 3    1   677
 8    2   22234
13    2   66778
(8)   3   00012334
19    3   555666677
10    4   1233
 6    4   56678
 1    5
 1    5   6
```

FIGURE 3.1 Stem-and-left plot for the health insurance claim data.

Easy to find percentiles of the data; see page 69

Plot of Data in Time Order



Marginal plot produced by MINITAB

FIGURE 3.2 A time series plot of the health insurance data in Table 3.1.

Also called a *run chart*

Histograms – Useful for large data sets

TABLE 3.2

Layer Thickness (Å) on Semiconductor Wafers

438	450	487	451	452	441	444	461	432	471
413	450	430	437	465	444	471	453	431	458
444	450	446	444	466	458	471	452	455	445
468	459	450	453	473	454	458	438	447	463
445	466	456	434	471	437	459	445	454	423
472	470	433	454	464	443	449	435	435	451
474	457	455	448	478	465	462	454	425	440
454	441	459	435	446	435	460	428	449	442
455	450	423	432	459	444	445	454	449	441
449	445	455	441	464	457	437	434	452	439

Group values of the variable into bins, then count the number of observations that fall into each bin

Plot frequency (or relative frequency) versus the values of the variable

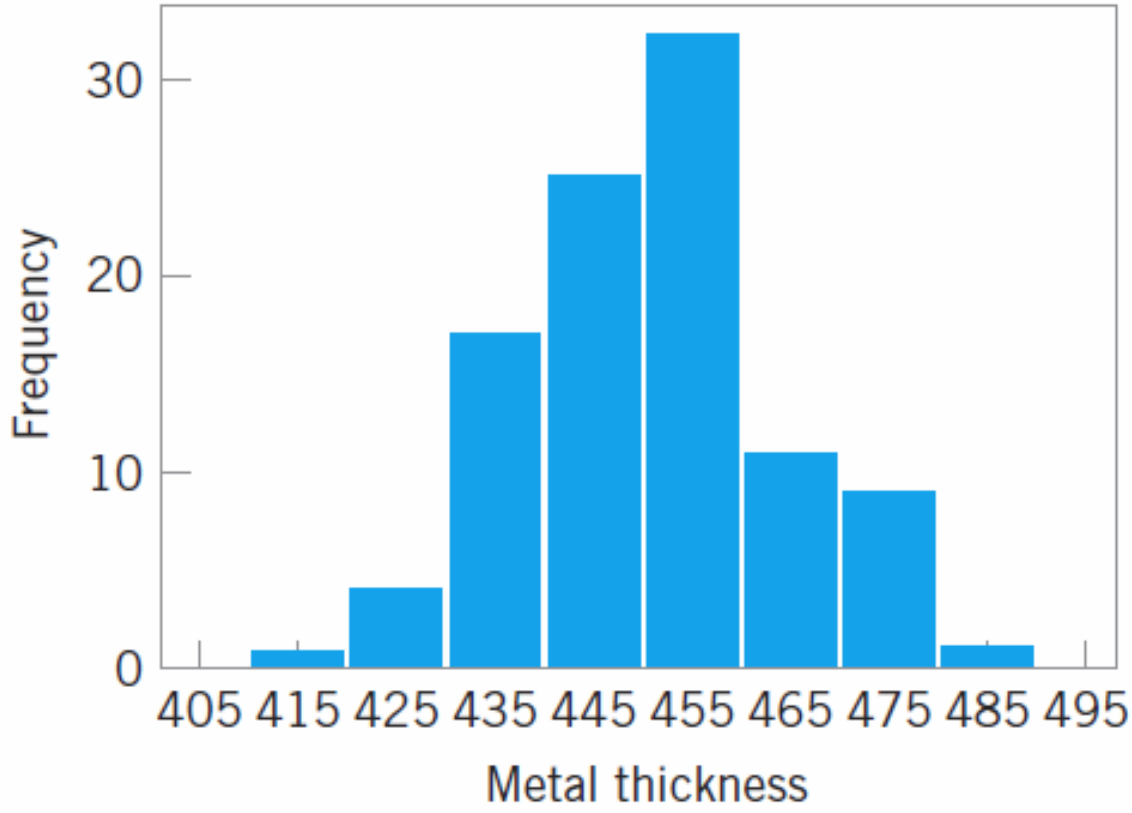


FIGURE 3.3 Minitab histogram for the metal layer thickness data in Table 3.2.

Additional Minitab Graphs

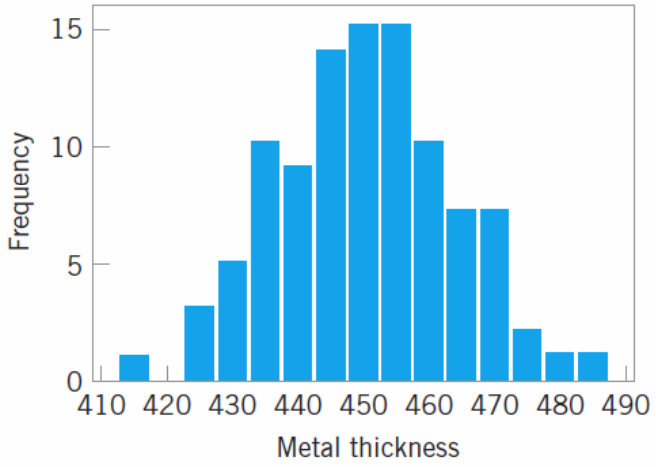


FIGURE 3.4 Minitab histogram with 15 bins for the metal layer thickness data.

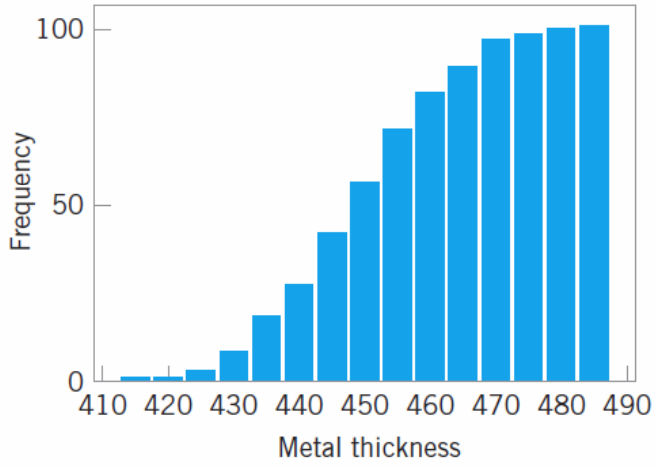


FIGURE 3.5 A cumulative frequency plot of the metal thickness data from Minitab.

TABLE 3.3

Surface Finish Defects in Painted Automobile Hoods

6	1	5	7	8	6	0	2	4	2
5	2	4	4	1	4	1	7	2	3
4	3	3	3	6	3	2	3	4	5
5	2	3	4	4	4	2	3	5	7
5	4	5	5	4	5	3	3	3	12

Figure 3.6 is the histogram of the defects. Notice that the number of defects is a discrete variable. From either the histogram or the tabulated data we can determine

$$\text{Proportions of hoods with at least 3 defects} = \frac{39}{50} = 0.78$$

and

Proportions of hoods with between 0 and

$$2 \text{ defects} = \frac{11}{50} = 0.22$$

These proportions are examples of relative frequencies.

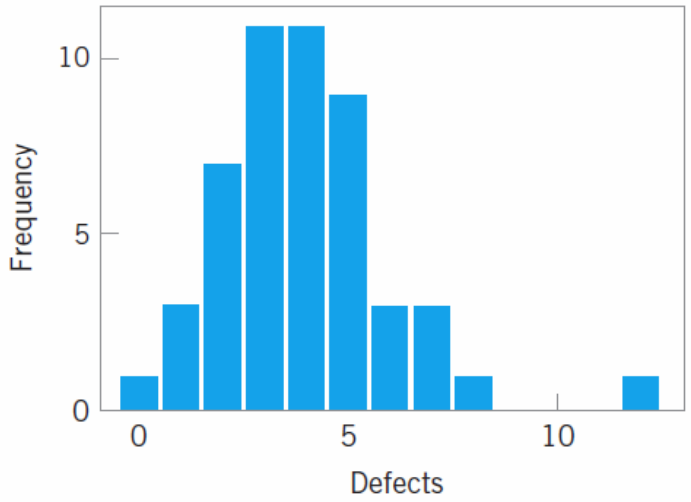


FIGURE 3.6 Histogram of the number of defects in painted automobile hoods (Table 3.3).

Numerical Summary of Data

Sample average:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (3.1)$$

Note that the sample average $\bar{x}$ is simply the arithmetic mean of the n observations. The sample average for the metal thickness data in Table 3.2 is

$$x = \frac{\sum_{i=1}^{100} x_i}{100} = \frac{45.001}{100} = 450.01 \text{Å}$$

Refer to Fig. 3.3 and note that the sample average is the point at which the histogram exactly "balances." Thus, the sample average represents the center of mass of the sample data.

The variability in the sample data is measured by the **sample variance:**

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \tag{3.2}$$

Note that the sample variance is simply the sum of the squared deviations of each observation from the sample average $\bar{x}$, divided by the sample size minus one. If there is no variability in the sample, then each sample observation $x_i = \bar{x}$, and the sample variance $s^2 = 0$. Generally, the larger is the sample variance s^2, the greater is the variability in the sample data.

The Standard Deviation

The units of the sample variance s^2 are the square of the original units of the data. This is often inconvenient and awkward to interpret, and so we usually prefer to use the square root of s^2, called the **sample standard deviation** s, as a measure of variability.

It follows that

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} \qquad (3.3)$$

The primary advantage of the sample standard deviation is that it is expressed in the original units of measurement. For the metal thickness data, we find that

$$s^2 = 180.2928 \text{ Å}^2$$

and

$$s = 13.43 \text{ Å}$$

The Box Plot
(or Box-and-Whisker Plot)

TABLE 3.4
Hole Diameters (in mm) in Wing Leading Edge Ribs

120.5	120.4	120.7
120.9	120.2	121.1
120.3	120.1	120.9
121.3	120.5	120.8

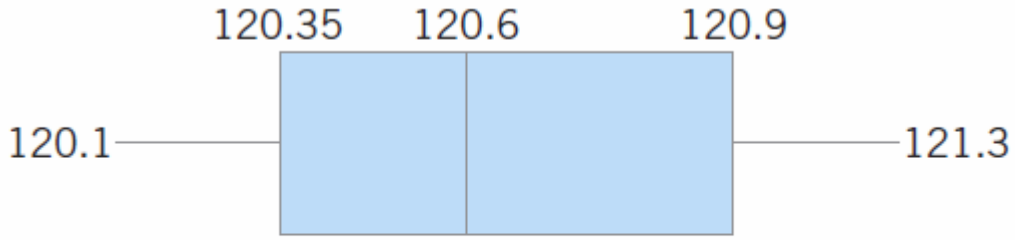


FIGURE 3.7 Box plot for the aircraft wing leading edge hole diameter data in Table 3.4.

Comparative Box Plots

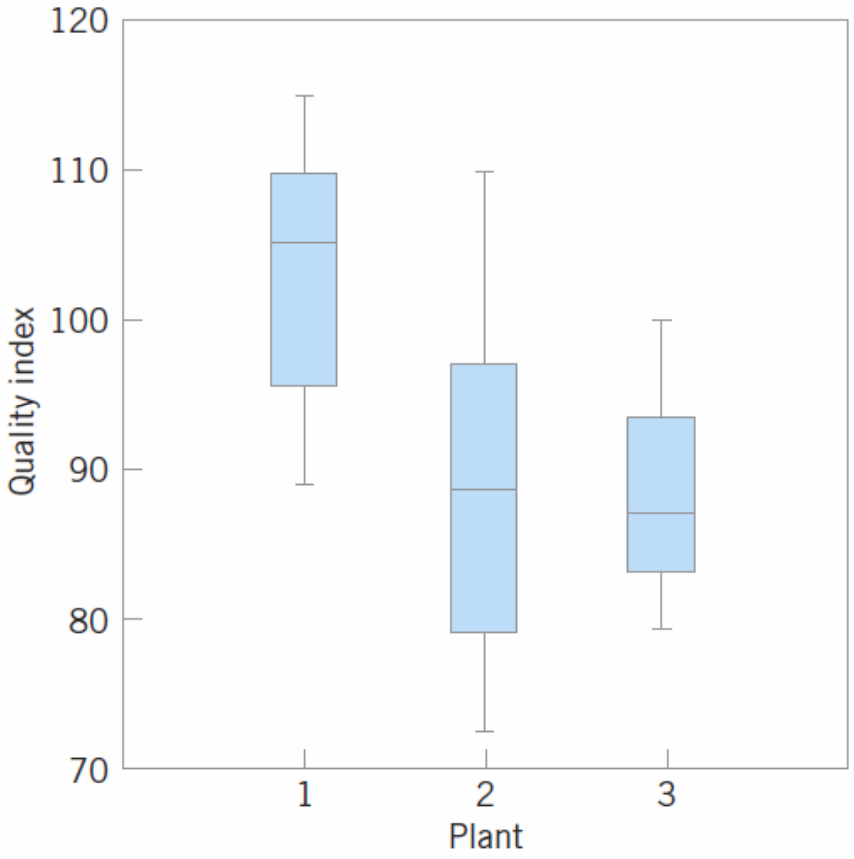


FIGURE 3.8 Comparative box plots of a quality index for products produced at three plants.

Probability Distributions

The histogram (or stem-and-leaf plot, or box plot) is used to describe *sample* data. A **sample** is a collection of measurements selected from some larger source or **population.** For example, the measurements on layer thickness in Table 3.2 are obtained from a sample of wafers selected from the manufacturing process. The population in this example is the collection of all layer thicknesses produced by that process. By using statistical methods, we may be able to analyze the sample layer thickness data and draw certain conclusions about the process that manufactures the wafers.

A **probability distribution** is a mathematical model that relates the value of the variable with the probability of occurrence of that value in the population. In other words, we might visualize layer thickness as a **random variable,** because it takes on different values in the population according to some random mechanism, and then the probability distribution of layer thickness describes the probability of occurrence of any value of layer thickness in the population. There are two types of probability distributions.

Definition

1. **Continuous distributions.** When the variable being measured is expressed on a continuous scale, its probability distribution is called a *continuous distribution.* The probability distribution of metal layer thickness is continuous.

2. **Discrete distributions.** When the parameter being measured can only take on certain values, such as the integers 0, 1, 2, . . . , the probability distribution is called a *discrete distribution.* For example, the distribution of the number of nonconformities or defects in printed circuit boards would be a discrete distribution.

Sometimes called a
probability mass function

Sometimes called a
probability density function

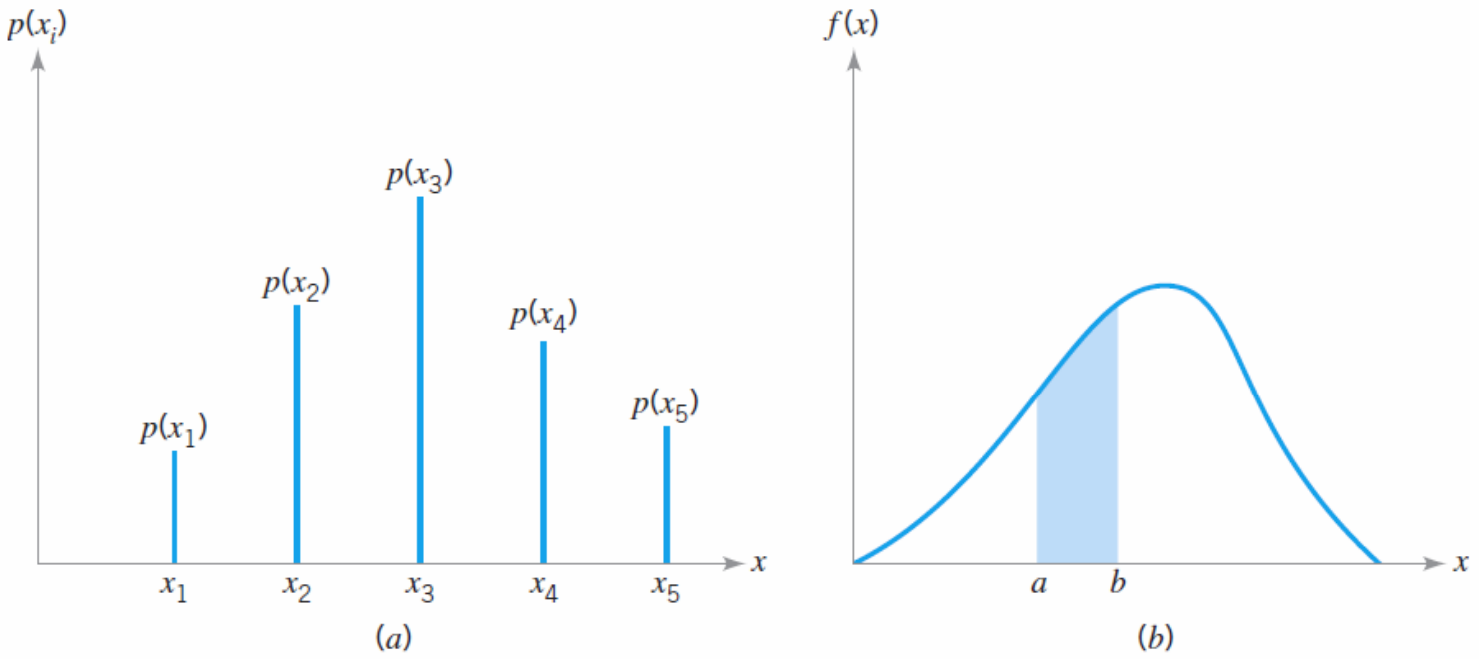


FIGURE 3.9 Probability distributions. (*a*) Discrete case. (*b*) Continuous case.

Will see many examples in the text

EXAMPLE 3.5 A Discrete Distribution

A manufacturing process produces thousands of semiconductor chips per day. On the average, 1% of these chips do not conform to specifications. Every hour, an inspector selects a random sample of 25 chips and classifies each chip in the sample as conforming or nonconforming. If we let x be the random variable representing the number of nonconforming chips in the sample, then the probability distribution of x is

$$p(x) = \binom{25}{x}(0.01)^x(0.99)^{25-x} \qquad x = 0,1,2,\ldots,25$$

where $\binom{25}{x} = 25!/[x!\,(25-x)!]$. This is a *discrete* distribution, since the observed number of nonconformances is $x = 0, 1, 2, \ldots, 25$, and is called the **binomial distribution.** We may calculate the probability of finding one or fewer nonconforming parts in the sample as

$$\begin{aligned} P(x \le 1) &= P(x=0) + P(x=1) \\ &= p(0) + p(1) \\ &= \sum_{x=0}^{1} \binom{25}{x}(0.01)^x(0.99)^{25-x} \\ &= \frac{25!}{0!25!}(0.99)^{25}(0.01)^0 + \frac{25!}{1!24!}(0.99)^{24}(0.01)^1 \\ &= 0.7778 + 0.1964 = 0.9742 \end{aligned}$$

EXAMPLE 3.6 A Continuous Distribution

Suppose that x is a random variable that represents the actual contents in ounces of a 1-pound bag of coffee beans. The probability distribution of x is assumed to be

$$f(x) = \frac{1}{1.5} \qquad 15.5 \le x \le 17.0$$

This is a *continuous* distribution, since the range of x is the interval [15.5, 17.0]. This distribution is called the **uniform distribution,** and it is shown graphically in Figure 3.10. Note that the area under the function $f(x)$ corresponds to probability, so that the probability of a bag containing less than 16.0 oz is

$$P\{x \le 16.0\} = \int_{15.5}^{16.0} f(x)dx = \int_{15.5}^{16.0} \frac{1}{1.5} dx$$

$$= \frac{x}{1.5}\Big|_{15.5}^{16.0} = \frac{16.0 - 15.5}{1.5} = 0.3333$$

This follows intuitively from inspection of Figure 3.9.

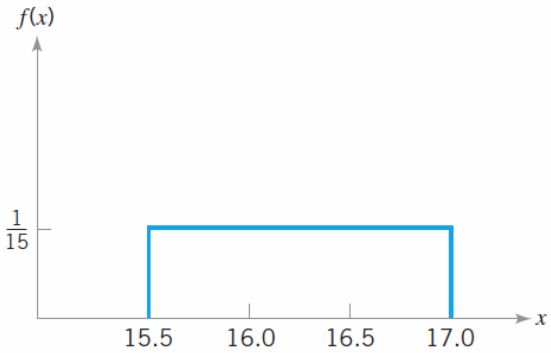


FIGURE 3.10 The uniform distribution for Example 3.6.

The **mean** μ of a probability distribution is a measure of the **central tendency** in the distribution, or its **location.** The mean is defined as

$$\mu = \begin{cases} \int_{-\infty}^{\infty} xf(x)\,dx, x \text{ continuous} & (3.5a) \\ \sum_{i=1}^{\infty} x_i p(x_i), x \text{ discrete} & (3.5b) \end{cases}$$

For the case of a discrete random variable with exactly N equally likely values [that is, $p(x_i) = 1/N$], then equation 3.5b reduces to

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

The mean is the point at which the distribution exactly "balances".

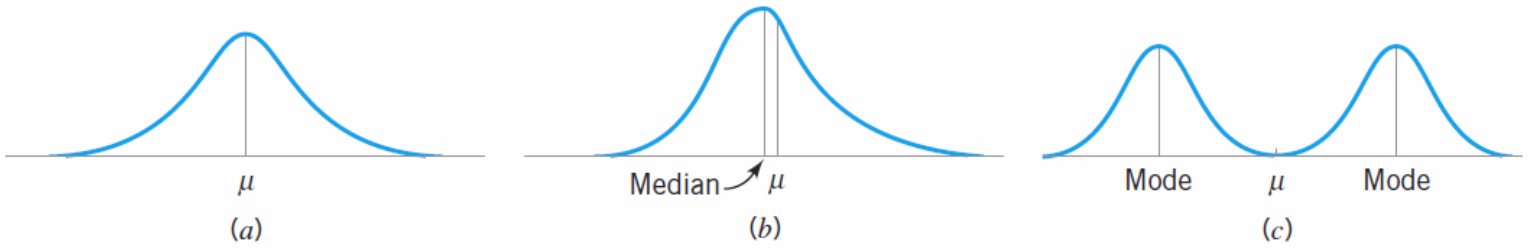


FIGURE 3.11 The mean of a distribution.

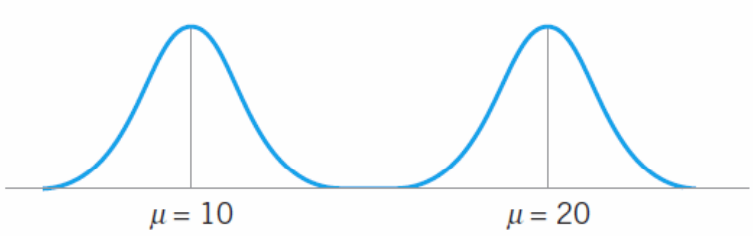


FIGURE 3.12 Two probability distributions with different means.

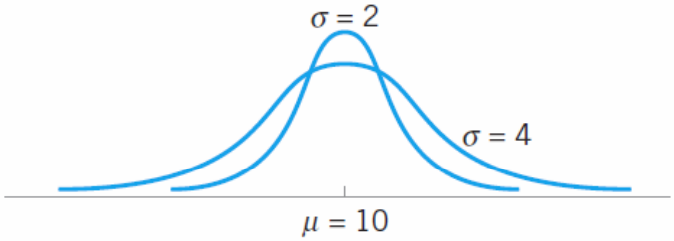


FIGURE 3.13 Two probability distributions with the same mean but different standard deviations.

The mean is not necessarily the 50th percentile of the distribution (that's the median)

The mean is not necessarily the most likely value of the random variable (that's the mode)

The scatter, spread, or variability in a distribution is expressed by the **variance** σ^2. The definition of the variance is

$$\sigma^2 = \begin{cases} \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx, x \text{ continuous} & (3.6a) \\ \sum_{i=1}^{\infty} (x_i-\mu)^2 p(x_i), x \text{ discrete} & (3.6b) \end{cases}$$

when the random variable is discrete with N equally likely values, then equation 3.6b becomes

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i-\mu)^2}{N}$$

and we observe that in this case the variance is the average squared distance of each member of the population from the mean. Note the similarity to the sample variance s^2, defined in equation 3.2. If $\sigma^2 = 0$, there is no variability in the population. As the variability increases, the variance σ^2 increases. The variance is expressed in the square of the units of the original variable. For example, if we are measuring voltages, the units of the variance are $(\text{volts})^2$. Thus, it is customary to work with the square root of the variance, called the **standard deviation** σ. It follows that

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{N} (x_i-\mu)^2}{N}} \qquad (3.7)$$

The standard deviation is a measure of spread or scatter in the population expressed in the original units. Two distributions with the same mean but different standard deviations are shown in Figure 3.13.

3.2 Important Discrete Distributions

The Hypergeometric Distribution

Definition

The **hypergeometric probability distribution** is

$$p(x) = \frac{\binom{D}{x}\binom{N-D}{n-x}}{\binom{N}{n}} \qquad x = 0,1,2,\ldots,\min(n,D) \tag{3.8}$$

The mean and variance of the distribution are

$$\mu = \frac{nD}{N} \tag{3.9}$$

and

$$\sigma^2 = \frac{nD}{N}\left(1-\frac{D}{N}\right)\left(\frac{N-n}{N-1}\right) \tag{3.10}$$

The hypergeometric distribution is the appropriate probability model for selecting a random sample of n items without replacement from a lot of N items of which D are nonconforming or defective. By a random sample, we mean a sample that has been selected in such a way that all possible samples have an equal chance of being chosen. In these applications, x usually represents the number of nonconforming items found in the sample. For example, suppose that a lot contains 100 items, 5 of which do not conform to requirements. If 10 items are selected at random without replacement, then the probability of finding one or fewer nonconforming items in the sample is

$$P\{x \le 1\} = P\{x = 0\} + P\{x = 1\}$$
$$= \frac{\binom{5}{0}\binom{95}{10}}{\binom{100}{10}} + \frac{\binom{5}{1}\binom{95}{9}}{\binom{100}{10}} = 0.923$$

Discrete distributions are used frequently in designing acceptance sampling plans – see Chapter 15

Some computer programs can perform these calculations. The display below is the output from Minitab for calculating cumulative hypergeometric probabilities with $N = 100$, $D = 5$ (note that Minitab uses the symbol M instead of D and $n = 10$). Minitab will also calculate the individual probabilities for each value of x.

Cumulative Distribution Function

Hypergeometric with N = 100, M = 5, and n = 10

x	P(X <= x)	x	P(X <= x)
0	0.58375	6	1.00000
1	0.92314	7	1.00000
2	0.99336	8	1.00000
3	0.99975	9	1.00000
4	1.00000	10	1.00000
5	1.00000		

The Binomial Distribution

Basis is in **Bernoulli trials**

Definition

The **binomial distribution** with parameters $n \geq 0$ and $0 < p < 1$ is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, \ldots, n \tag{3.11}$$

The mean and variance of the binomial distribution are

$$\mu = np \tag{3.12}$$

and

$$\sigma^2 = np(1-p) \tag{3.13}$$

The random variable x is the number of successes out of n Bernoulli trials with constant probability of success p on each trial

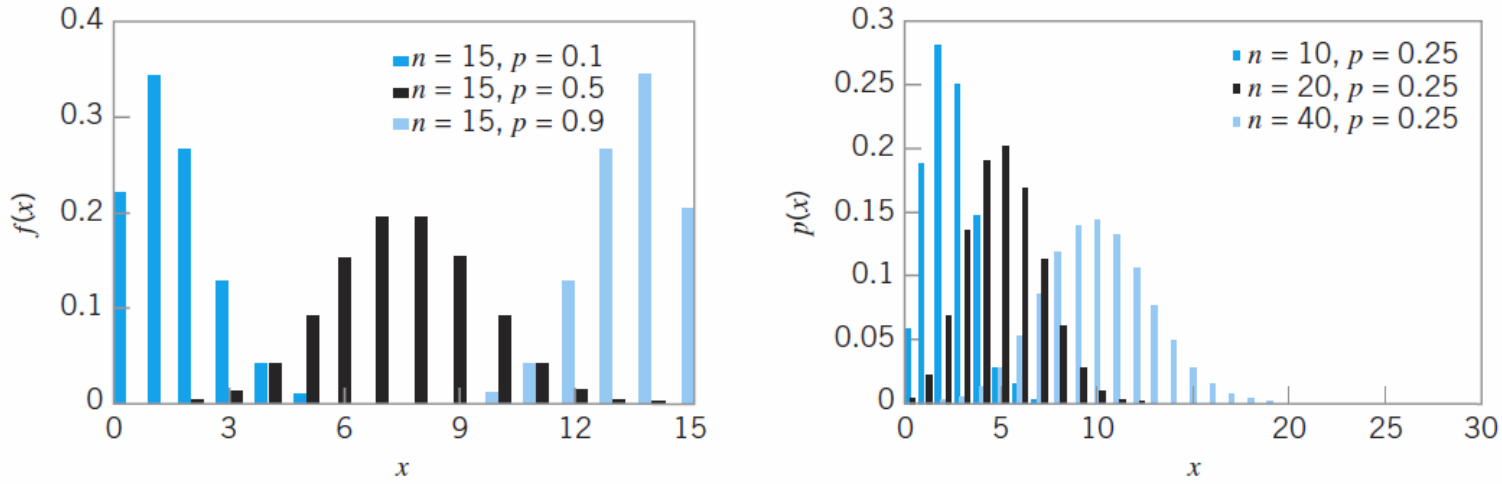
The binomial distribution is used frequently in quality engineering. It is the appropriate probability model for sampling from an infinitely large population, where p represents the fraction of defective or nonconforming items in the population. In these applications, x usually represents the number of nonconforming items found in a random sample of n items. For example, if $p = 0.10$ and $n = 15$, then the probability of obtaining x nonconforming items is computed from equation 3.11 as follows:

Probability Density Function

Binomial with n = 15 and p = 0.1

x	P(X = x)	x	P(X = x)
0	0.205891	6	0.001939
1	0.343152	7	0.000277
2	0.266896	8	0.000031
3	0.128505	9	0.000003
4	0.042835	10	0.000000
5	0.010471		

Binomial Distributions



(*a*) Binomial distributions for different values of p with $n = 15$.

(*b*) Binomial distributions for different values of n with $p = 0.25$.

FIGURE 3.14 Binomial distributions for selected values of n and p.

A random variable that arises frequently in statistical quality control is

$$\hat{p} = \frac{x}{n} \tag{3.14}$$

where x has a binomial distribution with parameters n and p. Often $\hat{p}$ is the ratio of the observed number of defective or nonconforming items in a sample (x) to the sample size (n) and this is usually called the **sample fraction defective** or **sample fraction nonconforming.** The "^" symbol is used to indicate that $\hat{p}$ is an estimate of the true, unknown value of the binomial parameter p. The probability distribution of $\hat{p}$ is obtained from the binomial, since

$$P\{\hat{p} \le a\} = P\left\{\frac{x}{n} \le a\right\} = p\{x \le na\} = \sum_{x=0}^{[na]} \binom{n}{x} p^x (1-p)^{n-x}$$

where $[na]$ denotes the largest integer less than or equal to na. It is easy to show that the mean of $\hat{p}$ is p and that the variance of $\hat{p}$ is

$$\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$$

The Poisson Distribution

Definition

The **Poisson distribution** is

$$p(x) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad x = 0,1,\ldots \tag{3.15}$$

where the parameter $\lambda > 0$. The **mean** and **variance** of the Poisson distribution are

$$\mu = \lambda \tag{3.16}$$

and

$$\sigma^2 = \lambda \tag{3.17}$$

Frequently used as a model for count data

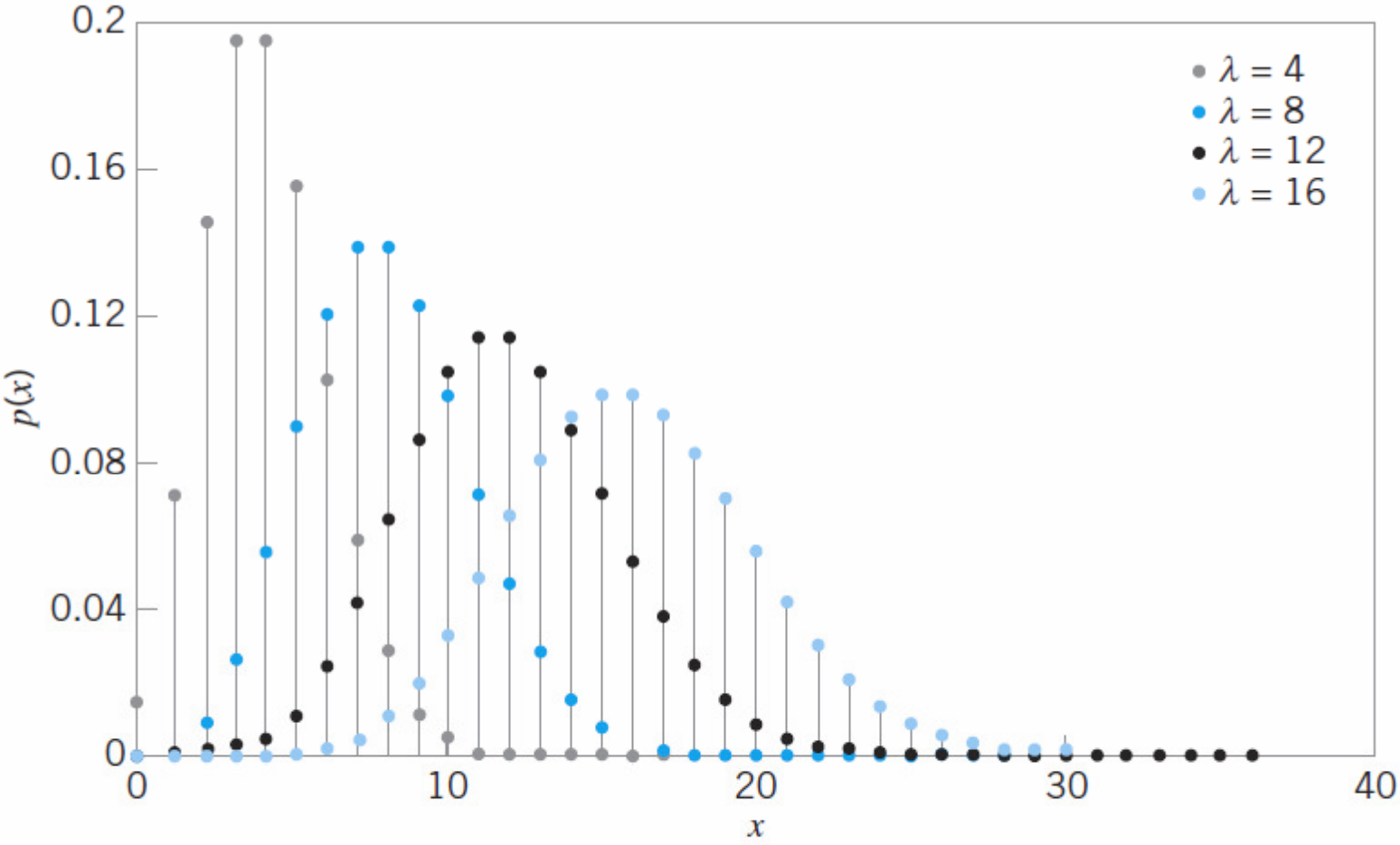


FIGURE 3.15 Poisson probability distributions for selected values of λ.

A typical application of the Poisson distribution in quality control is as a model of the number of defects or nonconformities that occur in a unit of product. In fact, any random phenomenon that occurs on a per unit (or per unit area, per unit volume, per unit time, etc.) basis is often well approximated by the Poisson distribution. As an example, suppose that the number of wire-bonding defects per unit that occur in a semiconductor device is Poisson distributed with parameter $\lambda = 4$. Then the probability that a randomly selected semiconductor device will contain two or fewer wire-bonding defects is

$$P\{x \le 2\} = \sum_{x=0}^{2} \frac{e^{-4}4^{x}}{x!} = 0.018316 + 0.073263 + 0.146525 = 0.238104$$

Probability Density Function

```
Poisson with mean = 4

x   P(X = x)
0     0.018316
1     0.073263
2     0.146525
```

The Negative Binomial Distribution

Definition

The **negative binomial distribution** is

$$p(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \qquad x = r, r+1, r+2, \ldots \tag{3.18}$$

where $r \geq 1$ is an integer. The ***mean*** and ***variance*** of the negative binomial distribution are

$$\mu = \frac{r}{p} \tag{3.19}$$

and

$$\sigma^2 = \frac{r(1-p)}{p^2} \tag{3.20}$$

respectively.

The random variable x is the number of Bernoulli trials upon which the rth success occurs

- The negative binomial distribution is also sometimes called the Pascal distribution
- When $r = 1$ the negative binomial distribution is known as the **geometric** distribution
- The geometric distribution has many useful applications in SQC

Geometric Distribution

$$p(x) = (1-p)^{x-1}p, \quad x = 1, 2, \ldots$$

The mean and variance of the geometric distribution are

$$\mu = \frac{1}{p} \quad \text{and} \quad \sigma^2 = \frac{1-p}{p^2}$$

respectively. Because the sequence of Bernoulli trials are independent, the count of the number of trials until the next success can be started from anywhere without changing the probability distribution. For example, suppose we are examining a series of medical records searching for missing information. If, for example, 100 records have been examined, the probability that the first error occurs on record number 105 is just the probability that the next five records are GGGGB, where G denotes good and B denotes an error. If the probability of finding a bad record is 0.05, the probability of finding a bad record on the fifth record examined is $P\{x = 5\} = (0.95)^4(0.05) = 0.0407$. This is identical to the probability that the first bad record occurs on record 5. This is called the **lack of memory property** of the geometric distribution. This property implies that the system being modeled does not fail because it is wearing out due to fatigue or accumulated stress.

3.3 Important Continuous Distributions

The Normal Distribution

Definition

The **normal distribution** is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \qquad -\infty < x < \infty \qquad (3.21)$$

The mean of the normal distribution is μ $(-\infty < \mu < \infty)$ and the variance is $\sigma^2 > 0$.

The normal distribution is used so much that we frequently employ a special notation, $x - N(\mu, \sigma^2)$, to imply that x is normally distributed with mean μ and variance σ^2. The visual appearance of the normal distribution is a symmetric, unimodal or **bell-shaped** curve and is shown in Figure 3.16.

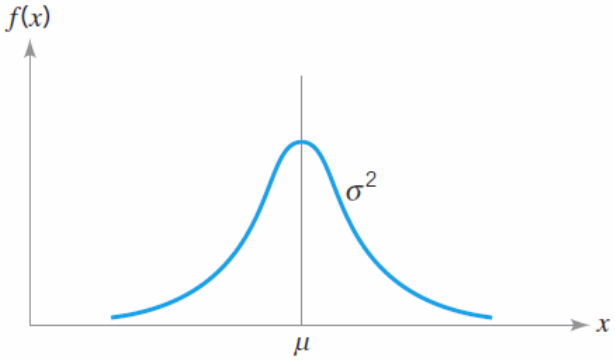


FIGURE 3.16 The normal distribution.

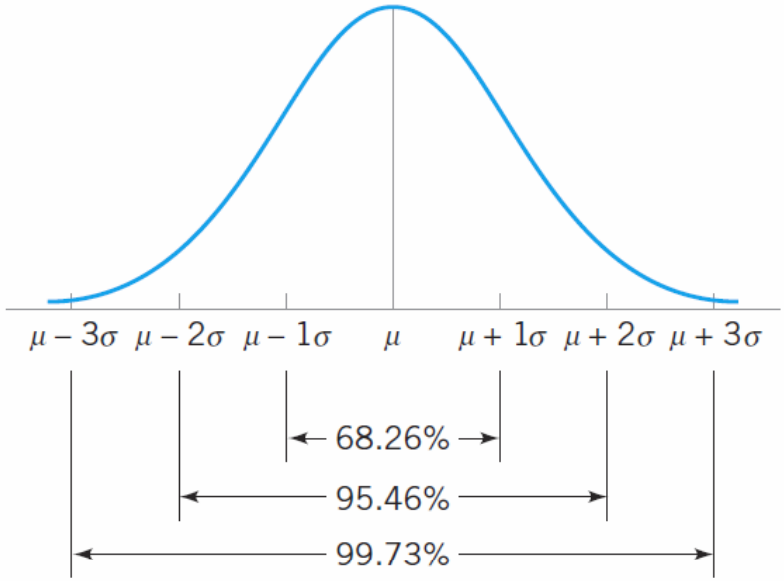


FIGURE 3.17 Areas under the normal distribution.

The cumulative normal distribution is defined as the probability that the normal random variable x is less than or equal to some value a, or

$$P\{x \le a\} = F(a) = \int_{-\infty}^{a} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \tag{3.22}$$

This integral cannot be evaluated in closed form. However, by using the change of variable

$$z = \frac{x-\mu}{\sigma} \tag{3.23}$$

the evaluation can be made independent of μ and σ^2. That is,

$$P\{x \le a\} = P\left\{z \le \frac{a-\mu}{\sigma}\right\} \equiv \Phi\left(\frac{a-\mu}{\sigma}\right)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the **standard normal distribution** (mean = 0, standard deviation = 1). A table of the cumulative standard normal distribution is given in Appendix Table II. The transformation (3.23) is usually called **standardization,** because it converts a $N(\mu, \sigma^2)$ random variable into an $N(0, 1)$ random variable.

EXAMPLE 3.7 Tensile Strength of Paper

The time to resolve customer complaints is a critical quality characteristic for many organizations. Suppose that this time in a financial organization, say, x—is normally distributed with mean $\mu = 40$ hours and standard deviation $\sigma = 2$ hours denoted $x \sim N(40, 2^2)$. What is the probability that a customer complaint will be resolved in less than 35 hours?

SOLUTION

The desired probability is

$$P\{x \le 35\}$$

To evaluate this probability from the standard normal tables, we standardize the point 35 and find

$$P\{x \le 35\} = P\left\{z \le \frac{35 - 40}{2}\right\} =$$

$$P\{z \le -2.5\} = \Phi(-2.5) = 0.0062$$

Consequently, the desired probability is

$$p\{x \ge 35\} = 0.0062$$

Figure 3.18 shows the tabulated probability for both the $N(40, 2^2)$ distribution and the standard normal distribution. Note that the shaded area to the left of 35 hr in Figure 3.18 represents the fraction of customer complaints resolved in less than or equal to 35 hours.

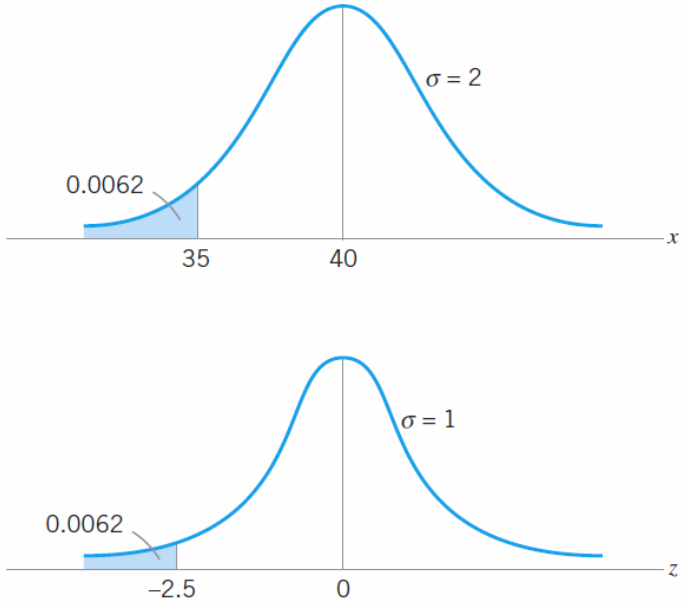


FIGURE 3.18 Calculation of $P\{x \le 35\}$ in Example 3.7.

EXAMPLE 3.9 Another Use of the Standard Normal Table

Sometimes instead of finding the probability associated with a particular value of a normal random variable, we find it necessary to do the opposite—find a particular value of a normal random variable that results in a given probability. For example, suppose that $x \sim N(10, 9)$. Find the value of x—say, a—such that $P\{x > a\} = 0.05$.

SOLUTION

From the problem statement, we have

$$P\{x > a\} = P\left\{z > \frac{a-10}{3}\right\} = 0.05$$

or

$$P\left\{z \leq \frac{a-10}{3}\right\} = 0.95$$

From Appendix Table II, we have $P\{z \leq 1.645\} = 0.95$, so

$$\frac{a-10}{3} = 1.645$$

or

$$a = 10 + 3(1.645) = 14.935$$

The normal distribution has many useful properties. One of these is relative to **linear combinations** of normally and independently distributed random variables. If $x_1, x_2 \ldots, x_n$ are normally and independently distributed random variables with means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, respectively, then the distribution of

$$y = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n$$

is normal with mean

$$\mu_y = a_1\mu_1 + a_2\mu_2 + \cdots + a_n\mu_n \tag{3.27}$$

and variance

$$\sigma_y^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \cdots + a_n^2\sigma_n^2 \tag{3.28}$$

where $a_1, a_2, \ldots, a_n$ are constants.

The Central Limit Theorem

Definition

The Central Limit Theorem If $x_1, x_2, \ldots, x_n$ are independent random variables with mean μ_i and variance σ_i^2, and if $y = x_1 + x_2 + \cdots + x_n$, then the distribution of

$$\frac{y - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}$$

approaches the $N(0, 1)$ distribution as n approaches infinity.

Practical interpretation – the sum of independent random variables is approximately normally distributed regardless of the distribution of each individual random variable in the sum

The Lognormal Distribution

Definition

Let w have a normal distribution mean θ and variance ω^2; then $x = \exp(w)$ is a **lognormal random variable,** and the lognormal distribution is

$$f(x) = \frac{1}{x\omega\sqrt{2\pi}} \exp\left[-\frac{(\ln(x)-\theta)^2}{2\omega^2}\right] \quad 0 < x < \infty \tag{3.29}$$

The mean and variance of x are

$$\mu = e^{\theta+\omega^2/2} \quad \text{and} \quad \sigma^2 = e^{2\theta+\omega^2}\left(e^{\omega^2}-1\right) \tag{3.30}$$

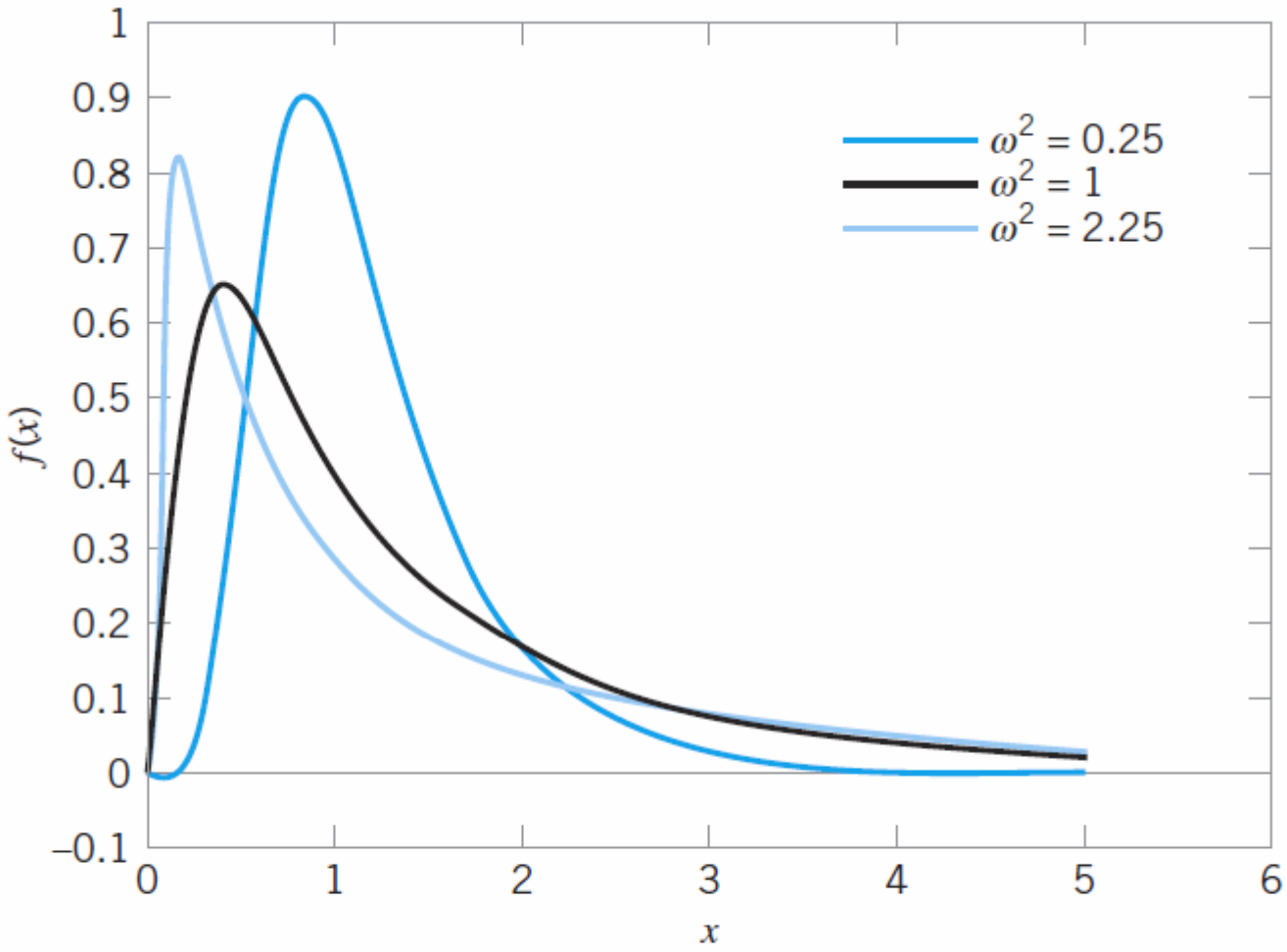


FIGURE 3.20 Lognormal probability density functions with $\theta = 0$ for selected values of ω^2.

EXAMPLE 3.10 Medical Laser Lifetime

The lifetime of a medical laser used in ophthalmic surgery has a lognormal distribution with $\theta = 6$ and $\omega = 1.2$ hours What is the probability that the lifetime exceeds 500 hours?

SOLUTION

From the cumulative distribution function for the lognormal random variable

$$P(x > 500) = 1 - P[\exp(w) \le 500] = 1 - P[w \le \ln(500)]$$

$$= \Phi\left(\frac{\ln(500) - 6}{1.2}\right) = 1 - \Phi(0.1788)$$

$$= 1 - 0.5710 = 0.4290$$

What lifetime is exceeded by 99% of lasers? Now the question is to determine a such that $P(x > a) = 0.99$. Therefore,

$$P(x > a) = P[\exp(w) > a] = P[w > \ln(a)]$$

$$= 1 - \Phi\left(\frac{\ln(a) - 6}{1.2}\right) = 0.99$$

From Appendix Table II, $1 - \Phi(a) = 0.99$ when $a = -2.33$. Therefore,

$$\frac{\ln(a) - 6}{1.2} = -2.33 \quad \text{and} \quad a = \exp(3.204) = 24.63 \text{ hours}$$

Determine the mean and standard deviation of the lifetime. Now,

$$\mu = e^{\theta + \omega^2/2} = \exp(6 + 0.72) = 828.82 \text{ hours}$$

$$\sigma^2 = e^{2\theta + \omega^2}\left(e^{\omega^2} - 1\right) = \exp(12 + 1.44)[\exp(1.44) - 1]$$

$$= 2,212,419.85$$

so the standard deviation of the lifetime is 1487.42 hours. Notice that the standard deviation of the lifetime is large relative to the mean.

The Exponential Distribution

Definition

The **exponential distribution** is

$$f(x) = \lambda e^{-\lambda x} \qquad x \geq 0 \tag{3.31}$$

where $\lambda > 0$ is a constant. The **mean** and **variance** of the exponential distribution are

$$\mu = \frac{1}{\lambda} \tag{3.32}$$

and

$$\sigma^2 = \frac{1}{\lambda^2} \tag{3.33}$$

respectively.

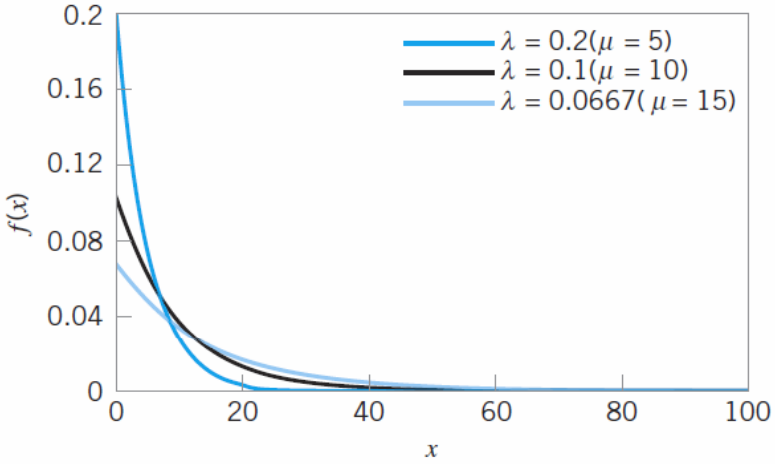


FIGURE 3.21 Exponential distributions for selected values of λ.

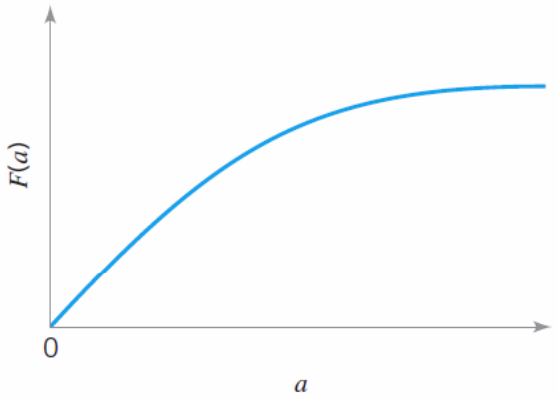


FIGURE 3.22 The cumulative exponential distribution function.

Relationship between the Poisson and exponential distributions

Lack-of-memory property

The exponential distribution has a **lack of memory** property. To illustrate, suppose that the exponential random variable x is used to model the time to the occurrence of some event. Consider two points in time t_1 and $t_2 > t_1$. Then the probability that the event occurs at a time that is less than $t_1 + t_2$ but greater than time t_2 is just the probability that the event occurs at time less than t_1. This is the same lack of memory property that we observed earlier for the geometric distribution. The exponential distribution is the only continuous distribution that has this property.

The Gamma Distribution

Definition

The **gamma distribution** is

$$f(x) = \frac{\lambda}{\Gamma(r)}(\lambda x)^{r-1} e^{-\lambda x} \qquad x \geq 0 \tag{3.36}$$

with **shape parameter** $r > 0$ and **scale parameter** $\lambda > 0$. The **mean** and **variance** of the gamma distribution are

$$\mu = \frac{r}{\lambda} \tag{3.37}$$

and

$$\sigma^2 = \frac{r}{\lambda^2} \tag{3.38}$$

respectively.[3]

[3] $\Gamma(r)$ in the denominator of equation 3.36 is the gamma function, defined as $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx$, $r > 0$. If r is a positive integer, then $\Gamma(r) = (r-1)!$

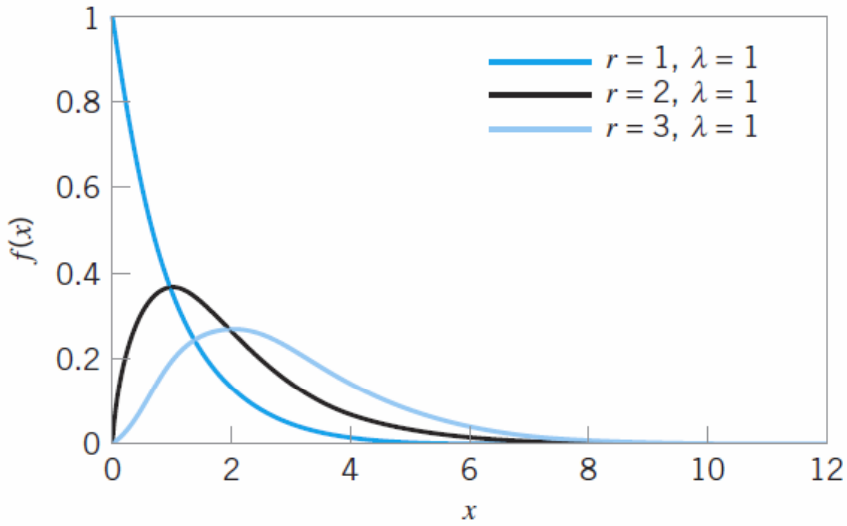


FIGURE 3.23 Gamma distributions for selected values or r and $\lambda = 1$.

When *r* is an integer, the gamma distribution is the result of summing *r* independently and identically exponential random variables each with parameter λ.

The gamma distribution has many applications in reliability engineering.

EXAMPLE 3.11 A Standby Redundant System

Consider the system shown in Figure 3.24. This is called a **standby redundant system,** because while component 1 is on, component 2 is off, and when component 1 fails, the switch automatically turns component 2 on. If each component has a life described by an exponential distribution with $\lambda = 10^{-4}$/h, say, then the system life is gamma distributed with parameters $r = 2$ and $\lambda = 10^{-4}$. Thus, the mean time to failure is $\mu = r/\lambda = 2/10^{-4} = 2 \times 10^4$ h.

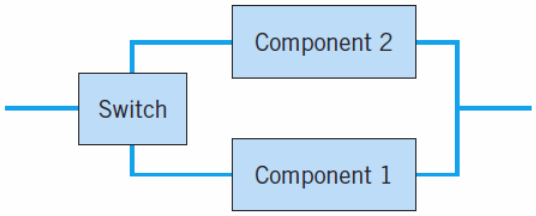


FIGURE 3.24 The standby redundant system for Example 3.11.

The Weibull Distribution

Definition

The **Weibull distribution** is

$$f(x) = \frac{\beta}{\theta}\left(\frac{x}{\theta}\right)^{\beta-1} \exp\left[-\left(\frac{x}{\theta}\right)^{\beta}\right] \qquad x \geq 0 \tag{3.41}$$

where $\theta > 0$ is the **scale parameter** and $\beta > 0$ is the **shape parameter.** The **mean** and **variance** of the Weibull distribution are

$$\mu = \theta\,\Gamma\left(1+\frac{1}{\beta}\right) \tag{3.42}$$

and

$$\sigma^2 = \theta^2\left[\Gamma\left(1+\frac{2}{\beta}\right) - \left\{\Gamma\left(1+\frac{1}{\beta}\right)\right\}^2\right] \tag{3.43}$$

respectively.

When β = 1, the Weibull reduces to the exponential

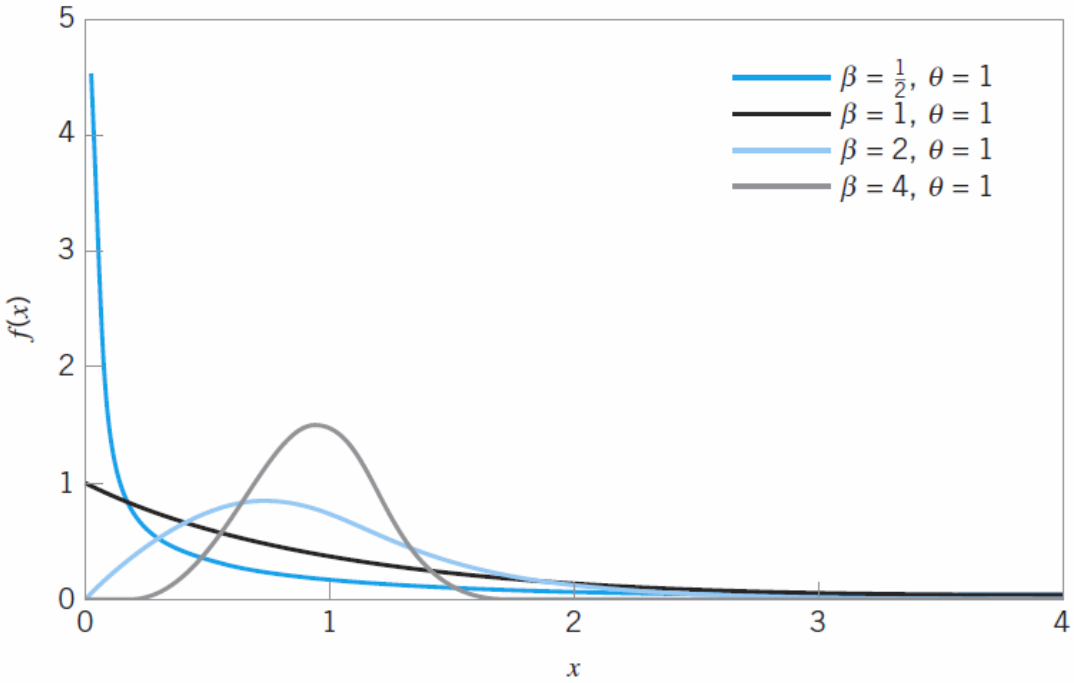


FIGURE 3.25 Weibull distributions for selected values of the shape parameter β and scale parameter $\theta = 1$.

An Application of the Weibull Distribution

EXAMPLE 3.12 Time to Failure for Electronic Components

The time to failure for an electronic component used in a flat panel display unit is satisfactorily modeled by a Weibull distribution with $\beta = \frac{1}{2}$ and $\theta = 5000$. Find the mean time to failure and the fraction of components that are expected to survive beyond 20,000 hours.

SOLUTION

The mean time to failure is

$$\mu = \theta\Gamma\left(1+\frac{1}{\beta}\right) = 5000\Gamma\left(1+\frac{1}{\frac{1}{2}}\right)$$

$$= 5000\Gamma(3) = 10{,}000 \text{ hours}$$

The fraction of components expected to survive $a = 20{,}000$ hours is

$$1 - F(a) = \exp\left[-\left(\frac{a}{\theta}\right)^{\beta}\right]$$

or

$$1 - F(20{,}000) = \exp\left[-\left(\frac{20{,}000}{5{,}000}\right)^{\frac{1}{2}}\right]$$

$$= e^{-2}$$

$$= 0.1353$$

That is, all but about 13.53% of the subassemblies will fail by 20,000 hours.

3.4 Probability Plots

- Determining if a sample of data might reasonably be assumed to come from a specific distribution
- Probability plots are available for various distributions
- Easy to construct with computer software (MINITAB)
- Subjective interpretation

Normal Probability Plot

EXAMPLE 3.13 A Normal Probability Plot

Observations on the road octane number of ten gasoline blends are as follows: 88.9, 87.0, 90.0, 88.2, 87.2, 87.4, 87.8, 89.7, 86.0, and 89.6. We hypothesize that the octane number is adequately modeled by a normal distribution. Is this a reasonable assumption?

SOLUTION

To use probability plotting to investigate this hypothesis, first arrange the observations in ascending order and calculate their cumulative frequencies $(j - 0.5)/10$ as shown in the following table.

j	$x_{(j)}$	$(j - 0.5)/10$
1	86.0	0.05
2	87.0	0.15
3	87.2	0.25
4	87.4	0.35
5	87.8	0.45
6	88.2	0.55
7	88.9	0.65
8	89.6	0.75
9	89.7	0.85
10	90.0	0.95

The pairs of values $x_{(j)}$ and $(j - 0.5)/10$ are now plotted on normal probability paper. This plot is shown in Figure 3.26. Most normal probability paper plots $100(j - 0.5)/n$ on the left vertical scale (and some also plot $100[1 - (j - 0.5)/n]$ on the right vertical scale), with the variable value plotted on the horizontal scale. A straight line, chosen subjectively as a "best fit" line, has been drawn through the plotted points. In drawing the straight line, you should be influenced more by the points near the middle of the plot than by the extreme points. A good rule of thumb is to draw the line approximately between the twenty-fifth and seventy-fifth percentile points. This is how the line in Figure 3.26 was determined. In assessing the systematic deviation of the points from the straight line, imagine a fat pencil lying along the line. If all the points are covered by this imaginary pencil, a normal distribution adequately describes the data. Because the points in Figure 3.26 would pass the fat pencil test, we conclude that the normal distribution is an appropriate model for the road octane number data.

(*continued*)

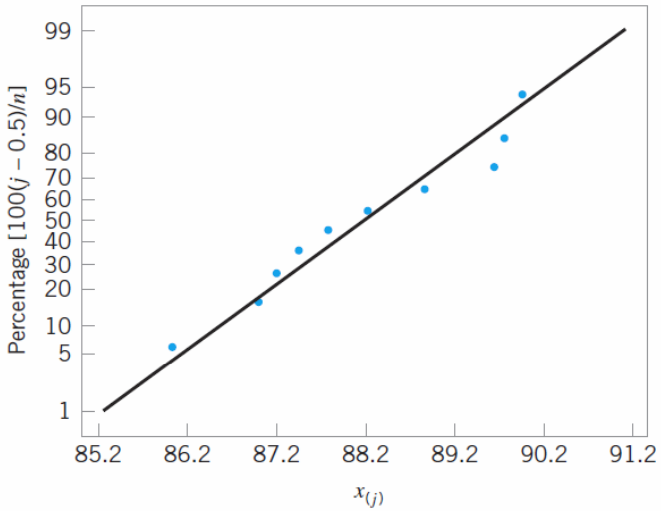


FIGURE 3.26 Normal probability plot of the road octane number data.

The Normal Probability Plot on Standard Graph Paper

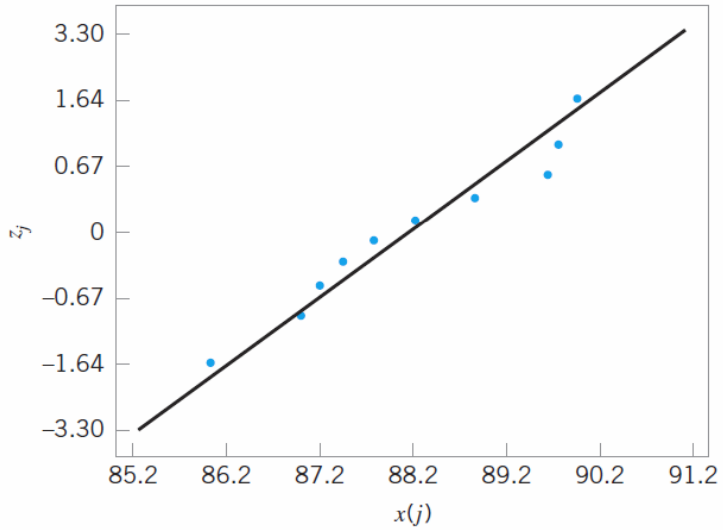


FIGURE 3.27 Normal probability plot of the road octane number data with standardized scores.

j	$x_{(j)}$	$(j - 0.5)/10$	z_j
1	86.0	0.05	–1.64
2	87.0	0.15	–1.04
3	87.2	0.25	–0.67
4	87.4	0.35	–0.39
5	87.8	0.45	–0.13
6	88.2	0.55	0.13
7	88.9	0.65	0.39
8	89.6	0.75	0.67
9	89.7	0.85	1.04
10	90.0	0.95	1.64

Other Probability Plots

- What is a reasonable choice as a probability model for these data?

TABLE 3.5

Aluminum Contamination (ppm)

30	30	60	63	70	79	87
90	101	102	115	118	119	119
120	125	140	145	172	182	
183	191	222	244	291	511	

From "The Lognormal Distribution for Modeling Quality Data When the Mean Is Near Zero," *Journal of Quality Technology*, 1990, pp. 105–110.

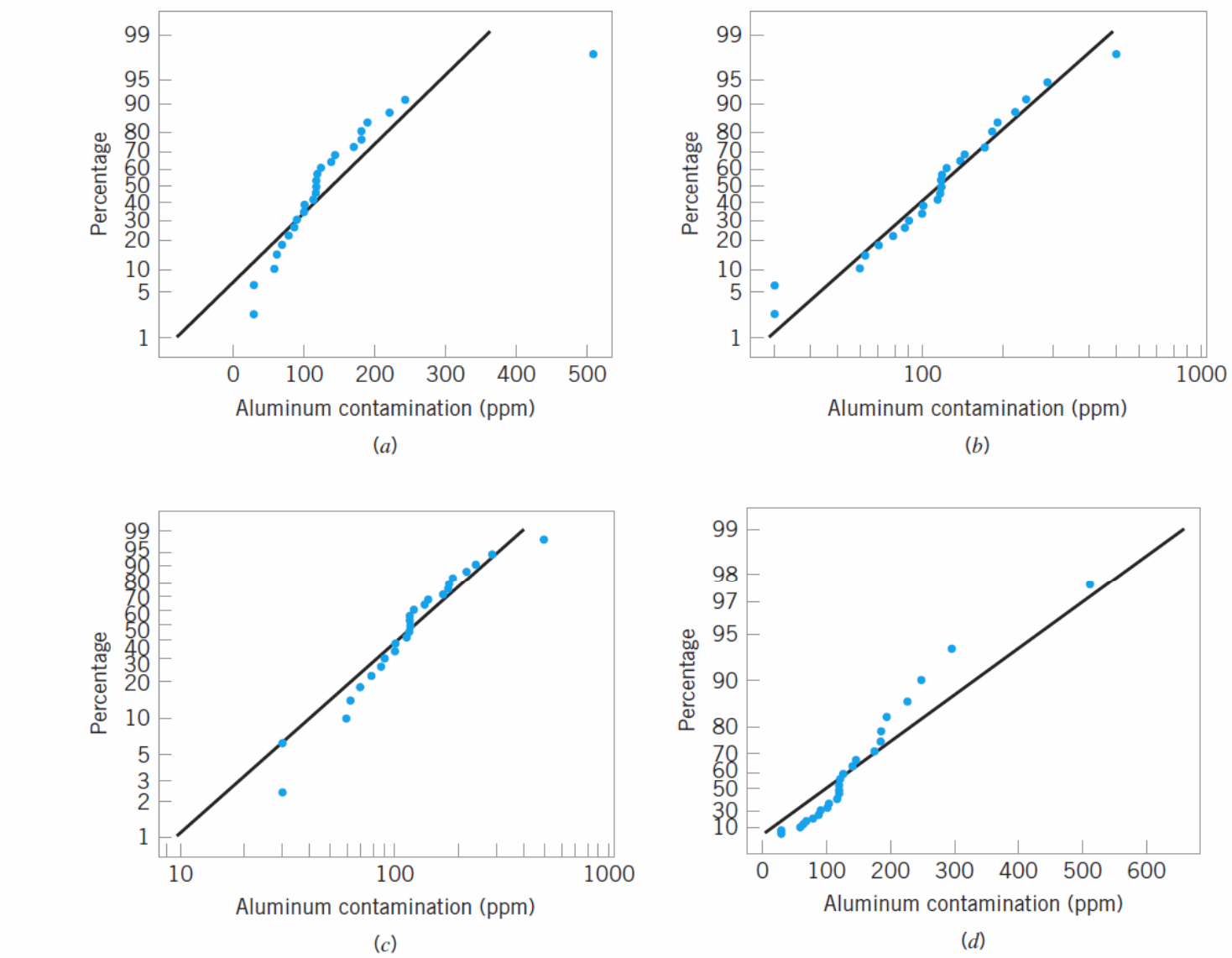


FIGURE 3.28 Probability plots of the aluminum contamination data in Table 3.5. (*a*) Normal. (*b*) Lognormal. (*c*) Weibull. (*d*) Exponential.

3.5 Some Useful Approximations

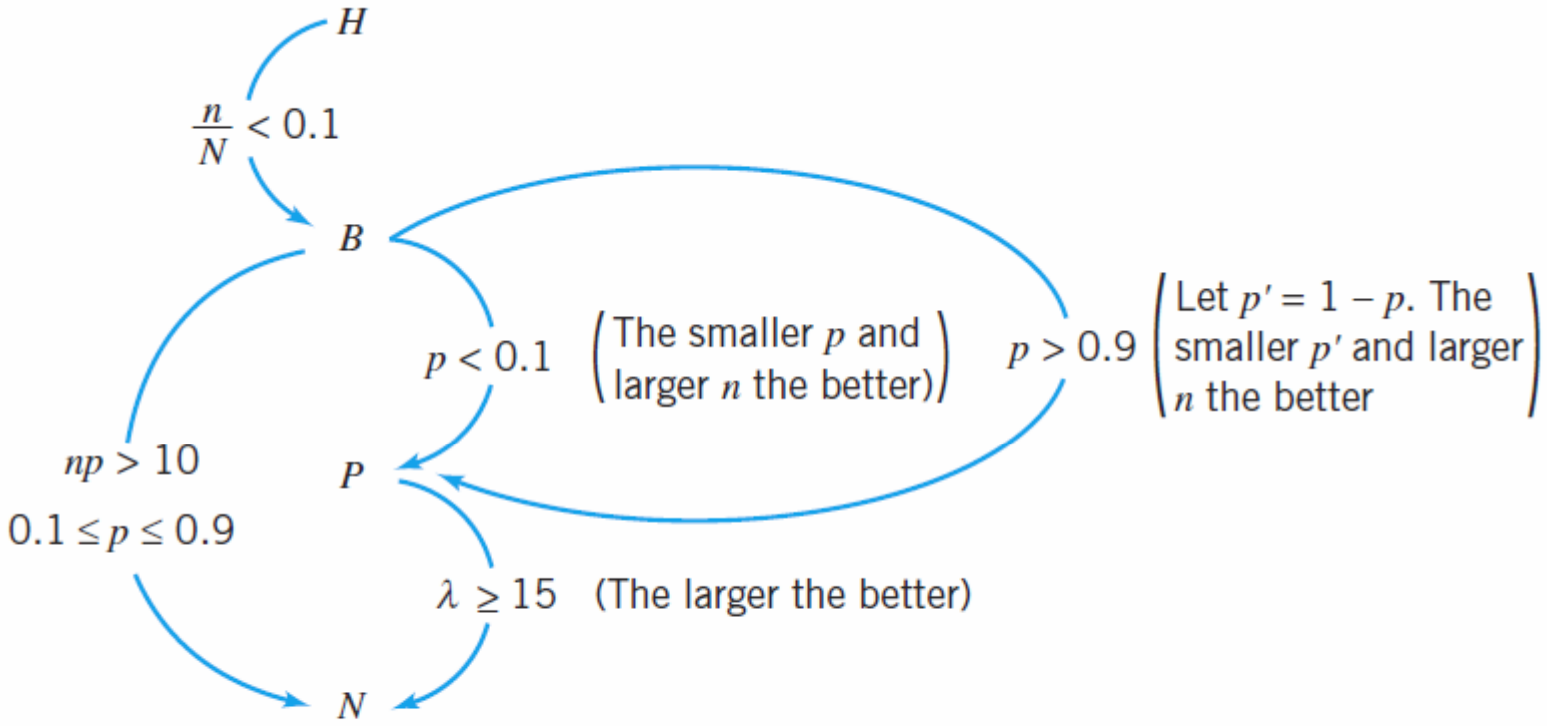


FIGURE 3.29 Approximations to probability distributions.

Important Terms and Concepts

Approximations to probability distributions
Binomial distribution
Box plot
Central limit theorem
Continuous distribution
Control limit theorem
Descriptive statistics
Discrete distribution
Exponential distribution
Gamma distribution
Geometric distribution
Histogram
Hypergeometric probability distribution
Interquartile range
Lognormal distribution
Mean of a distribution
Median
Negative binomial distribution
Normal distribution
Normal probability plot
Pascal distribution
Percentile
Poisson distribution
Population
Probability distribution
Probability plotting
Quartile
Random variable
Run chart
Sample
Sample average
Sample standard deviation
Sample variance
Standard deviation
Standard normal distribution
Statistics
Stem-and-leaf display
Time series plot
Uniform distribution
Variance of a distribution
Weibull distribution

Learning Objectives

1. Construct and interpret visual data displays, including the stem-and-leaf plot, the histogram, and the box plot
2. Compute and interpret the sample mean, the sample variance, the sample standard deviation, and the sample range
3. Explain the concepts of a random variable and a probability distribution
4. Understand and interpret the mean, variance, and standard deviation of a probability distribution
5. Determine probabilities from probability distributions
6. Understand the assumptions for each of the discrete probability distributions presented
7. Understand the assumptions for each of the continuous probability distributions presented
8. Select an appropriate probability distribution for use in specific applications
9. Use probability plots
10. Use approximations for some hypergeometric and binomial distributions